Warsaw University of Technology

Faculty of Electronics and Information Technology,

Institute of Electronic Systems

Alicia López Dot

Senior Design Project Report

DATA MANAGEMENT FOR WEB-BASED E-HEALTH EXPERT SYSTEM

ADVISING TOOL RULE-BASED ON OUTCOMES OF EPIDEMIOLOGIC SURVEY FOR USERS WITH POSSIBLE ALLERGIC DISEASES

Project done under the supervision of

Dr inż. Zbigniew Wawrzyniak, PhD

Warszawa, March 2012

Index

Contents table

Abstract

Spanish:

El conocimiento tiene su base en la información, pero ésta necesita de habilidades de razonamiento, deducción, intuición y experiencia para ser útil. El tratamiento de datos es un conjunto de técnicas encaminadas a la extracción de conocimiento procesable implícito en las bases de datos con el objetivo de buscar solución a problemas de predicción y clasificación.

Éste proyecto ha supuesto la inmersión en el mundo de la minería de datos con el objetivo de crear una herramienta de ayuda para el tratamiento de datos de forma que se obtenga una mejora en el proceso de razonamiento para el diagnóstico de enfermedades crónicas como afecciones alérgicas, cuyo núcleo está basado en datos reales recogidos de una investigación de carácter epidemiológico. Estos datos reales recogidos en encuestas realizadas a pacientes contando con sus correspondientes diagnósticos médicos han sido usados para construir unas reglas capaces de describir el estado de salud a partir de unos síntomas.

Ésta herramienta toma su estructura de los sistemas expertos y el modelado inteligente, capaces de simular problemas similares a la realidad basándose en el conocimiento integrado. Así, a partir de datos médicos, se han estudiado sus características y se han realizado análisis mediante herramientas estadísticas (SPSS y Statistica) con el objetivo de encontrar unas reglas capaces de gobernar un modelo de la actividad real, extrayendo el conocimiento de los datos recogidos, para dotar al sistema de capacidad de razonamiento y decisión.

La herramienta simula el entorno real en cuanto a salud se refiere, basándose en reglas construidas para dar consejo a posibles usuarios de un sistema que, por medio de la web, pueden conocer si es necesario acudir a la consulta de un especialista por diferentes razones conectadas con síntomas (datos de entrada en el sistema experto) o su estado de salud.

English:

Knowledge has its base in information, but also abilities of reasoning, deduction, intuition and experience are needed to achieve it. Data mining is a set of methods designed to extraction of machine-processable knowledge implicit in data bases, with the aim of finding solutions to problems of prediction and classification.

This report has supposed an immersion in data mining world with the target of building an advice tool to manage information in order to improve the reasoning process in diagnosis of chronic illnesses such as allergic diseases based on outcomes from epidemiologic research. Real data collected from surveys and medical checks is used to build rules describing health status reasoning from some symptoms.

This tool uses the knowledge structure extracted from medical data by an expert system and intelligent modeling capable of responding to similar problems based on integrated knowledge rules. This way, from real medical data, statistical features and discriminant classification have been undertaken by statistical packages (statistica and Statistica) with the aim of finding some rules able to govern the model of real activity. Therefore, extracting knowledge from real data acquired by surveys we have been able to build a system with capacity of reasoning and taking decisions.

The tool simulates health outcome reality on the basis of collected specific rules to give an advice to an user tested by web-based survey, telling him if necessary to go and see the doctor due to different medical reasons connected with symptoms (as input data for expert system) or his health status.

1. Introduction

The real processes can be described by the data and rules taken from knowledge structure constructed from results of experimental measurements and checks. It is essential to gain knowledge as much as possible from the data based on parameters of real activity to construct model as errorless as acceptable to simulate such processes in order to control and anticipate future reaction with stimulated condition.

Modeling and simulation of real biological processes can approach governing some parts of our reality and are especially interested in the case of our healthcare and especially taking care of chronic illness. In control and simulation of real reactions the data management is a crucial issue, however the data and their management cause practical activity (medical, social) and the ways of its extent. The raw data from experiments can be organized in the form of datasets (databases) with mined knowledge structure with indirect and direct rules and procedures.

In fact, the knowledge describing the health status are useful in assessment of the status of a patient in the process of medical diagnosis made by medical doctor. But the structure of the health status knowledge could be known, even by direct rules as in data mining methods or only in the form indirect data boundary estimations by statistical and classification methods. And in the case of healthcare the automatic categorization of the risk of a disease on a basis of information on the symptoms and reactions can be done.

The results of surveys based on answers to specially asked questions about the symptoms and reactions, confirmed by some medical checks and final doctor diagnosis, may describe the characteristic features of diseases or health status in the form of categories or labels (nominal variable names with medical meaning and full understanding) resulting in the doctor reactions known as medical procedures or other activities of the process of taking health care. Such categories in the final outcomes of knowledge structure can be used in constructing advising tool similar to but less serious than system known as expert systems. Practically such advising aid can tell if the user state is (or could be) serious and he should go and see the doctor or he has not any reason to do this at the moment, or he may go to the doctor. This is done in similar way as process of doctor's diagnosis (not only as the instant decision but during some time necessary to collect required data to decide) but at lower level of the knowledge and confidence due to the lack of all medical checks conforming the doctor's decision.

My report is devoted to analysis of the health related data taken from the epidemiological survey (with some medical checks)on one of the chronic and important diseases of allergy and asthma and correlated symptoms, extending doctor's reasoning and confirming the diagnosis.

Result of various studies show that asthma is usually underdiagnosed especially in developing countries, because of limitations on the access to medical specialists and laboratory facilities. [1]

Epidemiological surveys have shown asthma-like symptoms to be far more prevalent than physician diagnosed asthma, and under diagnosis of asthma has repeatedly been suspected during the past two decades, especially in children and young adults. Screening studies that used a combination of symptoms and objective indicators of asthma have confirmed this view.

Besides, there has been a huge increase in the global number of people suffering from Asthma over the last 40 years, particularly in children. Approximately 300 million people worldwide currently have asthma, and its frequency increases by 50 per cent every decade.

From this lack of information and the importance of asthma being diagnosed at an early stage so treatment can be undergone to avoid hard consequences, grows up the need of an intelligent system being able to efficiently give a diagnosis.

This project is part of a research scientific experiment of using the rules taken from data mining to achieve a knowledge structure of medical and life science origin, with the aim of building a model of expert system equipped with real medical knowledge.

2. Main goals

The aim of this project is to manage the information for an expert knowledge for the use in advising tool. The knowledge describing the health status can be useful in automatic categorization of the risk of a disease on a basis of answers to specially tailored questions about the symptoms and reactions. The categories in the final outcome of such advising tool can assess if the user state is (could be) serious and he should go and see the doctor or he has not any reason to do this, or he may better go to the doctor for further diagnosis. This is done in similar process as doctor's diagnosis but at lower level of the knowledge and confidence due to the lack of medical checks and full doctor's medical knowledge conforming the medical decision as a classical doctor's diagnosis.

Targets of the project to be considered are:

I. data preparation for database engine resulted from analysis of ECAP data
   a. preparing important issues for decision (advice) of the disease (allergy, asthma)
   b. choosing problems and exact question from the ECAP questionnaire questions
   c. estimation of conditional distributions for the answers from the ECAP outcomes database
   d. preliminary assessment of answer relations with final outcome variables resulting doctor's diagnosis (confirmed by medical check-ups)

II. model obtained from discriminative procedure by statistical package SPSS

III. model obtained from decision tree using calculations made by statistical package Statistica, some analyses of choosing answers and comments on the assessments of accuracy for specific cases .

3. Theoretical introductory knowledge

3.1 Knowledge structure

This project has as target the achievement of a knowledge structure constructed from results of experimental measurements and checks that are input data for such consideration. This knowledge will be represented as rules, and acquired data will be the origin of these rules. Real data, provided by a survey, counting on diagnosed results of real patients will allow us carry on this process. As far as possible the strict medical knowledge is not necessary but for some reasons the application of medical ontology, reasoning rules or real practical experience can give an advantage of expert view on outcomes reserved from modeling and real assessment of the parameter values.

Knowledge has a far more complex nature than simple data and information and requires the active contribution of people to manage knowledge systems. Therefore, for proper knowledge models implementation it is essential to clarify at an early stage, the main differences between data, information and knowledge. [2]

The relationship between data, information, knowledge and wisdom form a pyramid. The pyramid has data as its base, followed in the hierarchy by information, then knowledge, with wisdom at the top. Figure 1 below shows the relationships between data, information knowledge and wisdom.

Figure 1: Relations between data, information, knowledge and wisdom.

Data: a set of discrete objective facts about an event or a process which have little use by themselves unless converted into information. Data for example are numerical quantities or other attributes derived from observation, experiment, or calculation.

Information: data provided with relevance and purpose. It has meaning and it is organized for some purpose. Information for example, is a collection of data and associated explanations, interpretations, and other textual material concerning a particular object, event, or process.

Data could be converted into information using 5 main processes:

- Condensation – items of data are summarized into a more concise form and unnecessary depth is eliminated
- Contextualization –the purpose or reason for collecting the data in the first place is known or understood;
- Calculation - data is processed and aggregated in order to provide useful information
- Categorization – is a process for assigning a type or category to data;
- Correction – is a process for removal of errors.

Knowledge: a fluid mix of framed experience, values, contextual information, expert insight and solid intuition that provides an environment for evaluating and incorporating new experiences and information. It originates and is applied in the minds of people. Knowledge represents a state or potential for action and decisions in a person, organization or an intelligent system, in our case. It could be changed in the process of learning which causes changes in understanding, decision or action.

Wisdom: the ability to identify truth and make correct judgments on the bases of previous knowledge, experience and insight. Within an organization, intellectual capital or organizational wisdom is the application of collective knowledge.

The whole process can be resumed in Figure 2: data from the survey will allow us extract useful information. This information will undergo a process of data mining, in order to get a knowledge structure that will supply the appropriate logical rules to be able to give efficient diagnosis, even at the level of advisement. This way, from patient's symptoms and environmental information, and counting on real diagnosed results given by certain doctors, we will be able to extract knowledge enough to build a model of an intelligent system, so we hope the latter basing on the strong number of analyzed cases (person or patients) according probabilistic rules..

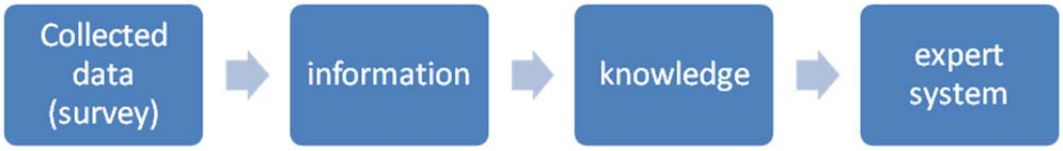


Figure 2: Data transformation from raw data to information (rules or probabilities) useful into expert systems.

The other project made by the other member of the research group student Hugo García Gonzalez connected with usage of the rules and other functional outcomes from this project made by the author of this report, student Alicia López Dot, going on with the experiment will be able to, from this knowledge structure (during so called learning phase), from this rules, perform a second phase (so called the testing phase in classification) in which the system will be set, and the data taken from new users as inputs will test the system and will act as feedback for the system to keep on learning. The whole process for the purpose of this project can be resumed in Figure 3.

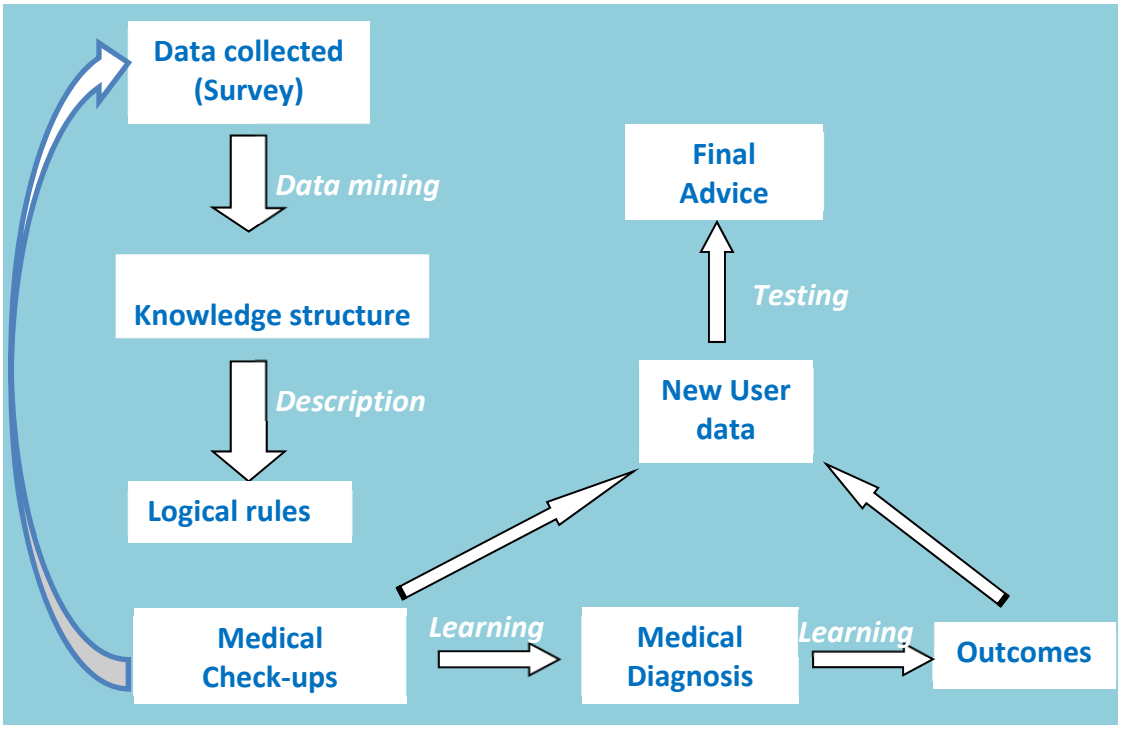


Figure 3: Structure of the experiment

Environment and symptoms conditions are input data provided by the survey, and this way we will satisfy our need of data understanding in health aspect, been able to find a model of relations (relational dependence of variables) in data management process and some simulation of outcomes for tested new answers from survey on line as an aid to patients and to the public.

3.2 Data origin

The base of the knowledge structure and of this whole two connected projects, is the data acquired from a survey called ECAP (Epidemiology of Allergic Disease in Poland), which has its origins in the ECRH surveys, European community of respiratory health survey. These two significant projects seem to this report important sources of information in order to find and model knowledge.

The European community of respiratory health survey, ECRHS, is an international study initially set up in the 1990's to evaluate the prevalence of asthma and allergy in young to middle aged adults. It was funded by the European Commission as part of their Quality of Life Program. Many research groups have followed up in the study in 2000-2002 (ECRHS II). ECRHS II was a survey handled for nine years, undertaken by more than 10,000 young adults that began in 1998. It was a collaborative study and aimed to collect data from 29 centers in 14 countries (mostly European). [3]

In ECRHS II, in 29 centers, individuals who took part in the clinical stage of ECRHS I were sent a short questionnaire and those who responded were invited to a local center, situated in an lung function laboratory in a local hospital or center.

Besides, environmental information was collected by home visits in a subsample of homes, and past and current exposure to air pollution was evaluated through recovery of air pollution records and by a program of air pollution monitoring.

The goals of ECRHS II were:

1. To determine the incidence and prognosis of allergy and allergic disease (asthma, and eczema) in adults.

2. To describe the distribution of exposure to environmental risk factors associated with the incidence and prognosis of allergy and allergic disease.

3. To determine the risk attributable to chronic exposure to these environmental risk factors for the incidence and prognosis of allergy and allergic disease.

4. To identify subgroups within the population based on gender, diagnosed diseases, bronchial response and genetic risk that may be more susceptible to these environmental risk factors.

5. To establish a bank of blood samples ready for DNA extraction taken from representative samples of the population.

The other necessary medical information from data point of view that takes part of the basis of this work is the project called ECAP, Epidemiology of Allergic Disease in Poland. ECAP is the continuation of the European study European Community Respiratory Health Survey II (ECRHS II).

In the preparation of ECAP, the protocol and methodology of the International Study of Asthma and Allergy in Childhood (ISAAC) was also used.

The International Study of Asthma and Allergies in Childhood, ISAAC, is a unique worldwide epidemiological research program established in 1991 to investigate asthma, rhinitis and eczema in children due to the concern that these conditions were increasing in western and developing countries. ISAAC has become the largest worldwide collaborative research project ever undertaken, involving more than 100 countries and nearly 2 million children and its aim is to develop environmental measures and disease monitoring in order to form the basis for future interventions to reduce the burden of allergic and non-allergic diseases, especially in children in developing countries. [4]

The ECAP project includes adults 20-44 years old (ECRHS standard) and children 6-7 and 13-14 years old (ISAAC standard) living in the eight biggest polish urban centers as well as one rural area. The target of the study was to survey 22,500 people in order to measure to what level allergy and asthma affect the population. It was decided that 30% of those individuals who had been studied through the questionnaire should undergo further standard evaluative studies to determine the presence of allergy and asthma. [5]

The project had two basic phases:

- Phase I: Questionnaire (22,500 participants)
- Phase II: further studies with medical checks and final doctor's diagnosis as categorical classification in data mining meaning and understanding (30% of those who completed the ECAP questionnaire interview)

The basic goals giving us data and useful general healthcare reasons are the following :

- Describe the frequency of allergic disease and the response to the most common allergens. The methodology used is similar to that used around the world so that international comparisons are able to be made.
- Describe living conditions of the participants, taking into account, among other factors, type of housing and setting, humidity, ventilation, and contact with animals.

