UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

Máster Universitario en

Ingeniería de Sistemas y de Control

BIOINFORMATICS: USING STATISTICAL METHODS AND ARTIFICIAL INTELLIGENCE FOR COMPLEX METABOLIC DISORDERS UNDERSTANDING

Author: Víctor Javier Cerquera Parra
Director(s): María Guijarro Mata-García
Juan Jiménez Castellanos
María Insenser Nieto

Curso académico 2017 - 2018

Convocatoria: febrero 2018

Máster Universitario en Ingeniería de Sistemas y de Control

BIOINFORMATICS: USING STATISTICAL METHODS AND ARTIFICIAL INTELLIGENCE FOR COMPLEX METABOLIC DISORDERS UNDERSTANDING

Author: Víctor Javier Cerquera Parra
Director(s): María Guijarro Mata-García
Juan Jiménez Castellanos
María Insenser Nieto

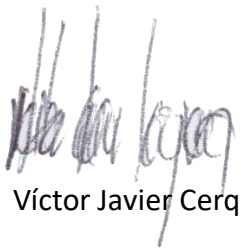
CALIFICACIONES

AUTORIZACIÓN

Autorizo a la Universidad Complutense y a la UNED a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a sus autores, tanto la memoria de este Trabajo Fin de Máster, como el código, la documentación y/o el prototipo desarrollado.

I hereby authorize the UNED to publish and use for academic, non-commercial purposes and expressly mentioning to its authors, both the memory of this master's degree, and the code, documentation and / or prototype developed.

Firmado / Signed:



Víctor Javier Cerquera

ABSTRACT

Polycystic ovary syndrome (PCOS), is a set of symptoms that affects women between the ages of 18 and 44 years [1], and is one of the leading causes of poor fertility [2]. The current work starts by trying to use bioinformatics in order to help on finding parameters that may indicate that a patient has PCOS, and then, by using classification techniques setup Support Vector Machines (SVM), that allows the doctors an early PCOS detection. In order to find out which are these parameters that may be linked to the source of the syndrome, a statistical analysis procedure is proposed. In order to create this procedure, first the current data was analysed, and then based on the specific restrictions imposed by the data, the procedure was depicted. Data restrictions made us think of a generic procedure, that could consider different existing analysis tests, in order to find these parameters.

Finally, after the procedure is followed, the tool will find a set of parameters that can be considered as interesting, and from these, a SVM classification machine is configured, so that it can be used to analyse new patients and help on PCOS detection.

KEYWORDS

Analysis of variance, Post hoc tests, False discovery rate, Support vector machines, Statistical models

INDEX

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Medical research is one of those fields that contributes to the welfare of people. In particular, medical investigation that tries to find the cause of complex diseases or find procedures in order to early detect medical conditions has an extra value given the social factor that it addresses.

It is natural to use computer science in order to improve the type and quality of the analysis performed. Nevertheless, bioinformatics will only get the researcher up to certain point guiding the medical research into a way or the another.

The definition of biostatistics can be said to be “Biostatistics is the science of both, the theory and methodology for the acquisition and use of quantitative evidence in the biomedical research” [3 p. 10]. This statement indicates that Biostatistics is such a tool needed in all kind of biomedical research, which helps to prove that the steps given in a particular research are correct according to the data collected by the researchers.

It is needed that a researcher supports the conclusions of the investigation, even when the lack of data can’t fully support or prove a theory. Once more data is available, it will further support or disprove the hypothesis being evaluated, but the lack of data should not prevent the research or investigation.

This by itself, shows how important the whole cycle of data collection, documentation, analysis, inference, etc., is. Biostatistics in particular, in the current project framework, deals with the methodologies used for data analysis and how significance tests can be applied in order to search for different variables in our data that may not be easily spotted and that could show indicators that lead to the detection of a particular disease.

Biostatistics as the science providing the methodologies for data analysis, will help in the current work to create a methodology of study, with several steps that once followed will validate the existence (or not), of variables that may be worth to be further studied in the medical research.

Such methodology will be described in this document, including terms, definitions and examples to better understanding the procedure and steps done.

Together with researchers from the Hospital Ramón y Cajal, from Madrid – Spain, the current work tries to advance in the early detection of a particular disease.

In this work, we will make use of biostatistics, in order to guide and help the researchers with the analysis of hundreds of different variables that may or may not indicate a specific condition being studied.