- Describe patients' age, sex, socioeconomic status, and family/genetically inherited risk for allergic disease.
- Evaluate the availability of specialized care for patients suffering from allergies and the quality of the pharmacological care they receive.

Taking in account the data acquired from the survey that about 22500 people answered and, particularly, the data from the second part which in addition counts with diagnosis from specialized doctors, we will be able to build some rules. These rules will take part of the core of the knowledge base of our expert system and will allow it to get efficient predictions. However, the accuracy of categorization in classification in the phase of testing are limited to the accuracy the final outcomes confirmed in the survey by compliance of the survey procedures regarding exact rules of diagnosis made by reviewing doctors. Such effect can results in lower bounds for true positive and negative categorization.

It is needed to underline that as data comes directly from ECAP survey, we will treat the answers to the questions done in the ECAP survey as variables (with the same numbering). As regards output we use general simplification of transformation some output variables (medical diagnosis confirmed by the doctor) in to new wider meaning, but not decreasing the specific knowledge structure. As the number of output variables decreases about 20 specific diagnosed states we will consider some projections of the outputs in the four outcomes in the form of categorical variables (classes as present/absent) as below:

- Healthy (H)
- Rhinitis (Rh)
- Allergy (Al)
- Asthma (As)

The medical meaning of the above outcome categories is correct based on the evidence-based medicine (EBM), however we are interested in assessment of the relation between input and output variables from knowledge structure point of view (data mining understanding) and purposes of classification results.

3.3 Expert system

The final goal of the whole experiment is the design of an expert system that aims to provide the patient for diagnosis (in the classification meaning not exactly in doctor's diagnosis by medical expert) of these illnesses.

An *expert system* is a computing system that is able to express and reason about some domain of knowledge. The purpose of the expert system is to be able to solve problems or offer advice in that domain of the structured knowledge.

Expert systems can be distinguished from other kinds of programs because:

- They deal with subjects of considerable complexity, almost of full internal relations.
- These subjects normally require a considerable quantity of human knowledge and experience.
- They provide expert-level solutions to complex problems.
- They must be fast and reliable;
- They must be capable of explaining and justifying solutions
- They must be understandable and easy to work with.
- They must be sufficiently flexible so that new information may be easily accommodated.

Structure

The structure of a common expert system is showed in Figure 4. [6]

The *knowledge base* stores all relevant information, data, rules (conditional statement that links given conditions to actions), and relationships used by the expert system. A knowledge base can combine the knowledge of multiple human experts. Our knowledge base is going to have its nucleus in some data original from a survey or other confirmed source we will deal with later on.

The purpose of the *inference engine* is to seek information and relationships from the knowledge base and provide answers, predictions, and suggestions in the way a human expert would. The inference engine must find the right facts, interpretations, and rules and assemble them correctly.

The *explanation facility* allows a user to understand how the expert system arrived at certain results.

The purpose of the *knowledge acquisition facility* is to provide efficient means for capturing and storing all components of the knowledge base.

The purpose of the *user interface* is to make the use of the expert system easy for developers, users, and administrators.

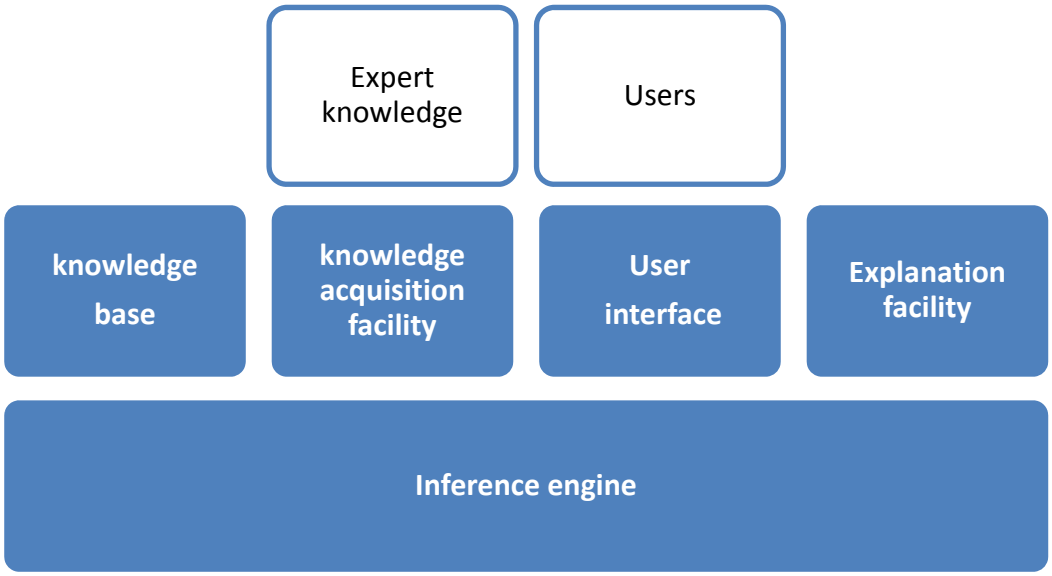


Figure 4: Architecture of a simple expert system from functional point of view.

Inference Engine in Rule-Based Systems

A classical so called rule-based system consists of *if-then rules*, a bunch of *facts*, and an *interpreter* controlling the application of the rules, given the facts.

These *if-then* rule statements are used to formulate the conditional statements that constitute the complete knowledge base. A single *if-then* rule assumes the form 'if *x* is *A* then *y* is *B*' and the if-part of the rule '*x* is *A*' is called the *antecedent* or *premise*, while the then-part of the rule '*y* is *B*' is called the *consequent* or *conclusion*. There are two general kinds of inference engines used in rule-based systems: *forward chaining* and *backward chaining* systems. In a *forward chaining* system, the initial facts are processed first, and keep using the rules to draw new conclusions given those facts. In a *backward chaining* system, the hypothesis (or solution/goal) we are trying to reach is processed first, and keep looking for rules that would allow concluding that hypothesis. As the processing progresses, new sub-goals are also set for validation. Forward chaining systems are primarily data-driven, while backward chaining systems are goal-driven.

Conceptual Model

Knowledge in an expert system can be represented in several ways such as production rules (classical manner), semantic networks, frames, data mining outcomes etc. Methods with natural graphical visualization as different tree classification algorithms can be used to represent knowledge and classify objects if the problem domain is hierarchical in nature, like in medical data structures. In this type of domain higher level alternatives at the top nodes are examined first and then a narrowing process begins until an answer is found. A decision tree could be both a knowledge representation and a reasoning method. Both the reasons help reader the deeper understanding and intuitive imagination for information mined from the data of medical origin.

Web-based Expert System

From the meaning of the classical expert system, our system will differ by the use of the Internet as the communication channel between the User Interface and the other modules of the Expert System architecture and as a dissemination media for healthcare outcomes and other results. The remote topology and structure e-media engaged in such projects (server, platforms, web-based interface between nodes of the system) needs Internet and Communication Technology (ITC) as an integral part at the technical and data management layers. The functional integration of the different modules is shown in Figure 5.

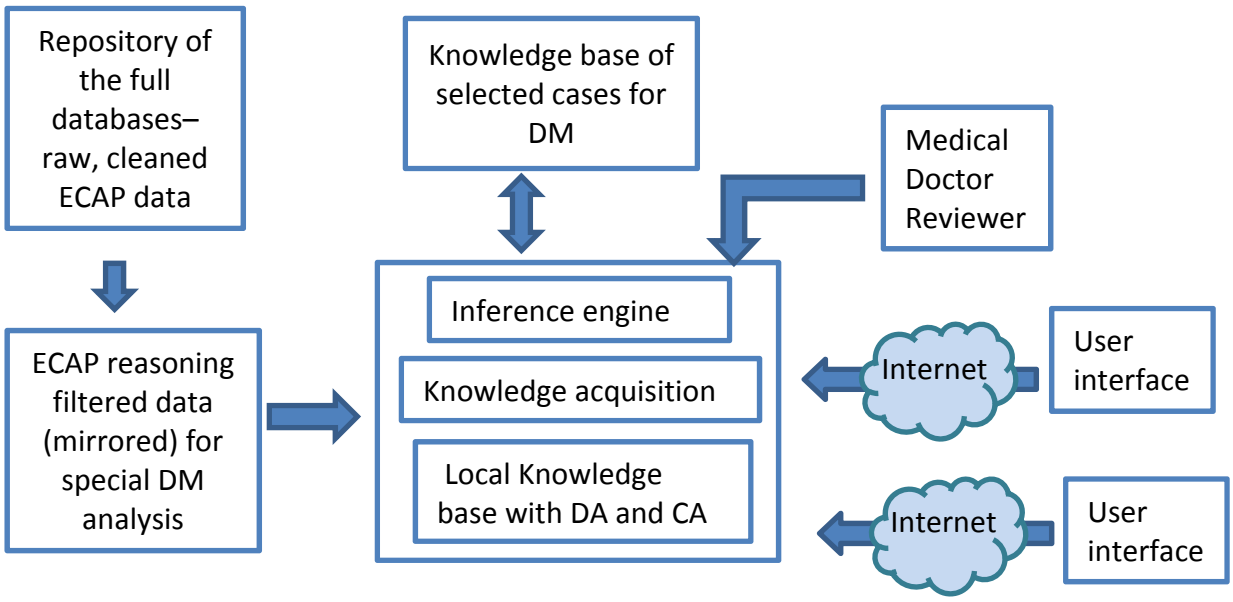


Figure 5: Web-based Expert System Model

DA: Discrimination analysis based on nominal dichotomous variables (2-state variables)
CA: Classification analysis
DM: data mining based on full set of cases with survey output and medical checks results

3.4 Practical cases

Basic statistic parameters

Preliminary data study will be carried out regarding three statistic parameters: conditional probabilities between the variables and the outcomes (diagnosed results), correlation of data and odd ratio. There are typical epidemiologic assessment measures applicable to relation and reasoning analyses, prevalence and other healthcare parameters.

- ***Conditional probabilities***

In probability theory, the conditional probability of an event A conditioned to B, is the probability of *A* happening if *B* has already happened. This way, we are interested in the probability of the outcomes (healthy, asthma, rhinitis or allergy) considering the value of the variables.

Regarding the case in which v_x can only be given value 0 or 1 (dichotomous classes of absent/present, is / is, not/yes, etc.) we will define conditional probability as following:

$$P(H|vx(0)) = \frac{P(H \cap vx(0))}{P(vx(0))} \qquad P(H|vx(1)) = \frac{P(H \cap vx(1))}{P(vx(1))}$$

Conditional probabilities approaching value 1 indicates a high relation between the variable and the outcome, so we can consider that variable as a good descriptive of the healthy/asthma/allergy/rhinitis state.

- ***Correlation***

Correlation is an empirical indicator of possible relationships between variables. In probability and statistics, correlation indicates the strength and direction of a linear relationship between two variables. Two quantitative variables are considered correlated when one of their values change systematically respect the homonym other

one's values: if A and B are variables, correlation exists if when A values rise, B's do it too.

Plenty of coefficients measure correlation grade in classical statistics, adapted to the origin of the treated data. The most known is the Pearson's correlation coefficient for numerical variables, and it is obtained by dividing the covariance of the two variables and the product of their standard deviation:

$$\mathbf{r} = \frac{\sum \mathbf{xy} - \mathbf{N}\bar{\mathbf{x}}\bar{\mathbf{y}}}{\sqrt{(\sum \mathbf{x^2} - \mathbf{N}\bar{\mathbf{x}}^{\mathbf{2}})(\sum \mathbf{y^2} - \mathbf{N}\bar{\mathbf{y}}^{\mathbf{2}})}}$$

Where $\bar{x}$ and $\bar{y}$ are means and N the number of samples of the variables. We will use this coefficient to calculate correlation in non-nominal numerical variables, but not in binary variables (those whose values are different from 0 and 1).

The kind of correlation that is applied to two binary variables is the phi correlation. It is in fact just a simplification of the Pearson's coefficient. We can show it considering Figure 6:

Vx / Vy	0	1	total
0	A	B	r1=A+B
1	C	D	r2=C+D
total	c1=A+C	c2=B+D	A+B+C+D

Figure 6: Table of calculations

Considering the values in the figure, we can calculate the Phi value this way:

$$\varphi = \frac{AD - BC}{\sqrt{r_1 r_2 c_1 c_2}}$$

This way we can calculate correlation for all our variables, those binary and those with more than two possible values. Taking account that correlation coefficient ranges from −1 to 1, we will interpret correlation this way:

- A value of 1 implies that a linear equation describes the relationship between *X* and *Y* perfectly, with all data points lying on a line for which *Y* increases as *X* increases.
- A value of −1 implies that all data points lie on a line for which *Y* decreases as *X* increases.
- A value of 0 implies that there is no linear correlation between the variables.

In our case, X will be one of the numeric variables (v_x) and Y will be one of the outputs (H, Rh, As, All). This way we will be able to choose variables to build our rules as those with highest correlation.

- ***odd ratio***

Odd ratio is a way of comparing whether the probability of a certain event is the same for two groups. It is a statistical parameter that measures effect size, describing the strength of association or non-independence between two binary data values. The odds ratio is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. The odds of success are defined as the ratio of the probability of success over the probability of failure.

Considering figure 5, the expression for odds ratio:

$$OR = \frac{A/C}{B/D}$$

Descriptively, odds ratio is a way of showing the proportion of times of an event happening opposite to not happening. This way, OR=2.5 must be read as 2.5:1, an events occurs in presence of another variable 2.5 more times than it does when the variable is not present. OR=1 means that the number of times that the event occurs is not related with the presence of the variable.

Discriminant analysis

Discriminant analysis is a statistic technique capable of determine which variables can make the difference between groups and how many variables are needed to achieve the best classification possible. Membership to the groups is used as a dependent variable (a categorical variable with so many discrete values as groups). Variables supposed to make the difference between groups are used as independent variables or classification variables (also known as discriminant variables).

The main target of the discriminant analysis is to find a linear combination of the independent variables that allows making the difference between groups (discriminate) the best. It is a multivariate analysis technique that makes possible to take profit of the existing relationships between a large number of independent variables to maximize the ability of discrimination.

Discriminant analysis can tell the difference between any number of groups, but we will explain it with two cases in order to better understand. Figure 7 shows the bivariate space defined by variables X1 and X2, cluster of dots corresponding to two hypothetic groups. We can also notice function D, a linear combination of both variables. Upon D function, projection of both clusters of dots is represented in a histogram way, as if D cut both clusters in its axe direction. Dot lines in each histogram show the projected location of mean points of each group (centroids).

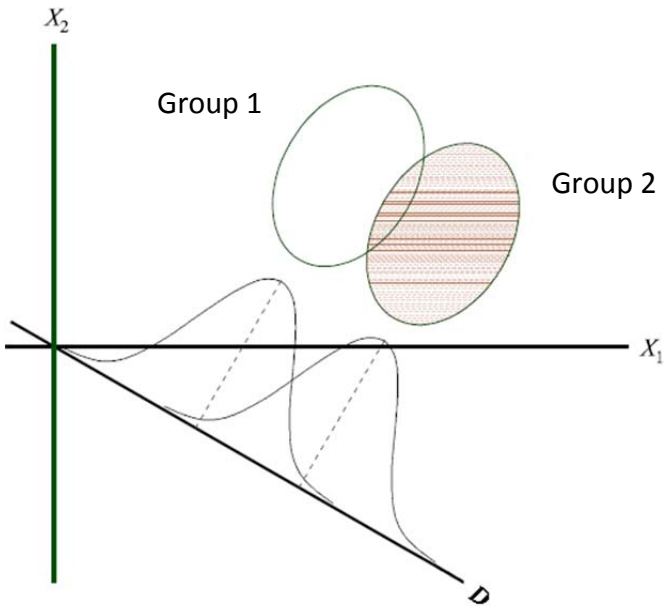


Figure 7: Space defined by two variables

Discriminant analysis purpose is to take profit of information contained in the independent variables in order to build a D function, as a linear combination of X1 and X2, capable of differentiating between the groups as possible as it can be.

Discriminant function has such a shape:

$$D = b_1 X_1 + b_2 X_2$$

Where b_1 and b_2 are the adjustments of the independent variables that make the punctuations of subjects of one of the groups highest and those of the other groups lowest.

Found discriminant function D, it is nonsense trying to represent situation of the groups in space defined by X1 and X2 variables. Representation of D function in p dimensions when p>2 is rather complicated, but in this case is really useful: groups appear represented by their histograms and centroids are shown as dotted lines, as we can see in Figure 8.