1.1 Justification

The medical research, and also the fact that a bioinformatics work can have an impact in the society by aiding in the research of a medical condition, is justification enough to perform this work.

On top of that, creating a tool that can be reused by researchers (given a set of initial conditions), is a good motivation to do the work presented in this document.

1.2 Objectives

The project main objective is to explore, research and develop a strategy that determines a set of significative variables which could lead to the detection of a specific medical condition related with metabolic disorders, specifically polycystic ovary syndrome (PCOS). This strategy must also include the creation of an AI tool, which helps with the detection of the medical condition, based on the medical and laboratory analyses done to a patient.

A secondary objective of the current work is to provide a generic tool, which automatically applies the strategy developed, so that researches can use it when the variables changes, the number of patients increase or even for other medical condition being studied, which share similar parameters.

Another objective is to make it easier for the end users (medical researchers, or other field researchers), to use AI in order to help classifying new data. This is done, by helping on the data analysis and detecting the interesting variables for a research and by also determining automatically a set of parameters for a classification machine, that helps bridging the gap of the specific knowledge needed to configure a SVM (Support Vector Machine), by making it automated and more practical and easier to use.

2.1 Fundamentals

The statistical analysis of data collected in an experiment, or as in the case being studied, the medical and laboratory analyses performed on a set of defined groups, allows to search for those variables that may indicate a statistical difference, and such that may be worth of further studies by the medical researchers.

In order to perform such statistical analysis, we need to go back to an important concept in biostatistics: the theory of probability.

In medical research, it is common to talk about that a set of parameters, conditions or variables, shows a difference statistically not significant or that the difference is significant at a P value of 0.05 [3 pp. 13 - 16].

On any experiment performed (like a medical experiment), exist always a probability that the results obtained are due to the natural variation, also referred as a chance factor [3 p. 11].

This factor should not be more than 5%, or if it is said the other way around, the confidence in the conclusions should be greater than 95%, for the obtained results to be significative.

Groups and Tests

The groups in a research, are those different sets of the population subject to study that share a common characteristic. One group could be the control group, whereas another group is the group being tested.

Another example could be group differentiation based on sex, or based on a relevant feature for the research like race (if we are studying blood RH or some blood conditions, for instance, race is relevant).

In the current work, we have a GROUP variable, which indicates either: Women with PCOS, healthy women and men.

Tests, it refers to the different variables or experiments being conducted on the population, and the results obtained. It could be to use a drug or a placebo, or the presence or absence of a protein in the body. In this work, we have over 400 tests, with data collected from the different patients.

Here is where the main research objective is: to try and detect which of these variables or tests are relevant in the studied group (women with the metabolism disorder). In the present document, *variables* term is used.

Natural variation

If in a dice roll experiment, 5 consecutive ones are rolled out the first time the experiment is attempted, we call that a natural variation. A researcher could wrongly think that these results are due to an external cause, but after testing the data obtained in the experiment, we can find out if the results obtained are due to the natural variation.

In the roll dice case, we know that our expected frequency is 1/6, but in a real experiment control groups are needed because the frequency is unknown, hence we need the control group to compare the results. This comparison should have a significative difference between the control group and the studied group.

This significance level (or P value), should be less than 0.05. This means that the property (significant difference), being studied may have occurred by natural variation in less than 5% of the cases, or in other words, the confidence that we have that there is a significative difference in the property being studied is more than 95% [3 pp. 13 - 16]. In the current context, when a property has a significative difference between 2 groups, then we say that this property or variable is interesting to be studied further.

In the given case, 3 groups are being considered: healthy women, men and women with the metabolism condition. It could be considered the healthy women group as a control group, but also, the men group can be considered as such. In this case, the researchers are looking for features where the medical condition may be affecting the women who has it, and changing their normal parameters and turning them into parameters with no significative difference with the men group.

There are several significance tests that can be used, to calculate this P value, which will be detailed in this document. However, a question arises: why it is considered significant properties those at $P < 0.05$?

The answer to this is related to the normal distribution. In this distribution the studied properties values fall around the median in a symmetrical way. Usually, medical parameters fall in a normal distribution curve [3 p. 13].

For instance, in the next graph, we can see the distribution of soldier's chest thickness which follows a normal distribution [4].