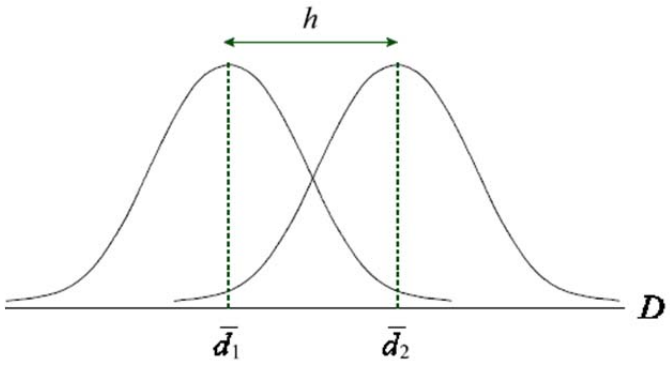


Figure 8: Histogram of each group characterized by Gauss distribution and centroids represented over discriminant function

Replacing in discriminant function the value of means of group one in X1 and X2 variables we can obtain centroid of group 1. Same manner, replacing means of group 2, we obtain centroid of group 2.

$$\bar{d}_1 = b_1 \bar{x}_1^{(1)} + b_2 \bar{x}_2^{(1)}$$

$$\bar{d}_2 = b_1 \bar{x}_1^{(2)} + b_2 \bar{x}_2^{(2)}$$

D function must be build considering distance h between both centroids should be maximum, so that we get groups as distanced as possible. Distance h can be given this way:

$$h = \bar{d}_1 - \bar{d}_2$$

As shown in figure 8, the goal is decreasing dimensionality of p independent variables into one only dimension (corresponding to linear combination D) in which groups differ as much as possible. Punctuation of subjects in this new dimension (named as discriminant coefficients) will allow us classify subjects.

It is important to remark that groups should be different in advance in independent variables. Discriminant analysis seeks mark the difference between groups combining independent variables but if groups don't differ respect to independent variables, analysis will be unsuccessful: we won't find a dimension in which groups differ.

In other words, if overlapping between two groups' cases is overloaded, centroids will be not located in the same or similar place in p-dimensional space, and under those conditions it won't be possible to find a useful discrimination function to classify. It is to say, if centroids are very close, means of the groups in discriminant function will be so similar (d will be so small) that it won't be possible to distinguish subjects from one and another group.

As a conclusion of this part for our purposes for data mining, building discriminant function D for every one of the state of illnesses (however Healthy is inverse in comparison to Sick state), we will have a simple criteria to know the belonging or not of a new user to a group. Knowing the distances to the centroids will also help us in this target and Mahalanobis method will be the chosen to calculate the distance as similarity and membership function for inclusion of multivariate effect size of the groups (classes). For these reasons the measure of the Mahalanobis distance is widely used in cluster analysis and classification techniques and it used for multivariate statistical testing and Fisher's Linear Discriminant Analysis for such case as ours being a supervised classification.

Decision trees

A decision tree is composed of nodes, leaves and branches. Inside the nodes some decisions occurs that transfer control via any of its branches to other nodes or leaves. Typically, these trees are used to implement knowledge systems.

The approach is used of the nodes *decision nodes* and the leaves *answer nodes*. A decision node has associated with it a decision question. An answer node has associated with it an answer. All the nodes can be identified with labels in order to keep track of tree traversals and ease the implementation. A binary tree (nominal values of yes/no) with these characteristics has the general structure as shown in the Figure 9 below.

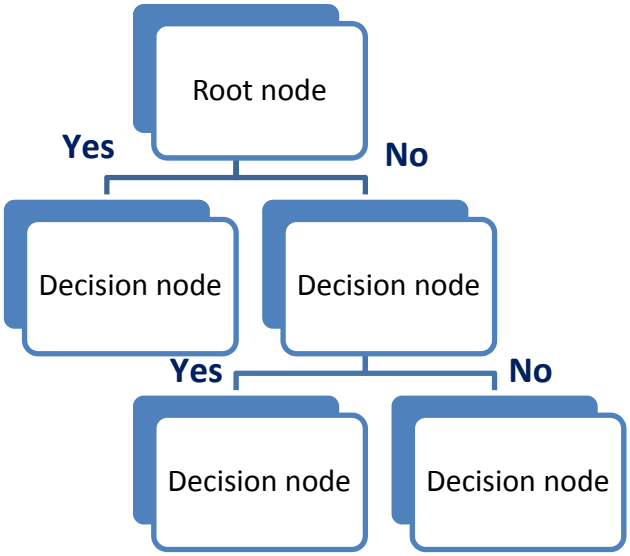


Figure 9:Part of structure of decision trees with two states of decision (yes/no).

The inference process in decision trees starts by setting an initial location, typically the root node, and following the yes/no transitions until we arrive at an answer node. If the current location is a decision node then the question associated with that node is presented. If the question is answered with a yes then the current location is set to the child node associated with the yes branch. If this location is a decision node then the process is repeated. If it is an answer node then the process comes to an end and the answer is presented. To begin the process again the current location is set to the root node.

4. Results

To discuss procedure results, will consider our specially designed 4 output variables mapped from 29 original diagnosis variables of ECAP database. The reason of decreasing the number of output diagnosis classes is a result of our consideration to demonstrate effective rule-based reasoning and assesses the prevalence of the most common diseases applied to advising tool for web-based test. For the convenience in presenting the results the following abbreviations for output variables and for other input variables:

H =healthy
P =probability
As =asthma
Rh =rhinitis
All =allergy

$V_x(y)$ =Variable number X with value Y

The labels of the input variables $V_x(y)$ in classification will be used for final description of the results to study the context from the side of healthcare specialist and for sociological purposes.

4.1 Description of files of data

In general ECAP survey provides us with 18627 patients' data, data acquired from survey undertaken by both adults and children. In these results, adults and children collaborate 50-50%, and considering children, half of them belong to 6/7 years group and half of them to 13/14 years group.

Besides, even if we count on 18627 patient's answers, only 4783 of them own a doctor's diagnosis, so we will be limited to that number of samples for, through data mining, find the rules governing our knowledge system. From that data, we will select adults' answers to analyze discrimination as children may present differences respecting to symptoms. Considering total area corresponding to 18627, the total number of patients who answered the survey, the description of data is shown in a graphic mode in Figure 10 to clarify dependence of the data partitioning and active part for different calculations and modeling in classification stages :

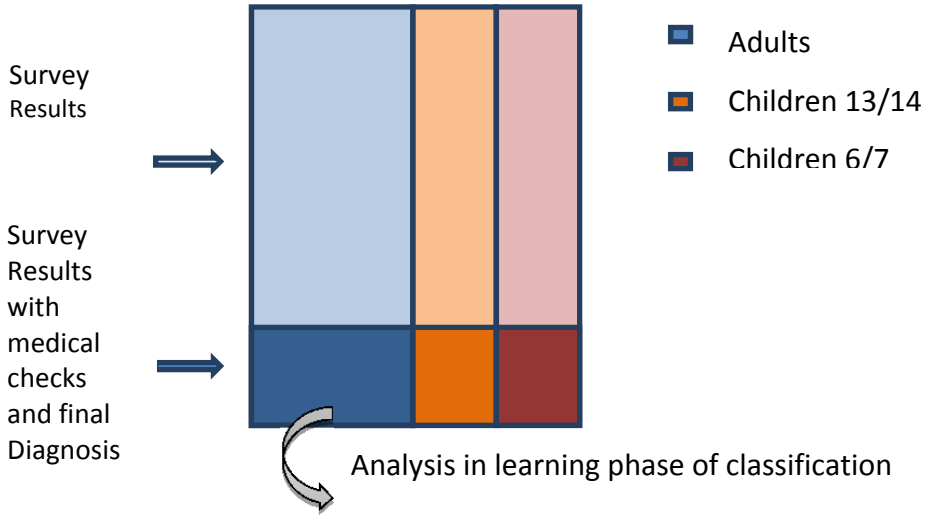


Figure 10: Full Data set description used in project

4.2 Logical analysis of data

In this section we will describe how data has been selected and checked, considering its origin comes from a real survey undertaking by patients and diagnosis has been made by different specialists – medical and data analysts.

The survey undertaken by the ECAP pursued numerous goals as we already commented in the point 2. However our purpose virtually focuses in the first of them, "Describe the frequency of allergic disease and the response to the most common allergens and their relations", that's why not every answer to question appearing in the ECAP survey is useful in our case. This way, before the statistic treatment of the data, shows up the need of having this questions filtered so we will consider just those questions giving us useful information about the state of illnesses or the strength of it. These selected questions will be later treated as variables that we will study deeper in order to find the most suitable ones for the engine of our expert system that gives us the strongest impact or dependence.

Filtering was made in a traditional way, selecting those questions in which key words appeared, such as sneezing, wheezing, tightness or cough. A list of the pre-sectioned questions/variables in shown in Appendix 1.

Once useless variables are filterer, is needed to check data consistency. Given that data origin is human, real, the probability of coherence mistakes shown up could be high. In order to avoid it and gain the true rules governing relation between variables ,

data “manual” reviewing and checking is needed to confirm data internal consistency and nontrivial intelligence.

The logical rules followed to check them seem to be the following:

$$If\ H = 1, then\ \ Rh = 0 \cap All = 0 \cap As = 0$$

$$else\ if\ H = 0, then\ \ Rh = 1 \cup All = 1 \cup As = 1$$

Disposing of a raw database mirrored from repository in the form of standard MsOffice Excel document where patients answers to the questions/variables are shown, with their correspondent diagnosed state (outputs: healthy, asthma, rhinitis, allergy) we can determine some incoherence appear: even if the rate of appearing is low, some diagnosis seem to be somewhat wrong, not respecting the rules above mentioned. The reason why this incoherence show up lies in the real origin of the data, and that can be averted.

4.3 Analysis of basic statistic parameters

Conditional probabilities, correlation, and odd ratio have been calculated as a preliminary study of data. Considering theoretical information given, results are shown below. For the clarity from data mining point only the names of variables will be used instead of full labels – the questions are published in Appendix 1. [7] [8]

Conditional probabilities

Probabilities of outputs (H, Rh, All, As) conditioned to independent variables (V_x) have been calculated for independent variables with two states (0/1) and with a considerable number of answers (as real data, answers to all the questions aren’t available from all the patients). In Figure 11 we can see conditional probabilities for variables with full number of patients.

Figure 11: Conditional probabilities of variables with full number of patients.

Question	number of observations	Healthy	Rhinitis	Asthma	Allergy
v132_002	18617	40,3%	34,4%	13,5%	56,3%
v132_003	18617	34,8%	40,7%	15,6%	62,6%
v136	18617	31,9%	41,4%	27,9%	61,0%
v139	18617	35,3%	40,6%	25,8%	58,9%
v140	18617	34,5%	36,2%	28,8%	58,8%
v141	18617	39,0%	35,1%	22,2%	54,9%

v142	18617	30,3%	37,8%	34,0%	60,1%
v146	18617	42,7%	33,7%	17,4%	52,4%
v147	18617	39,4%	33,8%	19,7%	54,1%
v148	18617	43,1%	32,4%	16,1%	52,4%
v150	18617	41,0%	32,4%	83,5%	53,9%
v151	18617	38,5%	35,5%	19,2%	56,6%
v153	18617	33,9%	39,4%	26,5%	59,2%
v160	18617	15,4%	52,2%	61,0%	68,0%
v176	18617	23,1%	56,4%	20,0%	73,0%
v178	18617	31,1%	47,4%	16,7%	65,5%
v182	18617	37,7%	40,6%	16,1%	59,0%
v191	18617	39,8%	31,4%	12,7%	57,7%
v192	18617	36,2%	30,7%	14,4%	61,5%
v195	18617	27,0%	40,4%	24,7%	68,5%
v199	18617	35,9%	39,1%	10,9%	57,8%
v209	18616	17,0%	53,6%	51,0%	74,5%
v250	18616	54,4%	25,9%	1,9%	44,0%
v256	18616	51,0%	29,6%	9,0%	47,0%
v269	18616	36,5%	32,4%	13,3%	61,5%
v281	18616	50,4%	27,8%	11,4%	46,3%
v304	18617	28,2%	43,7%	35,6%	63,8%
v404	18617	30,8%	42,1%	29,2%	62,2%
v424	18617	28,9%	43,2%	30,5%	64,4%
v429	18617	14,9%	52,6%	62,3%	68,7%
v430	18617	28,4%	44,0%	37,2%	63,1%
v431	18617	30,5%	42,2%	24,1%	63,7%
v432	18617	30,3%	47,4%	17,1%	65,9%
v437	18617	23,9%	55,7%	20,3%	72,2%
v438	18617	32,7%	33,0%	14,6%	64,3%
v444	18617	38,6%	32,0%	12,5%	59,4%
v478	18617	44,1%	31,0%	11,4%	53,7%
v479	18617	41,0%	32,8%	12,0%	55,9%
v480	18617	30,0%	35,0%	14,2%	67,8%
v481	18617	46,4%	30,7%	11,3%	51,6%
v504	18617	28,9%	42,9%	29,5%	64,3%
v507	18617	30,2%	48,3%	17,2%	66,4%
v509	18617	39,5%	38,0%	2,4%	58,5%
v514	18617	46,0%	30,8%	10,6%	51,6%
v515	18617	24,9%	45,0%	15,9%	72,0%
v517	18617	51,4%	27,6%	10,0%	45,9%

In Figure 12, we can see results for variables with lower number of cases (after filtration in sequence of questions in survey) which may not be used in the first approach but can be useful due to their possible discriminative meaning.

Figure 12: conditional probabilities for variables with lower number of patients

Question	Number of observations	Healthy	Rhinitis	Asthma	Allergy
v137	2519	28,0%	38,5%	39,9%	59,6%
v138	2519	25,1%	49,5%	35,1%	65,9%
v149	6457	32,9%	36,1%	24,0%	58,7%
v152	1832	34,5%	37,9%	19,7%	58,1%
v179	6735	29,5%	49,4%	16,9%	66,8%
v180	5550	21,6%	60,7%	19,6%	74,6%
v183	6753	35,2%	43,2%	18,9%	61,8%
v187	6753	29,2%	50,0%	16,9%	67,3%
v193	1643	33,1%	31,9%	15,0%	64,1%
v194	1040	28,8%	34,0%	16,8%	68,0%
v251	3266	53,2%	26,7%	8,6%	45,5%
v254	6250	48,1%	29,2%	12,8%	48,5%
v270	3907	33,4%	33,3%	13,6%	64,7%
v273	9376	48,4%	28,5%	10,8%	48,6%
v275	4402	49,0%	27,3%	11,5%	47,2%
v280	4402	48,4%	28,5%	10,4%	48,4%
v283	4455	54,0%	25,0%	11,2%	40,6%
v385	2534	9,4%	71,9%	50,0%	84,4%
v396	2534	23,3%	43,3%	35,0%	70,0%
v405	3098	26,4%	38,7%	41,5%	63,2%
v409	3096	25,1%	45,9%	38,1%	65,9%
v425	2929	26,9%	45,8%	33,8%	65,5%
v428	1804	24,0%	48,0%	45,0%	68,0%
v433	6367	28,8%	49,2%	17,5%	66,9%
v434	5524	21,6%	58,8%	20,0%	73,8%
v439	2220	29,8%	34,4%	14,0%	66,8%
v440	1565	25,9%	35,3%	15,9%	70,9%
v483_002	9386	54,7%	29,3%	14,7%	42,7%
v483_003	9386	36,9%	30,8%	9,2%	56,9%
v483_004	9386	31,6%	39,5%	7,9%	65,8%
v483_005	9386	26,1%	39,1%	4,3%	73,9%
v483_006	9386	30,6%	33,3%	11,1%	69,4%
v483_007	9386	50,3%	2,1%	9,1%	47,4%
v483_008	9386	44,6%	33,8%	11,3%	52,4%
v505	3160	19,7%	48,3%	46,1%	69,7%
v532	10263	41,9%	33,7%	13,7%	53,5%

v533	3225	41,1%	35,0%	13,5%	54,5%
v550	10263	43,6%	31,3%	12,6%	52,8%
v552_002	2168	41,6%	32,6%	12,1%	55,0%
v552_003	2168	41,0%	33,3%	14,0%	55,9%
v552_004	2168	42,1%	26,3%	5,3%	52,6%
v552_005	2168	17,6%	35,3%	0,0%	76,5%
v552_006	2168	44,4%	33,3%	0,0%	44,4%
v554	10263	33,3%	32,2%	13,9%	64,0%

Conditional probabilities of independent variables should approach value of 100% for a good description of outputs, and most of values seem not to give us high-quality information. Despite that, some values show quite good values, so we can start making an idea of which of them are likely to give us good information for rules. Besides, a look at the table we can see that values are higher for allergy, what means that allergy output is better described by these variables. Variables with higher values (P>65%) are resumed in Figure 13.