Just by looking at the bar graph, it can be inferred that most of the men has a chest thickness around 40 inches.

Over the bar chart, we have the normal distribution curve. On this distribution, if we take two standard deviations on each side of the mean it will cover up to 95% of the area under the curve.

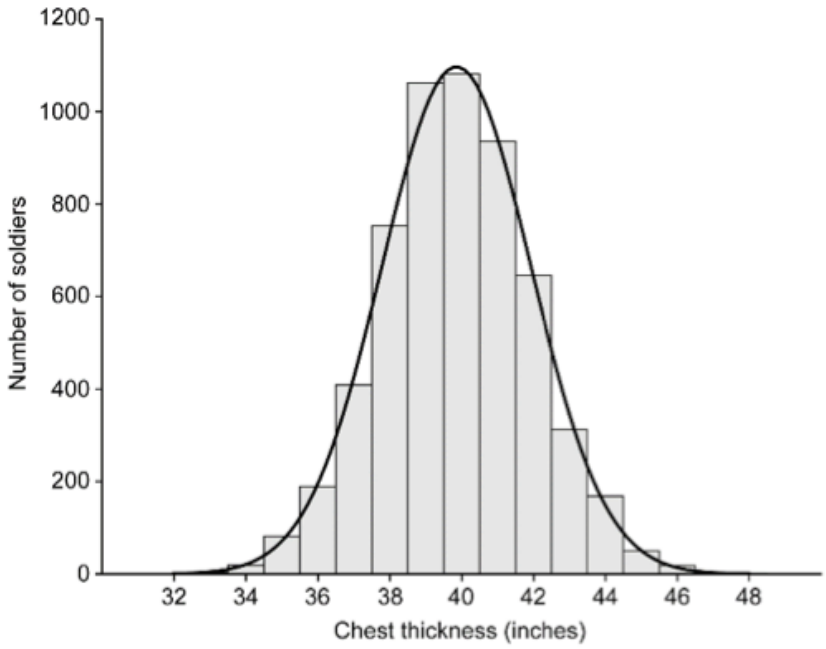


Figure 1. Bar chart representing the normal distribution curve for soldier's chest thickness [5]

This means, that 95% of the data collected are under this area (median plus 2 times the standard deviation on each side). The remaining 5% will fall outside the area of two standard deviations. Hence, this 95% or area up to two standard deviations is known as the confidence range [6] [3 pp. 13 - 16]. It means, that the value of a property has a 95% confidence of falling in this area. Any property value outside of this area is assumed to differ significantly [3 pp. 13 - 16].

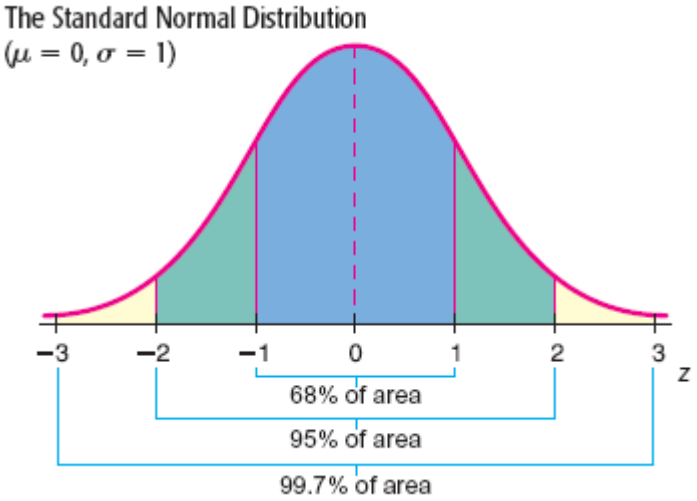


Figure 2. The standard normal distribution graph [6]

This is the reason why on a significance test, if the P value of the test is under 0.05, it is considered as significant, the lesser the value, the stronger the significance. The value of 0.05 is arbitrary, by choosing a lesser significance value ($P < 0.01$, for instance), stronger significance will be achieved, but it may lead to increase the false negative errors, making it difficult to have enough proof to support the hypothesis being studied [3 p. 16]. To avoid these false negatives, the P value is increase and usually $P \leq 0.05$ is used, although it is important to mention the value used in the research.