Figure 13: Input variables with highest conditional probability for dependent output variables

Variable	Value	Outcome
v138	65,9%	Allergy
v150	83,5%	Asthma
v160	68,0%	Allergy
v176	73,0%	Allergy
v178	65,5%	Allergy
v179	66,8%	Allergy
v180	74,6%	Allergy
v187	67,3%	Allergy
v194	68,0%	Allergy
v195	68,5%	Allergy
v209	74,5%	Allergy
v385	84,4%	Allergy
v396	70,0%	Allergy
v409	65,9%	Allergy
v425	65,5%	Allergy
v428	68,0%	Allergy

Variable	Value	Outcome
v429	68,7%	Allergy
v432	65,9%	Allergy
v433	66,9%	Allergy
v434	73,8%	Allergy
v437	72,2%	Allergy
v439	66,8%	Allergy
v440	70,9%	Allergy
v480	67,8%	Allergy
v483_004	65,8%	Allergy
v483_005	73,9%	Allergy
v483_006	69,4%	Allergy
v505	69,7%	Allergy
v507	66,4%	Allergy
v515	72,0%	Allergy
v552_005	76,5%	Allergy

Correlation

A input independent variable can be considered descriptive when the correlation with variables describing illnesses as outputs is high but even not so small, and the correlation with healthy status should be low or lower than in the case of other symptoms of sickness . But for the discriminative analyses we need to ask not to have internal correlation between independent variables due to lower discriminative power. Correlation values have been calculated by methods given in the theoretical part, for variables with two states (classes).

Results are shown below in Figure 14 for variables with high number of observations.

Figure 14: Correlation of input variables with output variables

Question	Observation	r(v;H)	r(v;Rh)	r(v;As)	r(v;Al)
v132_002	18617	-7,9%	5,6%	3,9%	7,7%
v132_003	18617	-9,8%	9,0%	5,3%	10,2%
v136	18617	-15,0%	12,6%	24,5%	12,1%
v139	18617	-9,2%	8,4%	15,3%	7,4%
v140	18617	-9,2%	5,0%	16,8%	6,7%
v141	18617	-10,8%	7,6%	18,9%	8,0%
v142	18617	-9,1%	4,9%	17,5%	6,0%
v146	18617	-9,4%	7,9%	14,6%	7,2%
v147	18617	-9,7%	5,8%	14,0%	6,6%
v148	18617	-11,1%	7,3%	14,2%	8,6%
v150	18617	-7,1%	3,5%	8,2%	5,3%
v151	18617	-8,0%	5,7%	9,9%	6,8%
v153	18617	-17,8%	13,9%	29,2%	13,7%
v155	18617	1,8%			
v160	18617	-16,4%	13,1%	40,4%	10,3%
v176	18617	-33,6%	39,4%	19,3%	32,9%
v178	18617	-34,4%	38,2%	18,1%	33,3%
v179	6735	-31,5%	36,1%	15,9%	30,6%
v180	5550	-18,3%	25,9%	8,3%	18,9%
v182	18617	-21,7%	24,4%	16,1%	21,4%
v191	18617	-18,7%	6,4%	6,4%	19,7%
v192	18617	-9,6%	1,9%	4,4%	10,1%
v195	18617	-9,7%	5,3%	9,1%	8,5%
v199	18617	-3,0%	2,8%	0,2%	2,6%
v207	18617	1,8%			
v208	18616	-0,4%			
v209	18616	-14,9%	12,4%	25,1%	12,7%

v250	18616	2,9%	-2,4%	-2,8%	-2,8%
v256	18616	0,3%	0,9%	-1,6%	0,0%
v269	18616	-15,9%	5,3%	5,2%	16,8%
v273	9376	-2,3%	-2,8%	3,6%	0,8%
v281	18616	-0,1%	-0,9%	2,1%	-1,0%
v304	18617	-19,7%	15,6%	37,1%	15,4%
v404	18617	-19,4%	15,8%	31,2%	15,7%
v424	18617	-20,7%	16,5%	32,3%	17,3%
v429	18617	-16,5%	13,2%	41,1%	10,6%
v430	18617	-12,9%	10,7%	26,5%	9,8%
v431	18617	-19,8%	15,8%	22,7%	17,1%
v432	18617	-34,2%	36,6%	18,1%	32,6%
v437	18617	-32,1%	37,5%	19,7%	31,1%
v438	18617	-14,2%	4,3%	5,5%	14,4%
v444	18617	-15,3%	5,4%	4,1%	16,4%
v460	18617	0,3%			
v478	18617	-10,8%	5,1%	2,3%	11,5%
v479	18617	-12,3%	6,8%	3,3%	12,1%
v480	18617	-16,3%	6,0%	4,9%	17,0%
v481	18617	-3,3%	2,6%	1,2%	4,4%
v483_002	9386	2,2%	-0,3%	3,3%	-2,1%
v483_003	9386	-3,8%	0,3%	-0,2%	3,1%
v483_004	9386	-4,3%	2,8%	-0,7%	4,8%
v483_005	9386	-4,5%	2,1%	-1,8%	5,4%
v483_006	9386	-4,5%	1,0%	0,7%	5,6%
v483_007	9386	3,6%	-3,7%	-3,4%	-2,9%
v483_008	9386	-3,3%	2,9%	2,1%	3,0%
v504	18617	-21,2%	16,7%	31,9%	17,9%
v507	18617	-22,9%	25,8%	12,7%	22,6%
v509	18617	-15,7%	15,8%	7,6%	16,8%
v514	18617	-2,8%	1,7%	0,0%	2,9%
v515	18617	-10,0%	7,6%	3,6%	10,2%
v532	10263	-12,8%	10,1%	8,1%	12,5%
v550	10263	-7,5%	4,2%	3,9%	8,1%
v554	10263	-18,8%	5,0%	6,0%	20,2%

Correlation values should approach 100% in absolute value, and be negative for healthy outcome. Variables with good enough partial correlation coefficient values calculated for single variable are resumed in Figure 15:

Figure 15:Variables with higher possible correlation to illnesses

	r(v;H) [φ]	r(v;Rh)	r(v;As)	r(v;Al)
v436	-65,8%			
v166	-53,6%			
v167	-51,5%			
v154	-43,5%			
v178	-34,4%	38,2%	18,1%	33,3%
v432	-34,2%	36,6%	18,1%	32,6%
v176	-33,6%	39,4%	19,3%	32,9%
v437	-32,1%	37,5%	19,7%	31,1%
r179	-31,5%	36,1%	15,9%	30,6%
v427	-27,6%			
v507	-22,9%	25,8%	12,7%	22,6%
v182	-21,7%	24,4%	16,1%	21,4%
v443	-21,4%			
v504	-21,2%	16,7%	31,9%	17,9%
v424	-20,7%	16,5%	32,3%	17,3%
v431	-19,8%	15,8%	22,7%	17,1%
v304	-19,7%	15,6%	37,1%	15,4%
v404	-19,4%	15,8%	31,2%	15,7%
v434	-19,3%	23,0%	7,8%	17,6%
v554	-18,8%	5,0%	6,0%	20,2%
v191	-18,7%	6,4%	6,4%	19,7%
v180	-18,3%	25,9%	8,3%	18,9%

Odd ratio

Odd ratio is a statistical parameter that measures effect size, describing the strength of association or non-independence between two binary data values (as for states like present/absent or yes/no). Its value gives us the degree of dependence between independent variables (symptoms) and the disease variables (outputs).

Odd ratio has been calculated for independent variables with two states (0/1), and it's shown in Figure 16:

Figure 16: odd ratio calculations for input variables

Question	Observation	OR(v;S)	OR(V;Rh)	OR(V;As)	OR(V;Al)
v132_002	18617	1,58	1,40	1,39	1,55
v132_003	18617	1,99	1,86	1,66	2,01
v136	18617	2,36	2,03	4,91	1,96

v139	18617	1,96	1,83	3,46	1,69
v140	18617	2,08	1,48	4,05	1,67
v141	18617	1,73	1,50	3,45	1,49
v142	18617	2,44	1,58	5,03	1,75
v146	18617	1,51	1,45	2,56	1,37
v147	18617	1,67	1,38	2,64	1,41
v148	18617	1,59	1,39	2,50	1,42
v150	18617	1,54	1,25	1,95	1,37
v151	18617	1,69	1,46	2,26	1,54
v153	18617	2,37	1,98	6,16	1,91
v160	18617	5,65	2,98	19,33	2,52
v176	18617	5,12	6,28	3,40	4,68
v178	18617	4,26	6,14	3,42	4,02
v182	18617	2,43	3,01	2,94	2,40
v191	18617	2,14	1,33	1,52	2,23
v192	18617	1,89	1,14	1,50	1,93
v195	18617	3,10	1,76	2,95	2,54
v199	18617	1,71	1,65	1,05	1,56
v209	18616	3,72	2,40	5,55	2,72
v250	18616	0,86	0,87	0,78	0,86
v256	18616	0,98	1,07	0,83	1,00
v269	18616	2,12	1,30	1,44	2,19
v273	9376	1,10	0,88	1,28	1,03
v281	18616	1,00	0,96	1,15	0,96
v304	18617	3,06	2,32	9,74	2,29
v404	18617	2,72	2,22	6,92	2,19
v424	18617	3,02	2,33	7,37	2,42
v429	18617	5,88	3,03	20,47	2,60
v430	18617	2,69	2,14	6,74	2,04
v431	18617	2,78	2,23	4,25	2,35
v432	18617	4,27	5,51	3,35	3,93
v437	18617	4,82	5,77	3,50	4,33
v438	18617	2,31	1,30	1,56	2,29
v444	18617	1,97	1,29	1,32	2,06
v478	18617	1,55	1,26	1,16	1,59
v479	18617	1,70	1,38	1,25	1,68
v480	18617	2,69	1,44	1,50	2,73
v481	18617	1,18	1,16	1,11	1,26
v483_002	9386	0,79	0,97	1,67	0,80
v483_003	9386	1,57	1,04	0,97	1,44
v483_004	9386	1,99	1,53	0,81	2,10
v483_005	9386	2,60	1,51	0,43	3,09
v483_006	9386	2,09	1,17	1,19	2,49
v483_007	9386	0,83	0,81	0,75	0,86
v483_008	9386	1,24	1,22	1,24	1,21
v504	18617	3,00	2,32	7,20	2,44
v507	18617	3,03	3,40	2,31	2,91
v509	18617	1,95	2,05	1,64	2,04
v514	18617	1,21	1,14	1,00	1,23
v515	18617	3,09	2,15	1,65	3,03
v532	10263	1,70	1,57	1,69	1,67
v550	10263	1,42	1,24	1,33	1,46
v554	10263	2,51	1,29	1,53	2,65

Output variable noted as S has considered a sick outcome e.g. inverse of H (Healthy): S= H', including rhinitis, allergy and asthma outcomes, and has been calculated by the inverse of healthy odd ratio calculation. Values approaching 1 mean low relation between the independent variable and the outcome, so higher values are source of information. Variables with high values are shown in Figure 17.

Figure 17: variables with highest odd ratio values

	OR(v; S)
r161	5,899
v429	5,882
v160	5,654
v176	5,122
v437	4,816
v432	4,274
v178	4,255
r179	3,940

	OR(v; S)
v385	3,752
v209	3,716
v271_003	3,559
v556_011	3,551
v552_005	3,539
r138	3,140
v195	3,104
v515	3,092
v304	3,056

Selected variables by basic parameters

Final selection of most representative variables considering these basic parameters (correlation with outputs, odd ratio and conditional probabilities) is represented in Figure 18.

Figure 18: variables likely to be selected

	OR	r(v;H)	r(v;Rh)	r(v;As)	r(v;Al)	P(Vx|H)	P(Vx|R)	P(Vx|As)	P(Vx|Al)
v176	5,12	-33,6%	39,4%	19,3%	32,9%	23,1%	56,4%	20,0%	73,0%
v437	4,82	-32,1%	37,5%	19,7%	31,1%	23,9%	55,7%	20,3%	72,2%
v432	4,27	-34,2%	36,6%	18,1%	32,6%	30,3%	47,4%	17,1%	65,9%
v178	4,26	-34,4%	38,2%	18,1%	33,3%	31,1%	47,4%	16,7%	65,5%
r179	3,94	-31,5%	36,1%	15,9%	30,6%	29,5%	49,4%	16,9%	66,8%
v304	3,06	-19,7%	15,6%	37,1%	15,4%	28,2%	43,7%	35,6%	63,8%
v507	3,03	-22,9%	25,8%	12,7%	22,6%	30,2%	48,3%	17,2%	66,4%
v424	3,02	-20,7%	16,5%	32,3%	17,3%	28,9%	43,2%	30,5%	64,4%
v504	3,00	-21,2%	16,7%	31,9%	17,9%	28,9%	42,9%	29,5%	64,3%
v431	2,78	-19,8%	15,8%	22,7%	17,1%	30,5%	42,2%	24,1%	63,7%
v404	2,72	-19,4%	15,8%	31,2%	15,7%	30,8%	42,1%	29,2%	62,2%
v554	2,51	-18,8%	5,0%	6,0%	20,2%	33,3%	32,2%	13,9%	64,0%
v182	2,43	-21,7%	24,4%	16,1%	21,4%	37,7%	40,6%	16,1%	59,0%
v434	2,38	-19,3%	23,0%	7,8%	17,6%	21,6%	58,8%	20,0%	73,8%
v180	2,27	-18,3%	25,9%	8,3%	18,9%	21,6%	60,7%	19,6%	74,6%
v191	2,14	-18,7%	6,4%	6,4%	19,7%	39,8%	31,4%	12,7%	57,7%

4.4 Analysis by discriminative procedure

Discriminant analysis has been made by SPSS software (IBM SPSS Data Collection, v20) - the well-known statistic package for survey authoring, text analytics, statistical analysis, and collaboration and deployment scoring services. [10]

The discriminant model used for our purposes has the following assumptions:

• The predictors (input variables) are not highly correlated with each other (Assumed correlation < 0,75)

• The mean and variance of a given predictor are not correlated.

• The correlation between two predictors is constant across groups.

• The values of each predictor have a normal distribution.

Preliminary exclusion of variables has been made this way: As SPSS package will select as analysis size the number of samples corresponding to the variable with smallest number of observations, variables with **short number of samples** will be considered non-descriptive. This criteria will be done considering as descriptive variables with more than 4641 observations (lack of 3,0% from total 4783). This leads us to consider 54 variables:

v132_002 v132_003 v136 r137 r138 v139 v140 v141 v142 v146 v147 v148 r149 v150 v151 v153 v155 v160 r161 v176 v178 r179 v182 v191 v192 v195 v199 v207 v208 v250 v256 v269 v281 v304 v404 v424 v429 v430 v431 v432 v437 v438 v444 v460 v478 v479 v480 v481 v504 v507 v509 v514 v515 v517

By calculating means of variables values, those **variables with more than two states** have been excluded. Then, following variables are not considered in the analysis:

v155 v207 v208 v460 v517

Considering **correlation among independent variables** and similarity between questions, some of them have been excluded. In Figure 19 we can see cancelled variables due to the logical conflicts coming from repetition of question in different parts of survey or logical usefulness, and in Figure 20 high correlation between some variables.

cancelled	reason
v437	v176
r179	v178
v504	v404
v429	v160
r161	Logical

Figure 19: Cancelled variables due to logical conflicts with others

	v160	r161	v178	r179	v404	v424	v429	v504
v160	1	**0,96**	0,12	0,10	0,35	0,36	**0,96**	0,35
r161	**0,96**	1	0,11	0,09	0,35	0,35	**0,95**	0,35
v178	0,12	0,11	1	**0,84**	0,24	0,23	0,12	0,23
r179	0,10	0,09	**0,84**	1	0,23	0,21	0,10	0,22
v404	0,35	0,35	0,24	0,23	1	**0,76**	0,36	**0,87**
v424	0,36	0,35	0,23	0,21	**0,76**	1	0,36	**0,78**
v429	**0,96**	**0,95**	0,12	0,10	0,36	0,36	1	0,36
v504	0,35	0,35	0,23	0,22	**0,87**	**0,78**	0,36	1

Figure 20: High correlation between variables

In this point, only 44 variables are considered informative: v132_002 v132_003 v136 r137 r138 v139 v140 v141 v142 v146 v147 v148 r149 v150 v151 v153 v160 v176 v178 v182 v191 v192 v195 v199 v250 v256 v269 v281 v304 v404 v424 v430 v431 v432 v438 v444 v478 v479 v480 v481 v507 v509 v514 v515

Healthy

Discriminatory analysis is done to healthy outcome, considering the list of variables given, and taking in account samples of adults only. This way, we will consider adult samples as to undergo the analysis, and rest of samples will be used to test the finally selected variables.

In this first analysis, processing summary is as shown in Figure 21. This table shows a summary including the total number of processed cases, the number of valid cases for analysis and the number of excluded cases, explaining the reason of the exclusion.

Figure 21: Analysis case processing summary for healthy analysis

Analysis Case Processing Summary			
Unweighted Cases		N	Percent
Valid		1811	40
Excluded	Missing or out-of-range group codes	0	0,0
	At least one missing discriminating variable	476	10,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	0,0
	Unselected	2396	50,1
	Total	2872	60,0
Total		4783	100,0

To compare groups, the difference between means must be significant because of its ability to discriminate between different states (classes) of the outputs. The tests of equality of group means measure each independent variable's potential before the model is created. If the significance value is greater than 0.10, the variable probably does not contribute to the mode. Wilks' lambda is another measure of a variable's potential. Smaller values indicate the variable is better at discriminating between groups.

Test of equality of group means provided by SPSS is shown in Figure 22, already organized in increasing values of Wilks' lambda.

Figure 22: Test of equality of group means

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
v178	,881	583,265	1	4305	,000
v432	,884	567,541	1	4305	,000
v176	,890	531,493	1	4305	,000
r179	,898	490,616	1	4305	,000
v437	,900	477,713	1	4305	,000
v507	,949	229,549	1	4305	,000
v504	,954	208,342	1	4305	,000
v182	,956	198,997	1	4305	,000

v424	,958	189,561	1	4305	,000
v431	,959	184,040	1	4305	,000
v404	,961	173,263	1	4305	,000
v304	,962	172,370	1	4305	,000
v191	,965	156,989	1	4305	,000
v153	,969	138,429	1	4305	,000
v429	,973	119,258	1	4305	,000
v160	,974	117,091	1	4305	,000
v269	,974	113,748	1	4305	,000
r161	,975	111,213	1	4305	,000
v444	,976	107,199	1	4305	,000
v509	,977	102,268	1	4305	,000
v480	,978	98,617	1	4305	,000
r138	,979	90,351	1	4305	,000
v438	,980	86,224	1	4305	,000
v136	,981	85,291	1	4305	,000
v430	,984	71,793	1	4305	,000
v479	,984	69,402	1	4305	,000
r149	,988	51,913	1	4305	,000
v148	,988	50,228	1	4305	,000
v478	,989	49,950	1	4305	,000
v141	,989	46,574	1	4305	,000
v132_003	,989	45,838	1	4305	,000
v515	,990	42,132	1	4305	,000
v195	,991	38,308	1	4305	,000
v192	,991	37,353	1	4305	,000
v139	,992	36,355	1	4305	,000
v147	,992	36,062	1	4305	,000
v140	,992	35,684	1	4305	,000
v142	,992	32,921	1	4305	,000
v146	,993	31,837	1	4305	,000
v132_002	,993	28,826	1	4305	,000
r137	,993	28,364	1	4305	,000
v151	,994	25,628	1	4305	,000
v517	,995	19,760	1	4305	,000
v150	,995	19,616	1	4305	,000
v481	,999	4,397	1	4305	,036

The within-groups correlation matrix shows the correlations between the independent variables. Dependence between variables must be low, and high values of cross correlation (value > 0,75) show similarity between different variables. Counting with such a large number of variables, this matrix turns to be quite vast. It won't be included in the document as no wrong values appear.

Box's M tests the assumption of equality of covariance across groups. Test results allow decline the hypothesis of equality of covariance across groups if significance value is less than 0,05, and as Sig.=0,000<0,05, we can conclude that one of the groups is more variable than the other.

Figure 23: Covariance across groups

Test Results		
Box's M		3548,379
F	Approx.	33,681
	df1	105
	df2	57565351,604
	Sig.	0,000

Enter/remove procedure analyzes which variables are useful. Given the results, standardized canonical discriminant function coefficients of chosen variables are shown in Figure 24. The standardized coefficients allow comparison of variables measured on different scales. Coefficients with large absolute values correspond to variables with greater discriminating ability.

Figure 24: Standardized Canonical Discriminant function coefficients for healthy analysis

	Function
	1
v146	-,191
v160	,199
v176	,332
v178	,321
v191	,206
v431	,156
v432	,156
v437	,164
v438	,117
v504	,125
v515	,159

Classification function coefficients show the resulted coefficients to build the discriminant function.

Figure 25: Classification function coefficients for healthy analysis

Classification Function Coefficients		
	v622_Healthy	
	0	1
v146	,430	,873
v160	,671	-,240
v176	,806	,017
v178	1,545	,864
v191	1,528	1,104
v431	,213	-,197
v432	,787	,456
v437	,107	-,280
v438	,402	,038
v504	,431	,115
v515	1,127	,395
(Constant)	-2,182	-1,169

Finally, model validation is shown in figure 26. The classification table shows the practical results of using the discriminant model.

Of the cases used to create a model, 1693 of the 2246 patients' answers who previously defaulted are classified correctly. 1523 of the 2354 non-defaulters are classified correctly. Overall, 69,9% of the cases are classified correctly.

Classifications based upon the cases used to create the model tend to be too "optimistic" in the sense that their classification rate is inflated.

The cross-validated section of the table attempts to correct this by classifying each case while leaving it out from the model calculations; however, this method is generally still more "optimistic" than subset validation.

Subset validation is obtained by classifying past customers who were not used to create the model. These results are shown in the Cases Not Selected selection of the table.

Figure 26: Classification results for Healthy outcome

Classification Results[a,c]					
v622_Healthy			Predicted Group Membership		Total
			0	1	
Original	Count	0	1523	831	2354
		1	553	1693	2246
	%	0	64,7	35,3	100,0
		1	24,6	75,4	100,0
Cross-validated[b]	Count	0	1517	837	2354
		1	554	1692	2246
	%	0	64,4	35,6	100,0
		1	24,7	75,3	100,0
a. 69,9% of original grouped cases correctly classified.					
b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.					
c. 69,8% of cross-validated grouped cases correctly classified.					

As conclusion for healthy state, distance values are given by classification functions below using coefficients resulted from analysis:

Dis_hea0=
=v146 *0.430+v160 *0.671+v176 *0.806+v178 *1.545+v191 *1.528+v431 *0.213+v432 *0.787+v437 *0.107+v438 *0.402+v504 *0.431+v515 *1.127*-2.182

Dis_hea1=
v146*0.873+v160*(-0.240) +v176*0.017 +v178*0.864 +v191*1.104 +v431*(0.197) +v432*0.456+v437*(0.280) + v438*0.038 + v504*0.115 + v515*0.395 -1.169

Discrimination function is built from canonical discrimination function coefficients:

Dis_hea=
v146*(-0.446) +v160*0.916 +v176*0.794 +v178*0.685 +v191*0.426 +v431*0.412 +v432*0.333+v437*0.390 +v438*0.366 +v504*0.318 +v515*0.736 -1.114

This same analysis has been made for rest of outputs (Rhinitis, allergy and asthma), results are shown in the following paragraphs.

Rhinitis

Analysis for rhinitis output has resulted in the following choice of 14 variables, with their correspondent coefficients, as we can see in Figure 27.

Figure 27: Classification function coefficients for Rhinitis outcome

Classification Function Coefficients		
	v622_Rhinitis	
	0	1
v132_003	,38682	,66138
r137	,22678	-,42658
r138	-,06586	,51888
r149	,05064	-,32337
v150	,58094	,28931
r161	,05715	,67073
v176	-,12360	,94234
v178	,83172	1,35456
r179	-,28108	,14099
v182	1,00279	1,19171
v269	,86825	,65994
v432	,55262	1,13893
v437	-,09509	,45909
v515	,57490	1,10546
(Constant)	-,83501	-3,19878

Canonical discriminant function coefficients are shown in figure 28:

Figure 28: Canonical Discriminant Function coefficients

Canonical Discriminant Function Coefficients	
	Function
	1
v132_003	,23534
r137	-,56002
r138	,50121
r149	-,32058
v150	-,24997
r161	,52594
v176	,91368
v178	,44815
r179	,36178
v182	,16193
v269	-,17855
v432	,50256
v437	,47502
v515	,45478
(Constant)	-,96749

Classification results for Rhinitis status are shown in Figure 29. We can conclude the choice of variables for this outcome leads to a 73,8% of cases correctly classified.

Figure 29: Classification results for Rhinitis status

Classification Results					
v622_Rhinitis			Predicted Group Membership		Total
			0	1	
Original	Count	0	2500	846	3346
		1	378	942	1320
	%	0	74,7	25,3	100,0
		1	28,6	71,4	100,0
Cross-validated[b]	Count	0	2498	848	3346
		1	389	931	1320
	%	0	74,7	25,3	100,0
		1	29,5	70,5	100,0
a. 73,8% of original grouped cases correctly classified.					
b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.					
c. 73,5% of cross-validated grouped cases correctly classified.					

As conclusion for Rhinitis state, distance values are given by classification functions below using coefficients resulted from analysis:

Dis_Rhin0 =

= v132_003*0.38682 + r137*0.22678 + r138*(-0.06586) + r149*0.05064 + v150*0.58094 + r161*0.05715 + v176*(-0.12360) + v178*0.83172 + r179*(-0.28108) + v182*1.00279 + v269*0.86825 + v432*0.55262 + v437*(-0.09509) + v515*0.57490 -0.83501

Dis_Rhin1=

=v132_003*0.66138 + r137*(-0.42658) + r138*0.51888 + r149*(-0.32337) + v150*0.28931 + r161*0.67073 + v176*0.94234 + v178*1.35456 + r179*0.14099 + v182*1.19171 + v269*0.65994 + v432*1.13893 + v437*0.45909 + v515*1.10546 -3.19878

Discrimination function for Rhinitis is built from canonical discrimination function coefficients:

Dis_Rhin=

v132_003*0.23534 + r137*(-0.56002) + r138*0.50121 + r149*(-0.32058) + v150*(-0.24997) + r161*0.52594 + v176*0.91368 + v178*0.44815 + r179*0.36178 + v182*0.16193 + v269*(-0.17855) + v432*0.50256 + v437*0.47502 + v515*0.45478 -0.96749

Asthma

Analysis for asthma output has resulted in the following choice of 13 variables, with their correspondent coefficients.

Figure 30: Classification function coefficients for asthma outcome

Classification Function Coefficients		
	v622_Asthma	
	0	1
v140	,26238	,87238
v153	,52352	1,06422
r161	,42483	2,48943
v176	,48982	1,10861
v182	1,00675	1,29296
v199	,17985	-1,26034
v304	-,05990	1,53279
v429	-,62578	2,34696
v430	-,05674	1,29095
v431	,22568	,96340
v444	1,02404	,71114
v504	,06537	,53823
v509	,65221	,22128
(Constant)	-,70388	-5,29986

Canonical discriminant function coefficients are shown below.

Figure 31: Canonical discriminant Function coefficients

Canonical Discriminant Function Coefficients	
	Function
	1
v140	,33110
v153	,29349
r161	1,12064
v176	,33587
v182	,15535
v199	-,78172
v304	,86449
v429	1,61357
v430	,73151
v431	,40043
v444	-,16984
v504	,25666
v509	-,23390
(Constant)	-,58914
Unstandardized coefficients	

Classification results are shown in Figure 32.

Figure 32: Classification results for asthma outcome

Classification Results					
v622_Asthma			Predicted Group Membership		Total
			0	1	
Original	Count	0	3719	432	4151
		1	200	289	489
	%	0	89,6	10,4	100,0
		1	40,9	59,1	100,0
Cross-validated[b]	Count	0	3714	437	4151
		1	204	285	489
	%	0	89,5	10,5	100,0
		1	41,7	58,3	100,0
a. 86,4% of original grouped cases correctly classified.					
b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.					
c. 86,2% of cross-validated grouped cases correctly classified.					

As conclusion for Asthma state, distance values are given by classification functions below using coefficients resulted from analysis:

Dis_Asth0 =
v140 * 0.26238 +v153 * 0.52352 +r161 * 0.42483 +v176 * 0.48982 +v182 * 1.00675 +v199 * 0.17985 +v304 *(-0.05990) +v429 *(-0.62578) +v430 *(-0.05674) +v431 * 0.22568 +v444 * 1.02404 +v504 * 0.06537 +v509 * 0.65221 -0.70388

Dis_Asth1=
v140 *0.87238 + v153 * 1.06422 + r161 * 2.48943 + v176 * 1.10861 + v182 * 1.29296 + v199 *(-1.26034) + v304 * 1.53279 + v429 * 2.34696 + v430 * 1.29095 + v431 *0.96340 + v444 *0.71114 + v504 *0.53823 + v509 *0.22128 - 5.29986

Discrimination function is built from canonical discrimination function coefficients:

Dis_Asth=
v140 *0.33110 + v153 *0.29349 + r161 * 1.12064 + v176 *0.33587 + v182 *0.15535 + v199 * (-0.78172)+ v304 *0.86449 + v429 * 1.61357 + v430 *0.73151 + v431 *0.40043 + v444 *(-0.16984) + v504 *0.25666 + v509 *(-0.23390) -0.58914

Allergy

Analysis for allergy output has resulted in the following choice of 16 variables, with their correspondent coefficients.

Figure 33: Classification function coefficients for allergy outcome

Classification Function Coefficients		
	v622_Allergy	
	0	1
v132_003	,33108	,59042
v146	,75610	,40767
v147	,42463	,17377
v176	,05408	,76556
v178	,65133	1,25188
v191	,97381	1,37539
v269	,35790	,60671
v431	-,40476	-,11640
v432	,25585	,66594
v437	-,15834	,12110
v438	-,24813	-,01648
v478	1,21478	1,40746
v480	-,29084	-,01561
v504	-,02307	,22854
v509	,68447	,84354
v515	,29400	1,08019
(Constant)	-1,35783	-2,68745

Canonical discriminant function coefficients are shown in figure 34:

Figure 34: Canonical discriminant function coefficients

Canonical Discriminant Function Coefficients	
	Function
v132_003	,27090
v146	-,36396
v147	-,26204
v176	,74319
v178	,62731
v191	,41948
v269	,25990
v431	,30121
v432	,42837
v437	,29190
v438	,24198
v478	,20127
v480	,28750
v504	,26282
v509	,16616
v515	,82123
(Constant)	-1,22116

Classification results are shown in Figure 35.

Figure 35: Classification results for allergy outcome

Classification Results[a,c]					
v622_Allergy			Predicted Group Membership		Total
			0	1	
Original	Count	0	1862	570	2432
		1	840	1320	2160
	%	0	76,6	23,4	100,0
		1	38,9	61,1	100,0
Cross-validated[b]	Count	0	1858	574	2432
		1	845	1315	2160
	%	0	76,4	23,6	100,0
		1	39,1	60,9	100,0
a. 69,3% of original grouped cases correctly classified.					
b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.					
c. 69,1% of cross-validated grouped cases correctly classified.					

As conclusion for Allergy state, distance values are given by classification functions below using coefficients resulted from analysis:

Dis_Aller0=

v132_003 *0.33108 + v146 *0.75610 + v147 *0.42463 + v176 *0.05408 + v178 *0.65133 + v191 *0.97381 + v269 *0.35790 + v431 * (-0.40476) + v432 *0.25585 + v437 * (-0.15834) + v438 * (-0.24813) + v478 * 1.21478 + v480 * (-0.29084) + v504 * (-0.02307) + v509 *0.68447 + v515 *0.29400 -1.35783

Dis_Aller1=

v132_003 *0.59042 + v146 *0.40767 + v147 *0.17377 + v176 *0.76556 + v178 * 1.25188 + v191 * 1.37539 + v269 *0.60671 + v431 * (-0.11640) + v432 *0.66594 + v437 *0.12110 + v438 *(-0.01648) + v478 * 1.40746 + v480 * (-0.01561) + v504 *0.22854 + v509 *0.84354 + v515 * 1.08019 -2.68745

Discrimination function is built from canonical discrimination function coefficients:

Dis_Aller=

v132_003 *0.27090 + v146 * (-0.36396) + v147 *(-0.26204) + v176 *0.74319 + v178 *0.62731 + v191 *0.41948 + v269 *0.25990 + v431 *0.30121 + v432 *0.42837 + v437 *0.29190 + v438 *0.24198 + v478 *0.20127 + v480 *0.28750 + v504 *0.26282 + v509 *0.16616 + v515 *0.82123 -1.22116

Selected variables by discriminative procedure

In Figure 36 correctly classified percentage for different undertaken analysis considering age groups is shown. This way, Adults group will concern explained results, and Whole group will concern new analysis done considering a larger number of cases including children.

Figure 36:Results of classification by correctly classified cases by age groups

cases	healthy	Rhinitis	Asthma	Allergy
Whole group	70,4	76,7	92,1	66,7
Adults	69,9	73,8	86,4	69,3

In Figure 37 we can see final choice of variables considering the 4 outcomes.

Figure 37: Final choice of variables for outcomes

	Healthy	Rhinitis	Asthma	Allergy
v132_003		■		■
r137		■		
r138		■		
v140			■	
v146	■			■
v147				■
r149		■		
v150		■		
v153			■	
v160	■			
r161		■	■	
v176	■	■	■	■
v178	■	■		■
r179		■		
v182		■	■	
v191	■			■

	Healthy	Rhinitis	Asthma	Allergy
v199			■	
v269		■		■
v304			■	
v429			■	
v430			■	
v431	■		■	■
v432	■	■		■
v437	■	■		■
v438	■			■
v444			■	
v478				■
v480				■
v504	■		■	■
v509			■	■
v515	■	■		■

4.5 Analysis by tree methods

Tree procedure will be developed by using Statistica version 10, statistic software package from Statsoft. [11]

Algorithm used for tree procedure is called CHAID, Chi-squared Automatic Interaction Detector. It is one of the oldest tree classification methods, based on a relatively simple algorithm that is particularly well suited for the analysis of larger datasets. His name derives from the basic algorithm that is used to construct trees, which for classification problems relies on the Chi-square test to determine the best next split at each step.

Specifically, the algorithm proceeds as follows:

Preparing predictors: The first step is to create categorical predictors out of any continuous predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of observations. For categorical predictors, the categories (classes) are "naturally" defined.

Joining categories: The next step is to cycle through the predictors to determine for each predictor the pair of categories that is least significantly different with respect to the dependent variable; for classification problems (where the dependent variable is categorical as well), it will compute a Chi-square test (Pearson Chi-square), which is our case. If the test for a given pair of predictor categories is not statistically significant, then it will join the respective predictor categories and repeat this step (i.e., find the next pair of categories, which now may include previously joined categories). If the statistical significance for the respective pair of predictor categories is, then it will compute a Bonferroni adjusted p-value for the set of categories for the respective predictor.

Selecting the split variable: The next step is to choose the split the predictor variable with the smallest adjusted p-value, if the smallest (Bonferroni) adjusted p-value for any predictor is greater than some alpha-to-split value, then no further splits will be performed, and the respective node is a terminal node.

Continue this process until no further splits can be performed (given the alpha-to-merge and alpha-to-split values).

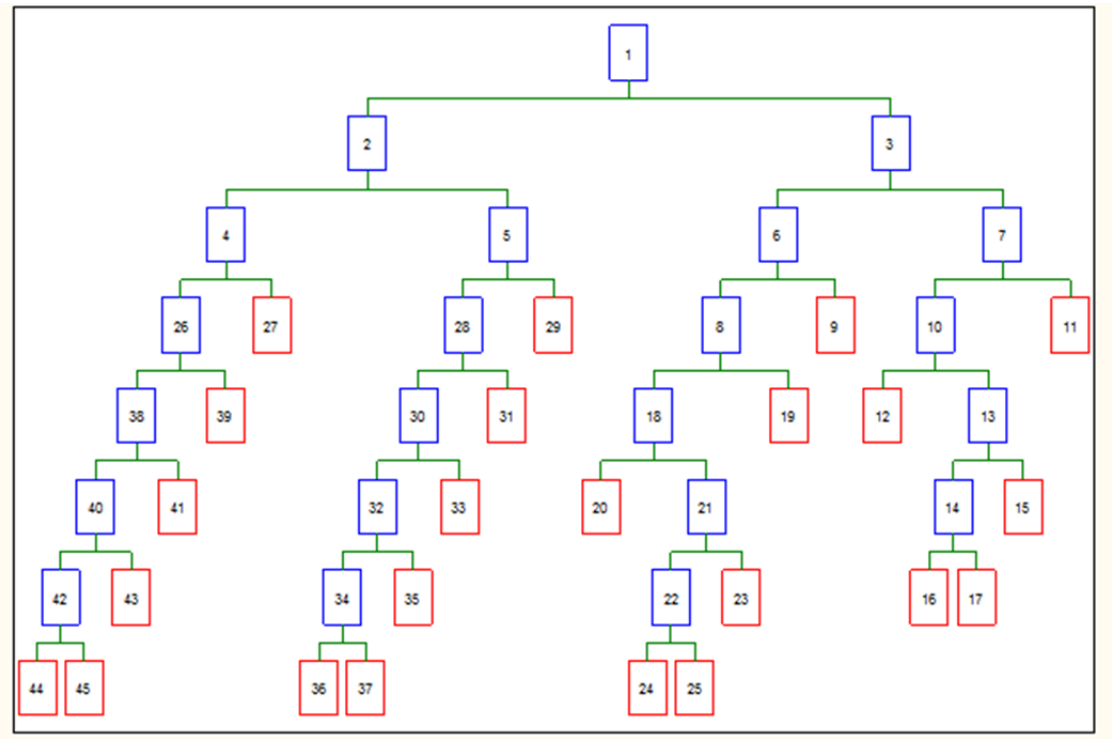
Trees are shown in the following pages.

Healthy tree

Different analyses by Statistica have been made considering different groups of data. Therefore trees for different analysis are shown. Results have been taking directly from software tool.

In a first option, all data has been used, considering data from all age groups. Tree is shown in Figure 38.

Figure 38: Tree for healthy outcome considering data from every age group



Data to understand tree is given in Figure 39. Name of the variable for the division will be the chosen variable to, answered by the patient, lead to a branch or another of the tree in order to get the healthy advice.

Figure 39: Healthy tree for every age group data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	4631	2373	2258	0	V178	No	Yes
2	2	2568	918	1650	1	V191	No	Yes
4	2	1604	466	1138	1	V431	No	Yes

26	2	1496	407	1089	1	V176	No	Yes
38	2	1424	368	1056	1	V504	No	Yes
40	2	1341	329	1012	1	V515	No	Yes
42	2	1314	317	997	1	V160	No	Yes
44		1307	313	994	1			
45		7	4	3	0			
43		27	12	15	1			
41		83	39	44	1			
39		72	39	33	0			
27		108	59	49	0			
5	2	964	452	512	1	V437	No	Yes
28	2	878	384	494	1	V431	No	Yes
30	2	767	319	448	1	V438	No	Yes
32	2	582	225	357	1	V504	No	Yes
34	2	524	194	330	1	V515	No	Yes
36		509	184	325	1			
37		15	10	5	0			
35		58	31	27	0			
33		185	94	91	0			
31		111	65	46	0			
29		86	68	18	0			
3	2	2063	1455	608	0	V176	No	Yes
6	2	916	544	372	0	V160	No	Yes
8	2	852	490	362	0	V438	No	Yes
18	2	704	386	318	0	V432	No	Yes
20		193	89	104	1			
21	2	511	297	214	0	V515	No	Yes
22	2	499	286	213	0	V191	No	Yes
24		260	136	124	0			
25		239	150	89	0			
23		12	11	1	0			
19		148	104	44	0			
9		64	54	10	0			
7	2	1147	911	236	0	V160	No	Yes
10	2	1008	785	223	0	V432	No	Yes
12		109	73	36	0			
13	2	899	712	187	0	V146	No	Yes
14	2	520	431	89	0	V437	No	Yes
16		79	58	21	0			
17		441	373	68	0			
15		379	281	98	0			
11		139	126	13	0			

Tree for different age groups are shown below.

Figure 40: Tree for healthy outcome considering data from adults

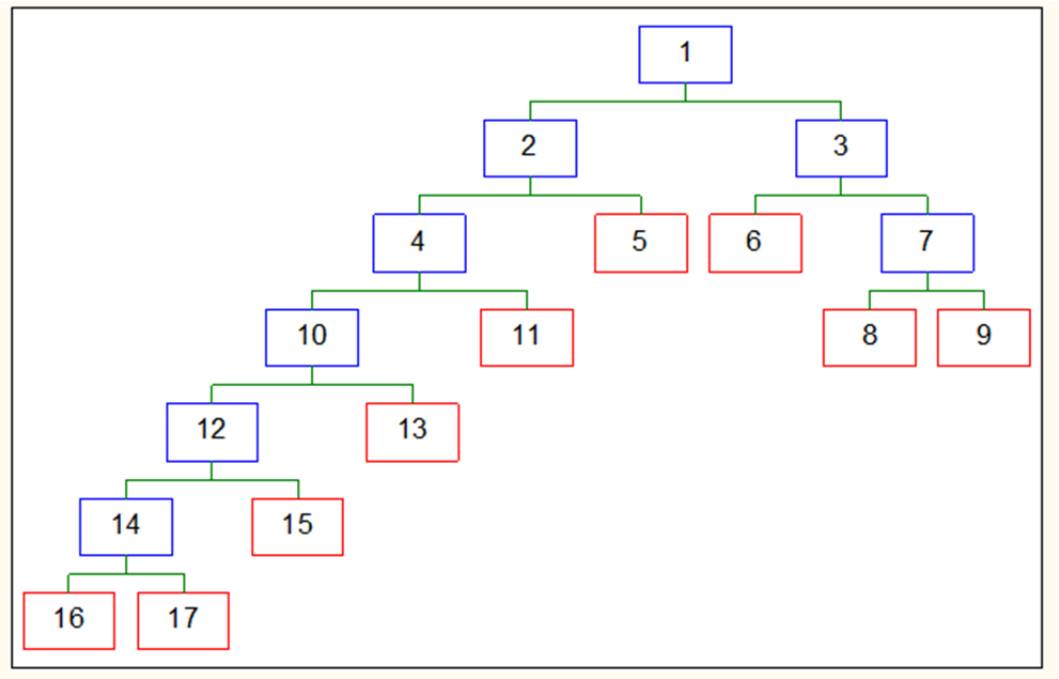


Figure 41: Healthy tree for adults group data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	2059	1080	979	0	V176	No	Yes
2	2	1484	618	866	1	V178	No	Yes
4	2	1043	357	686	1	V191	No	Yes
10	2	704	204	500	1	V160	No	Yes
12	2	691	192	499	1	V431	No	Yes
14	2	648	170	478	1	V515	No	Yes
16		628	160	468	1			
17		20	10	10	0			
15		43	22	21	0			
13		13	12	1	0			
11		339	153	186	1			
5		441	261	180	0			
3	2	575	462	113	0	V437	No	Yes
6		93	60	33	0			
7	2	482	402	80	0	V146	No	Yes
8		322	277	45	0			
9		160	125	35	0			

Figure 42: Tree for healthy outcome considering data from children (6/7 years)

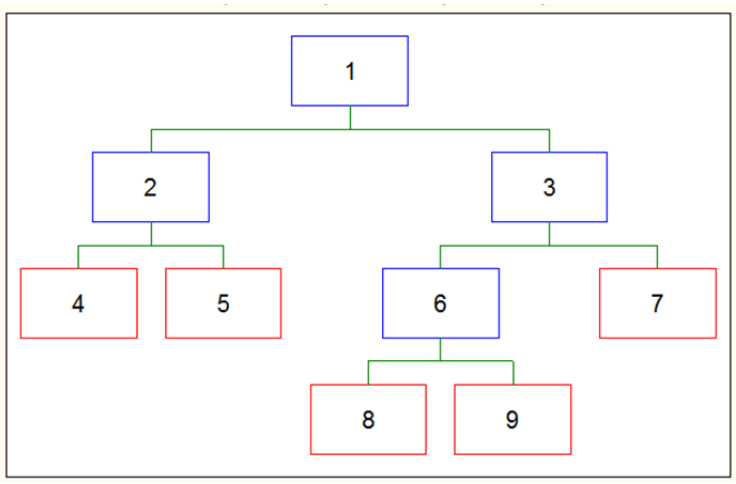


Figure 43: Healthy tree for children (6/7 years) group data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	1286	631	655	1	V178	No	Yes
2	2	727	253	474	1	V191	No	Yes
4		405	112	293	1			
5		322	141	181	1			
3	2	559	378	181	0	V160	No	Yes
6	2	511	334	177	0	V432	No	Yes
8		98	51	47	0			
9		413	283	130	0			
7		48	44	4	0			

Figure 44: Tree for healthy outcome considering data from adolescents (13/14 years old)

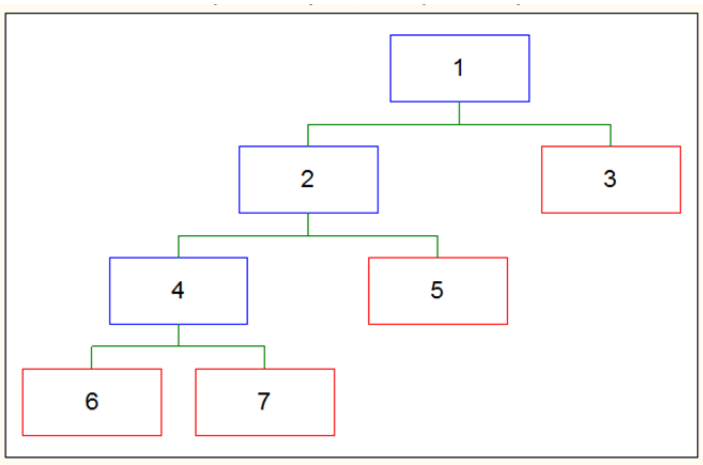


Figure 45: Healthy tree for adolescents (13/14 years old) group data

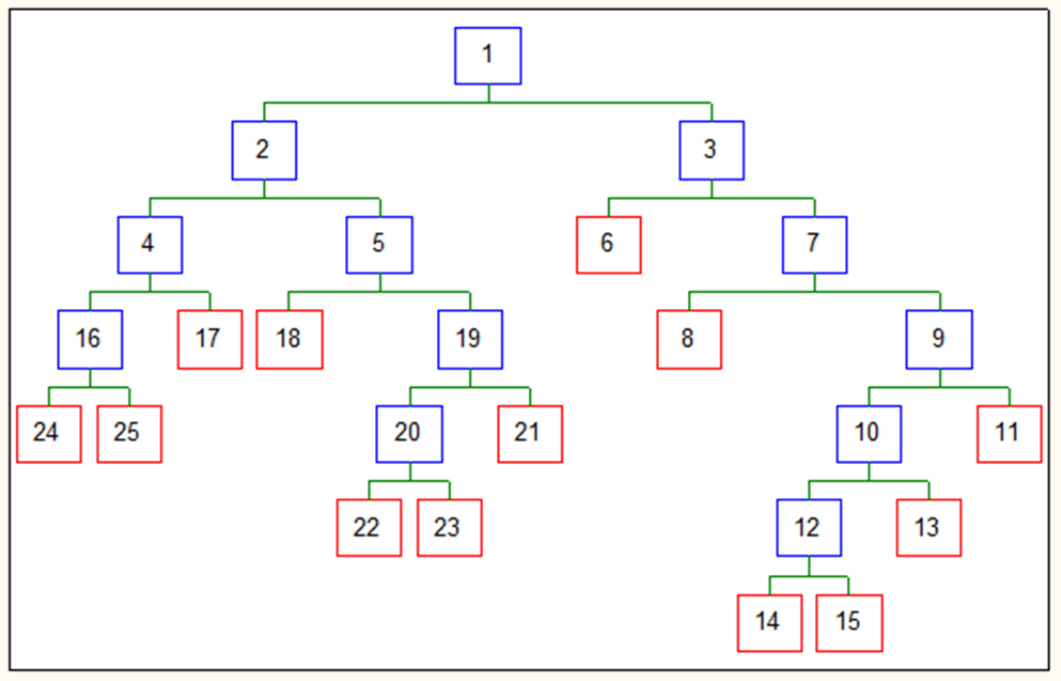
	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	1286	662	624	0	V176	No	Yes
2	2	907	356	551	1	V432	No	Yes
4	2	694	231	463	1	V191	No	Yes
6		445	118	327	1			
7		249	113	136	1			
5		213	125	88	0			
3		379	306	73	0			

Rhinitis tree

Different analyses by Statistica have been made considering different groups of data.

In a first option, all data has been used, considering data from all age groups. Tree is shown in Figure 46.

Figure 46: Tree for Rhinitis outcome considering data from every age group



Data to understand tree is given in Figure 44. Name of the variable for the division will be the chosen variable to, answered by the patient, lead to a branch or another of the tree in order to get the healthy advice.

Figure 47: Rhinitis tree for every age group data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	4667	3346	1321	0	V176	No	yes
2	2	3345	2771	574	0	V178	No	yes
4	2	2424	2157	267	0	V432	No	yes
16	2	2242	2012	230	0	V437	No	yes
24		2209	1987	222	0			
25		33	25	8	0			
17		182	145	37	0			
5	2	921	614	307	0	V432	No	yes
18		240	182	58	0			
19	2	681	432	249	0	V515	No	yes
20	2	661	427	234	0	R161	No	yes
22		615	404	211	0			
23		46	23	23	0			
21		20	5	15	1			
3	2	1322	575	747	1	V178	No	yes
6		163	100	63	0			
7	2	1159	475	684	1	V182	No	yes
8		244	130	114	0			
9	2	915	345	570	1	V150	No	yes
10	2	719	245	474	1	R149	No	yes
12	2	607	194	413	1	V432	No	yes
14		61	27	34	1			
15		546	167	379	1			
13		112	51	61	1			
11		196	100	96	0			

Tree for different age groups are shown below.

Figure 48: Tree for Rhinitis outcome considering data from adults

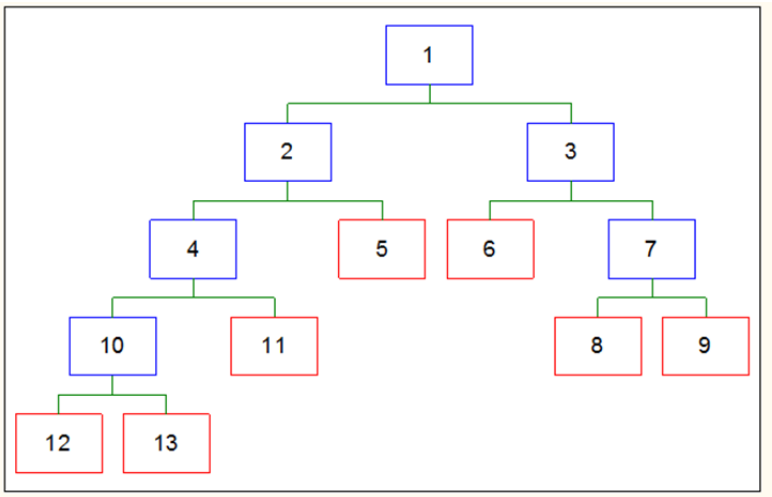


Figure 49: Rhinitis tree for adults group data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	2074	1450	624	0	V437	No	Yes
2	2	1488	1225	263	0	V178	No	Yes
4	2	1051	940	111	0	V182	No	Yes
10	2	868	767	101	0	V432	No	Yes
12		802	715	87	0			
13		66	52	14	0			
11		183	173	10	0			
5		437	285	152	0			
3	2	586	225	361	1	V176	No	Yes
6		103	62	41	0			
7	2	483	163	320	1	V182	No	Yes
8		133	60	73	1			
9		350	103	247	1			

Figure 50: Tree for Rhinitis outcome considering data from children (6/7 years)

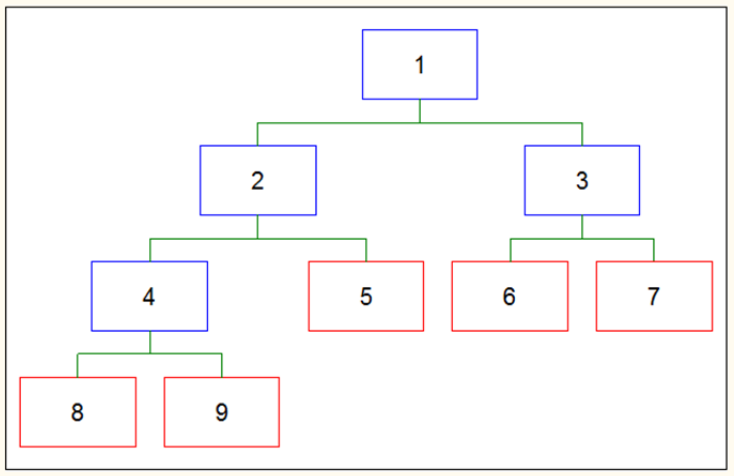


Figure 51: Rhinitis tree for children (6/7 years) group data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	1297	987	310	0	V432	No	Yes
2	2	770	676	94	0	V176	No	Yes
4	2	709	634	75	0	V178	No	Yes
8		641	578	63	0			
9		68	56	12	0			
5		61	42	19	0			
3	2	527	311	216	0	V176	No	Yes
6		230	156	74	0			
7		297	155	142	0			

Figure 52: Tree for healthy outcome considering data from adolescents (13/14 years old)

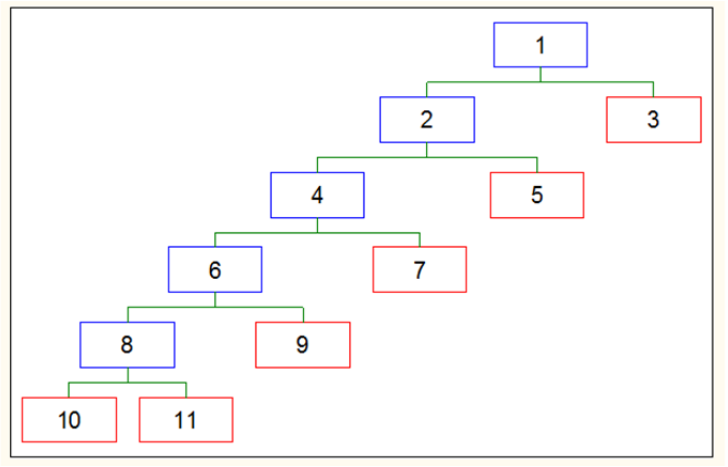


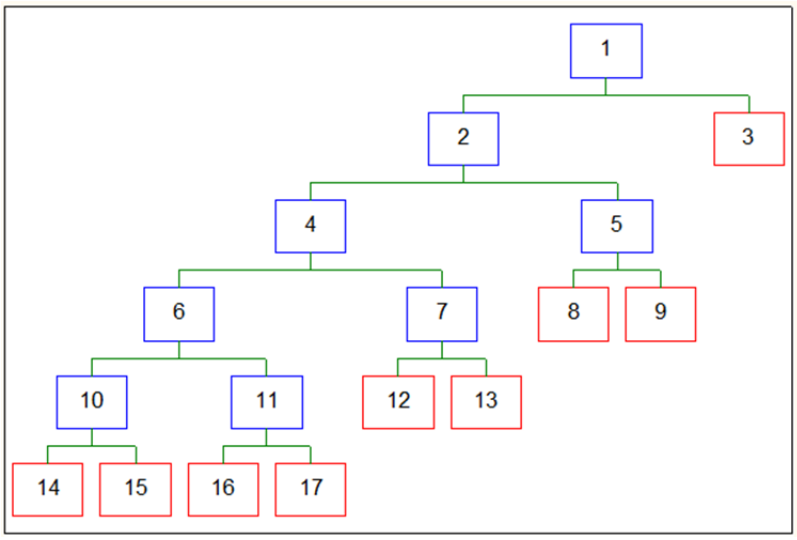
Figure 53: Rhinitis tree for adolescents (13/14 years old) group data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	1296	909	387	0	V176	No	Yes
2	2	908	751	157	0	V432	No	Yes
4	2	696	613	83	0	R161	No	Yes
6	2	679	602	77	0	V182	No	Yes
8	2	522	472	50	0	V150	No	Yes
10		506	460	46	0			
11		16	12	4	0			
9		157	130	27	0			
7		17	11	6	0			
5		212	138	74	0			
3		388	158	230	1			

Asthma tree

Different analyses by Statistica have been made considering different groups of data. In a first option, all data has been used, considering data from all age groups. Tree is shown in Figure 54.

Figure 54: Tree for Asthma outcome considering data from every age group



Data to understand tree is given in Figure 55. Name of the variable for the division will be the chosen variable to, answered by the patient, lead to a branch or another of the tree in order to get the healthy advice.

Figure 55: Asthma tree every age group data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	4640	4151	489	0	V429	No	Yes
2	2	4380	4052	328	0	V304	No	Yes
4	2	3825	3630	195	0	V431	No	Yes
6	2	3258	3127	131	0	V176	No	Yes
10	2	2614	2536	78	0	V504	No	Yes
14		2447	2382	65	0			
15		167	154	13	0			
11	2	644	591	53	0	V153	No	Yes
16		529	494	35	0			
17		115	97	18	0			
7	2	567	503	64	0	V504	No	Yes
12		397	359	38	0			
13		170	144	26	0			
5	2	555	422	133	0	V504	No	Yes
8		224	196	28	0			
9		331	226	105	0			
3		260	99	161	1			

Tree for different age groups are shown below.

Figure 56: Tree for Asthma outcome considering data from adults

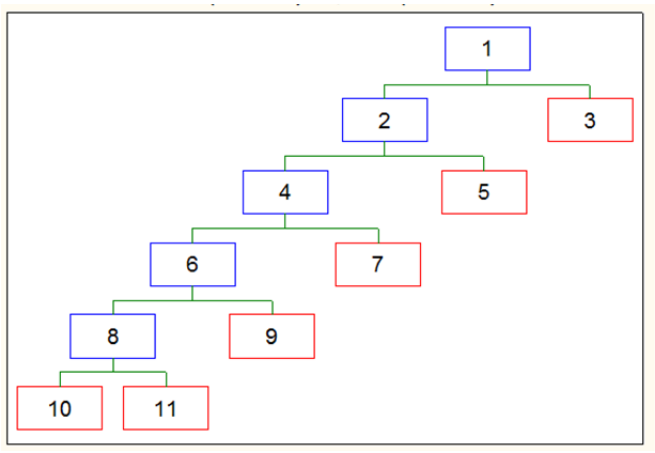


Figure 57: Asthma tree adults group data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	2059	1861	198	0	R161	No	Yes
2	2	1958	1831	127	0	V304	No	Yes
4	2	1788	1705	83	0	V431	No	Yes
6	2	1528	1474	54	0	V153	No	Yes
8	2	1338	1299	39	0	V176	No	Yes
10		1080	1057	23	0			
11		258	242	16	0			
9		190	175	15	0			
7		260	231	29	0			
5		170	126	44	0			
3		101	30	71	1			

Figure 58: Tree for Asthma outcome considering data from children (6/7 years)

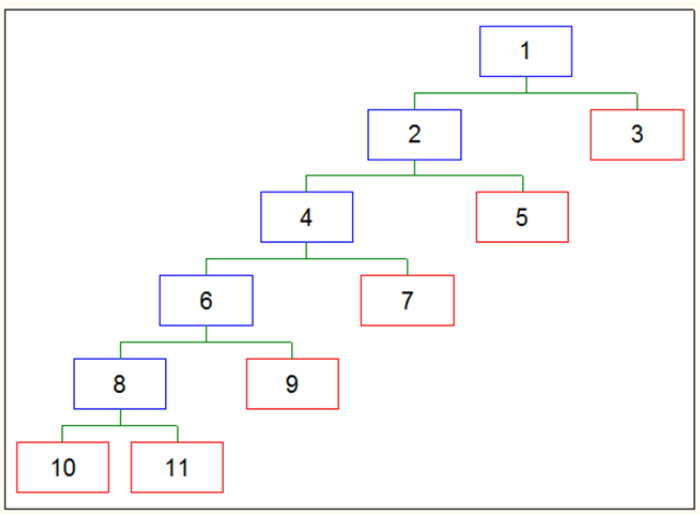


Figure 59: Asthma tree children (6/7 years) group data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	1295	1149	146	0	V504	No	Yes
2	2	999	940	59	0	V429	No	Yes
4	2	993	937	56	0	V176	No	Yes
6	2	791	758	33	0	V304	No	Yes
8	2	729	703	26	0	V431	No	Yes
10		642	623	19	0			
11		87	80	7	0			
9		62	55	7	0			
7		202	179	23	0			
5		6	3	3	0			
3		296	209	87	0			

Figure 60: Tree for Asthma outcome considering data from adolescents (13/14 years old)

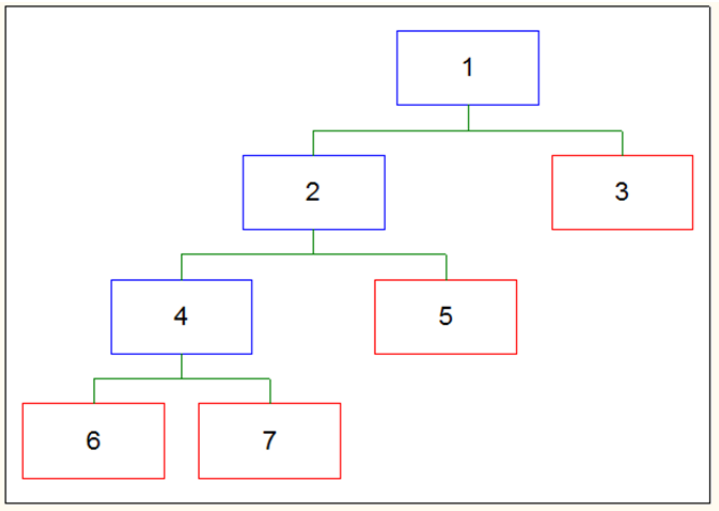


Figure 61: Asthma tree adolescents (13/14 years old) group data

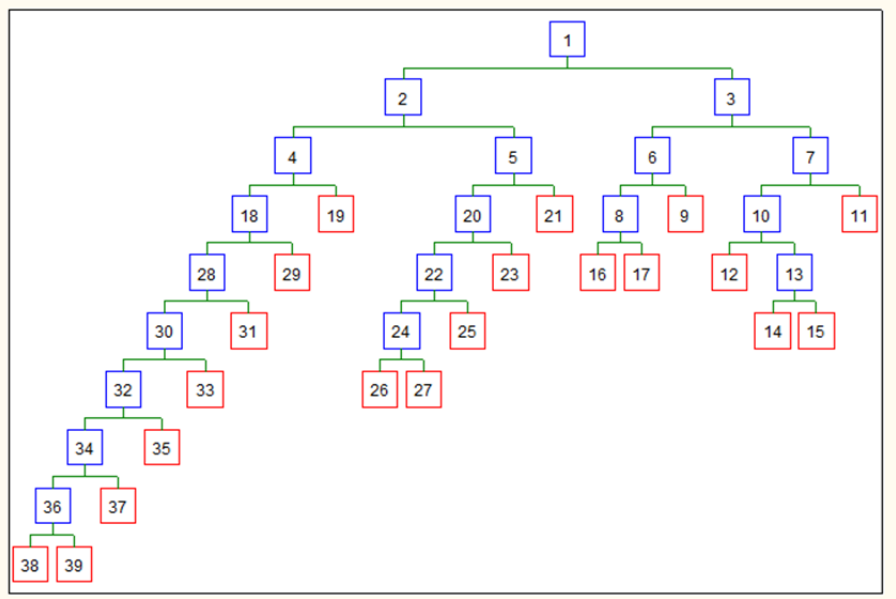
	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	1286	1141	145	0	V429	No	Yes
2	2	1192	1104	88	0	V304	No	Yes
4	2	1044	994	50	0	V153	No	Yes
6		894	863	31	0			
7		150	131	19	0			
5		148	110	38	0			
3		94	37	57	1			

Allergy tree

Different analyses by Statistica have been made considering different groups of data.

In a first option, all data has been used, considering data from all age groups. Tree is shown in Figure 62.

Figure 62: Tree for Allergy outcome considering data from every age group



Data to understand tree is given in Figure 63. Name of the variable for the division will be the chosen variable to, answered by the patient, lead to a branch or another of the tree in order to get the healthy advice.

Figure 63: Allergy tree data for all age groups data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	4592	2432	2160	0	V178	No	Yes
2	2	2546	1731	815	0	V191	No	Yes
4	2	1599	1199	400	0	V176	No	Yes
18	2	1520	1160	360	0	V431	No	Yes
28	2	1421	1102	319	0	V269	No	Yes
30	2	1302	1029	273	0	V504	No	Yes
32	2	1229	983	246	0	V432	No	Yes
34	2	1190	958	232	0	V515	No	Yes
36	2	1169	945	224	0	V478	No	Yes
38		870	715	155	0			
39		299	230	69	0			
37		21	13	8	0			
35		39	25	14	0			
33		73	46	27	0			

31		119	73	46	0			
29		99	58	41	0			
19		79	39	40	1			
5	2	947	532	415	0	V480	No	Yes
20	2	737	459	278	0	V437	No	Yes
22	2	675	438	237	0	V431	No	Yes
24	2	593	399	194	0	V478	No	Yes
26		313	229	84	0			
27		280	170	110	0			
25		82	39	43	1			
23		62	21	41	1			
21		210	73	137	1			
3	2	2046	701	1345	1	V176	No	Yes
6	2	905	415	490	1	V509	No	Yes
8	2	495	254	241	0	V191	No	Yes
16		260	148	112	0			
17		235	106	129	1			
9		410	161	249	1			
7	2	1141	286	855	1	V147	No	Yes
10	2	811	180	631	1	V432	No	Yes
12		90	36	54	1			
13	2	721	144	577	1	V146	No	Yes
14		489	83	406	1			
15		232	61	171	1			
11		330	106	224	1			

Trees for different age groups are shown below.

Figure 64: Tree for Allergy outcome considering data from adults

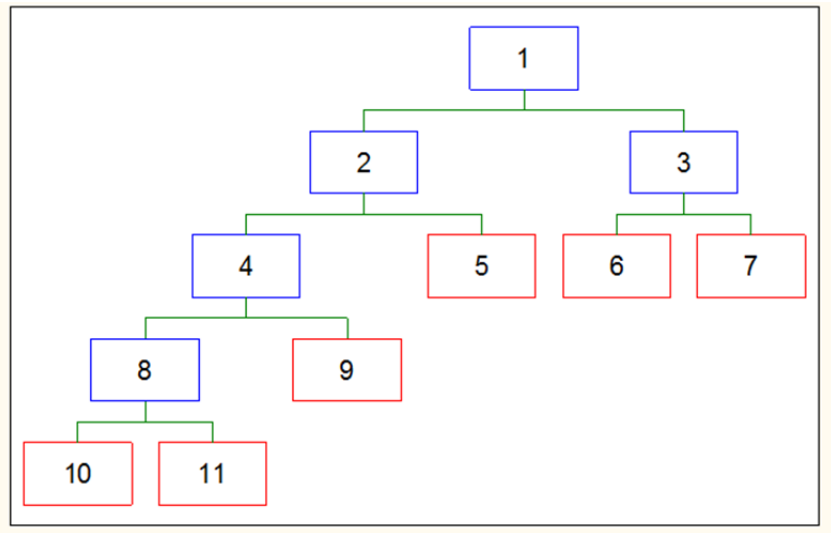


Figure 65: Allergy tree data for adults data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	2040	1055	985	0	V176	No	Yes
2	2	1469	917	552	0	V178	No	Yes
4	2	1032	718	314	0	V191	No	Yes
8	2	700	529	171	0	V478	No	Yes
10		439	350	89	0			
11		261	179	82	0			
9		332	189	143	0			
5		437	199	238	1			
3	2	571	138	433	1	V437	No	Yes
6		93	38	55	1			
7		478	100	378	1			

Figure 66: Tree for Allergy outcome considering data from children (6/7 years)

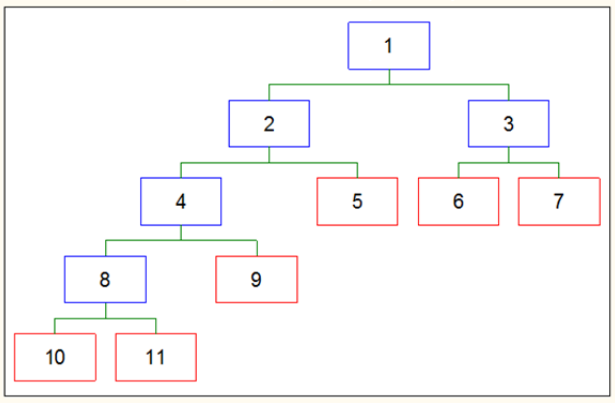


Figure 67: Allergy tree data for children (6/7 years) group data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	1276	702	574	0	V178	No	Yes
2	2	724	500	224	0	V480	No	Yes
4	2	611	451	160	0	V431	No	Yes
8	2	540	412	128	0	V269	No	Yes
10		456	359	97	0			
11		84	53	31	0			
9		71	39	32	0			
5		113	49	64	1			
3	2	552	202	350	1	V480	No	Yes
6		382	162	220	1			
7		170	40	130	1			

Figure 68: Tree for Allergy outcome considering data from adolescents (13/14 years old)

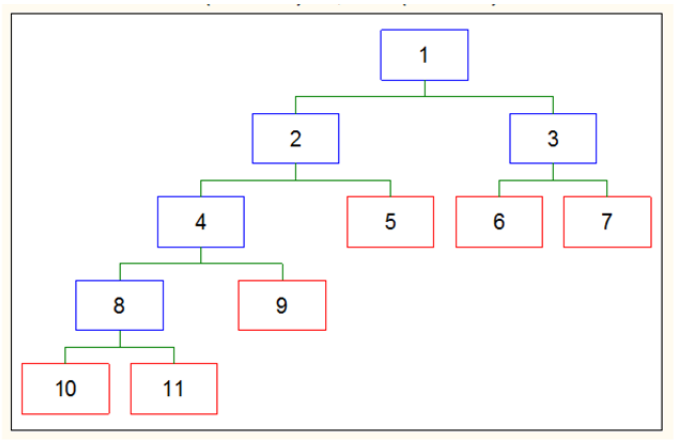


Figure 69: Allergy tree data for adolescents (13/14 years old) group data

	Number of children	Number of cases in node	N class 0	N class 1	Category	Name of variable for division	Category for children 1	Category for children 2
1	2	1276	675	601	0	V176	No	Yes
2	2	898	578	320	0	V269	No	Yes
4	2	699	491	208	0	V178	No	Yes
6	2	546	413	133	0	V431	No	Yes
8	2	519	400	119	0	V480	No	Yes
10	2	507	395	112	0	V515	No	Yes
12	2	498	391	107	0	V437	No	Yes
14	2	488	386	102	0	V504	No	Yes
16		452	363	89	0			
17		36	23	13	0			
15		10	5	5	0			
13		9	4	5	1			
11		12	5	7	1			
9		27	13	14	1			
7		153	78	75	0			
5		199	87	112	1			
3		378	97	281	1			

5. Conclusions

- Preliminary study by basic statistical parameters (means, correlation, prevalence OR) shows up to be efficient as results are later supported by the methods of data mining as discriminative and classification procedures. This way we get an intuitive idea of variables or group variables for different outputs before proceeding to analyze data deeply and assessing their accuracy and compactness.

- In discriminative analysis, accuracy achieved is good enough for qualitative discrimination but not so high enough for classification (starting at about 70% up to about 92% depending on the output variables) in comparison to other problems of machine-processable knowledge from healthcare data, due to the lack of some data (missing value in datasets for variables chosen for calculating membership functions) and enough independent variables to model efficiently the reality of health status parameters. We end up turning to the “nearest neighbor” of health status parameters based on real data instead of having a completely effective classification only based on data analysis without environmental and healthcare interpretation .

- Another cause of such “no perfect in 99,99%” results (not being ideally mapped by the group from the sampled population) is the origin of real raw data: data consistency is not perfect as diagnosis going with survey answers supported by medical checks is given by different doctors with probably somewhat fuzzy methods of reasoning and evaluating health status as known from some research and data mining literature devoted to the process of medical diagnosis decisions by human experts.

- Analysis for different data considering age groups of patients results in different conclusions, due to the differences in symptoms depending on age and characteristic features of allergic diseases and their development in different socio-economic environmental conditions .

6. Acknowledgements

First place I want to thank my parents, Javier and Montse, for raising me as good as they have done, this report wouldn't have been possible without their support, their effort in making a thinking and lively person out of me. I hope someday I will be able to give everything you did for me back.

I also want to thank my classmates, who have been by my side during 5 years of exams, practical and theoretical lessons and study hours in the library, and all the fantastic moments we have lived in these now ended academic years. Specially I want to thank Óscar Amatriain for helping me when I got lost, when I thought answers didn't exist, and when I was about to throw in the towel, as well as for so many things I couldn't be able to resume here.

It's also needed to name my uncles, Carlos and Jorge, who encouraged me to get involved with these technical studies, emphasizing my skills to face them. You were true. To my uncle Juan and my grandmother, who have played an important part these last years, showing them proud of me.

I also want to thank Dr. Zbigniew Wawrzyniak , from Politechnika Warszawska, for understanding and helping me in the development of this report, and also for his advices that have made my stay in Poland quite easier. To Miguel Angel Gomez Laso, a great professor and a better tutor, who has helped me since the first moment I decided to immerse in this fantastic experience that this six months in Warsaw have been.

To all those who, direct or indirectly, have been with me in this years of study, helping me, listening to my problems, making me laughing or just supporting me in the distance. To all my friends, in Spain and in Warsaw, thank you for being there.

Thank you very much. Dziękuję bardzo.

7. Bibliography

1. Zolnoori M, Fazel Zarandi MH, Moin M, Heidarnezhad H, Kazemnejad A., Computer-aided intelligent system for diagnosing pediatric asthma., J Med Syst. 2012 Apr;36(2):809-22. Epub 2010 Jul 10.
2. TRAINMOR-KNOWMORE, TRAINing Material in ORganisational KNOWledge Management for European ORganisations & Enterprises: ©2005-2008 TRAINMOR-KNOWMORE Partners. http://www.trainmor-knowmore.eu/root.en.aspx
3. ECRHS: http://www.ecrhs.org
4. ISAAC: http://isaac.auckland.ac.nz/
5. ECAP: http://ecap.pl/pdf/ECAP_metoda.pdf
6. Expert system: http://www.csc.liv.ac.uk/~tbc/comp210/17lectureESMYCIN.pdf
7. Cohen, J., and Cohen , *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Routledge Academic; Third edition (August 1, 2002), ISBN-13: 978-0805822236.
8. Hays, W. L., and Winkler, R. L., *Statistics: Probability, Inference, and Decision,* Holt, Rinehart and Winston, 1971.
9. Shegog R., Sockrider M.M., *Computer-Based Applications in the Management of Asthma*, in Harver A., Kotses H., Asthma, Health and Society: A Public Health Perspective, Springer, 2008, pp.153-178.
10. Statistica, analytics statistics platform package version 10: http://www.statsoft.com/
11. SPSS: IBM SPSS Statistics, version 20, http://www-01.ibm.com/software/analytics/spss/products/statistics/

8. Appendix

Appendix 1: preselected variables of the survey used in project

V132 Does any other member of the family suffer from allergies?

V132_002 Mother

V132_003 Father

V136 1. Have you had wheezing or whistling in your chest at any time in the last 12 months?

V137 1.1 When the wheezing was heard in the chest, were you at all unable to take a breath?

V138 1.2. Did you have this whistling and wheezing when you did not have a cold?

V139 2. Have you woken up with a feeling of tightness in your chest in the last 12 months?

V140 3. Have you had an attack of shortness of breath that came on during the day when you were at rest in the last 12 months?

V141 4. Have you had an attack of shortness of breath that came on following strenuous activity in the last 12 month?.

V142 5. Have you been woken by an attack of shortness of breath at any time in the last 12 months?

V146 6. Have you been woken by an attack of coughing in the last 12 months?

V147 7. Do you usually cough first thing in the morning in the winter?

V148 8. Do you cough during the day or at night in the winter?

V149 8.1. Does this cough persist on most days for at least 3 months each year?

V150 9. Do you bring up any phlegm from your chest first thing after waking up in the morning in the winter?

V151 10. Do you often bring up phlegm from the chest during the day or at night in the winter?

V152 10.1. Do you bring up phlegm like this for longer than 3 months?

V153 11. Have you ever had any problems with your breathing?

V154 11.1 Do you have these problems with breathing :
Continuously so that your breathing is never quite right / Repeatedly, but it always gets completely better / Rarely

V155 12. Do you have difficulty walking because of any condition other than heart or lung disease?
Yes, I have difficulty walking because of a condition other than heart or lung disease. / No, I only have difficulty walking because of heart or lung disease. / No, I do not have any difficulty walking.

V160 14. Have you ever had asthma?

V161 14.1 Was asthma confirmed by a doctor?

V164 14.5 Have you had an attack of asthma in the last 12 months?

V165_001 14.6-14.7 How many attacks of asthma have you had in the last 12 months?

V165_002 14.6-14.7 How many attacks of asthma have you had in the last 3 months?

V166 14.8 How many times have you woken up because of an attack of asthma in the last 3 months?
every night or almost every night / more than once a week, but not most nights / at least twice a month, but not more than once a week/ less than twice a month / not at all

V167 14.9 How often have you had problems with asthma in the last 3 months?
Continuously / about once a day / at least once a week, but less than once a day / less than once a week / not at all

V168 14.10 Are you currently taking any medicines for asthma, including inhalers, aerosols or tablets?

V176 15. Do you have any nasal allergies, including a runny nose caused by allergy to pollens (hay fever)?

V178 16. Have you ever had a problem with sneezing or a runny or blocked nose when you did not have a cold, fever or flu?

V179 16.1. Have you ever had a problem with sneezing or a runny or blocked nose when you did not have a cold or flu in the last 12 months?

V180 16.1.1 Have these symptoms been accompanied by itchy or watery eyes?

V182 17. Have you used any medication to treat nasal disorders?

V183 17.1 Have you used any nasal aerosols for the treatment of your nasal disorder?

V187 17.2. Have you used any tablets or capsules for the treatment of your nasal disorders?

V191 18. Have you ever had eczema or any kind of skin allergy?

V192 19. Have you ever had any itchy rash that was coming and going for at least 6 months ?

V193 19.1 Have you had this itchy rash in the last 12 months?

V194 19.1.1 Has this itchy rash at any time affected any of the following places: the folds of the elbows, behind the knees, in front of the ankles, under the buttocks or around the neck, ears or eyes?

V195 20. Have you ever had any difficulty with your breathing after taking medicines?

V199 16.1.3. Do you suffer from nasal polyps?

V207 33. How often do you usually exercise so much that you get out of breath or sweat?
every day / 4-6 times a week / 2-3 times a week / once a week / once a month / less than once a month / never

V208 34. How many hours a week do you usually exercise so much that you get out of breath or sweat?
none / about ½ hour / about 1 hour / about 2-3 hours / about 4-6 hours / 7 hours or more

V209 35. Do you avoid taking vigorous exercise because of wheezing or asthma?

V250 61. Do you keep a cat/cats?

V253 62. Do you keep a dog?

V254 62.1. Is your dog allowed inside the house?

V256 63. Do you have any birds?

V269 73. Have you ever had an illness or trouble caused by eating a particular food?

V270 73.1. Have you nearly always had the same illness or trouble after eating this type of food?

V271 73.1.1 What type of food was this?
dairy products / cereals / eggs / fish /nuts / spices / vegetables / meats / citrus fruit / non-citrus fruit / none of the above / other

V272 73.1.2. Did this illness or trouble include
itchy skin and skin changes / diarrhea or vomiting / runny or stuffy nose / severe headaches / breathlessness / other / none of the above

V273 74. Have you ever smoked for as long as a year?

V274 74.1 How old were you when you started smoking?

V275 74.2. Do you still smoke [in the last month]?

V280 74.4 Do you or did you inhale the smoke?

V281 75. Have you been regularly exposed to tobacco smoke in the last 12 months?

V283 75.2 Do people smoke regularly in the room where you work?

V304 76.a Have you taken any medicines for breathlessness or problems with breathing?

V385 80. Have you had any other injections to treat your breathing problems in the last 12 months?

V396. 84.1. Have you sought help at the emergency medical service or intensive care unit in the last 12 months?

V404. Have you ever had breathing problems, wheezing or whistling or a feeling of tightness in your chest?

V405. 85. Have you ever spent at least one night in hospital because of your breathing problems?

V406. 85.1. Have you spent at least one night in hospital because of breathing problems in the last 12 months?

V409. 86. Have you been seen by a doctor because of breathing problems or because of shortness of breath?

V424. IM1 1. Have you ever had wheezing or wheezy chest in the past?

V425. IM1 2 Have you had wheezing or wheezy chest in the last 12 months?

V427. IM1 4. How many times have you been woken at night by breathlessness, attack of wheezing or wheezy chest?
Never / fewer than one night a week / one or more nights a week

V428. IM1 5. Has any attack of shortness of breath and wheezing over the last 12 months been severe enough to make speaking difficult or prevent you from saying more than one to two words between breaths?

V429. IM1 6. Have you ever had asthma?

V430. IM1 7. Have you had wheezing during or following strenuous activity in the last 12 months?

V431. IM1 8. Have you had dry coughing at night not related to a cold in the last 12 months?

V432. IM2 1 Have you ever had attacks of sneezing or abundant nasal discharge or stuffy nose when you did not have a cold or the flu?

V433. IM2 2 Have you had attacks of sneezing or abundant nasal discharge or stuffy nose when you did not have a cold or the flu in the last 12 months?

V434. IM2 3 Have these nasal symptoms been accompanied by itchy or watery eyes in the last 12 months?

V436. IM2 5 How much have these nasal symptoms interfered with your normal activity in the last 12 months?
Not at all / Not much / Moderately / Very much

V437. IM2 6 Have you ever had hay fever (allergic rhinitis)?

V438. IM3 1 Have you ever had itchy skin changes coming and going for at least 6 months?

V439. IM3 2. Have you had these itchy skin changes in the last 12 months?

V440. IM3 3. Have these itchy skin changes at any time affected any of the following places: the folds of the elbows, behind the knees, in front of the ankles, under the buttocks or around the neck, ears or eyes?

V443. IM3 6. On average, how often were you unable to sleep or were woken because of these itchy skin changes in the last 12 months?
Never / Less than one night a week / One or more nights a week

V444. IM3 7. Have you ever had eczema, atopic dermatitis or any other inflammation of the skin?

V460 IM4 4. How many times do you exercise so much that your breathing becomes labored/faster?
Never or occasionally / Once or twice a week / Three or more times a week

V478 A120. Is your skin dry and needs frequent moisturizing?

V479 A121. Does your skin (e.g. of the hands) tend to get irritated after exposure to chemicals (soap, detergents, organic solvents)?

V480 A122. Are skin changes made worse by certain foods?

V481 A123. Do any chronic skin disorders run in your family?

V483 A125. Have you had adverse reactions to alcohol such as nettle rash or angioedema?
Don't drink alcohol / to spirits (40%) / to red wine / to white wine / to beer / Haven't had such reaction / None of the above

v504 Have you ever had breathing problems, wheezing or whistling or a feeling of tightness in the chest?

v505 Have there been any days when you have had to give up activities other than work (e.g looking after children, the house, studying) because of your asthma, shortness of breath or wheezing in the last 12 months?

v507 U131 Do you tend to have a stuffy nose for a few weeks or months a year without any accompanying symptoms (itching, sneezing, RUNNY NOSE, conjunctivitis)?

v509 U131.2. Have you used any nasal anticongestants in the last 12 months?

V514 U132. Do you suffer from recurrent otitis media?

V515 U133. Have you ever had anaphylactic shock?

V517 U134. Do you get angioneurotic oedema(swelling of e.g. hands or lips)?
Don't know / Haven't had / Have had with reddened skin, rash and itching / Have had with pale skin and without itching

V532 EX1. Have you ever had the sensation of phlegm flowing down the back of your throat in the past? (post-nasal drip)?

V533 EX1.1 Have you had the sensation of phlegm flowing down the back of your throat over the past 12 months (post-nasal drip)?

V550 EX3. Have you ever had allergy after an insect bite in the past?

V552 EX3.2 What were the symptoms of this allergy?
Swelling over 10 cm, but only on one segment of the body over 24 hours
Urticaria, pruritus, malaise, anxiety
Nausea, vomiting, diarrhea, dizziness, pain in the abdomen /
Shortness of breath, wheezing, urticaria all over the body, difficulty speaking and swallowing, dead scare
Fainting, loss of consciousness, cyanosis
Others

V554 EX4. Has a DOCTOR ever diagnosed you with food allergy?

V556 EX4.2Which products make you allergic?
Categorical refusal to answer / Meat (e.g. Beef, lamb, chicken, pork) / Fish or seafood (e.g. shrimps) / Citrus fruit / Kiwi fruit / Other southern fruit (e.g. bananas) / Small-stone fruit (e.g. apples, pears) / Vegetables (green and root) / Leguminous vegetables (soya, beans, peas, lentils) / Cereal products (groats, flour) / Hazelnuts / Walnuts / Peanuts / Cow milk / Diary products (cheese) / Yolk of the egg / White of the egg / Others