Statistical inference

The whole purpose of the statistical inference is to generate new knowledge from our data. This can be done, by going one step further and express the data in statistical terms, trying to predict behaviours or results from the data or classify an individual based on the data collected and the information we have from the population. In general, we aim to generate new knowledge.

As we said before, significance tests can be used to find out the significant difference between two or more groups. Every time we run such a test we start with a null hypothesis, which assumes that there is no significant difference between the groups. Then, after calculating the tests and finding out the P value, we say that if $P > 0.05$, the null hypothesis is then accepted, or in other words, we say that there is no significant difference between the groups.

Likewise, if $P \leq 0.05$, the null hypothesis is rejected and we conclude that there is a significant difference between the groups.

This is what we will use for inferring information regarding the metabolic disorders.

Type I and Type II errors

False positive and false negative errors, also known as type I and type II errors respectively, are the error that occur when a null hypothesis is rejected, when it is actually true (type I), or vice versa, when a null hypothesis is not rejected when the alternate hypothesis is valid.

In simple words, type I errors refers to observe a statistically significative difference when in truth there is none. Type II errors then, is to fail to observe a statistically significative difference when in truth there is one. [7]

Significance tests: Chi test

This test is used when we need to check the significance difference between two groups. The key factor is that the test acts on discrete variables [3 p. 36].

These are the conditions that the data has to meet in order to be a candidate for Chi test:

a. Variables must be discrete. It means that the variables cannot take continuous values, but rather have to take fixed values, like our variable OBESE (either a patient is obese or not).
b. The sample must be random
c. There must be at least 5 elements, for all the different possible values that the variable being tested may take.
d. Values must be numbers. This means that we have to quantize any qualitative variables that we want to study (for instance: OBESE = 1, NO OBESE = 2).
e. No percentages can be used.

The variables of the present study are not discrete, this discards the Chi test as a viable test to find interesting variables.

Significance tests: Fisher test

Fisher's test is used when the condition **c** is not met for Chi test. In other words, when we have a small amount of data [8] [3 p. 40].

This is the case of our current study. However, like Chi test, Fisher test act on discrete variables and not on continuous ones, hence it is discarded as well.

Significance tests: T test (T Student's test) and Paired T tests

The main requisite with T test, is that it is used to compare data from two groups, which both must have normal distribution curves [9] [10] [11] [3 p. 45].

In the case of the Paired T test, the feature being analysed must have two values on the same person or element of the sample, like for instance the blood pressure before and after training [3 p. 49].

In the current study we could apply T test. Paired T tests do not apply, as we don't have several measurements on the same variable for each sample patient.

Significance tests: Mann-Whitney-Wilcoxon (MWW) test

Unlike T test which requires a normal distribution, we do use MWW test when the data doesn't follow a normal distribution curve [12] [13] [3 p. 54].

The rest of the requirements are:

a. The data from the different groups are independent
b. Data can be discrete or continuous
c. Each individual must be independent on its own group
d. Data can be ordered or ranked. It means that variable values can be compared
e. Based on the shapes of the distribution curves, the MWW test will be used to calculate if either there are differences with the groups medians in order to find out whether there are differences on the distributions or not.

MWW test won't yield a low P value, when you have 7 or less individuals. This, together with the fact that MWW test will find out whether the distribution differ on both groups, instead of

pointing out the statistical significant difference on a variable, it won't be used on this study.

Significance tests: ANalysis Of Variance (ANOVA)

With ANOVA, the difference between the means of two groups is tested. This is done by looking at the variance of the data [14 p. 736] [3 p. 67].

T test should be used when comparing two groups, in other case, ANOVA is used. Furthermore, when ANOVA is applied to two groups, it produces the same results than T test [3 p. 68].

Table 1 compiles the information regarding which test to use.

Test	Requirements
Chi test	Discrete variables Normal distribution Data are numbers Two or more groups At least 5 observations
Fisher's test	Chi test with no minimum observations
T test	Two or more groups (in the case of paired) Normal distribution Continuous variables
MWW test	T test conditions with not normal distribution
ANOVA test	T test conditions with 2 or more groups

Table 1. List of statistic test and its requirements

Post hoc tests in ANOVA

When we do apply ANOVA to multiple groups, it will indicate that the null hypothesis is rejected, hence there is a significative difference that indicates that at least a group is different from the others [15] [14 p. 746]. However, ANOVA won't tell us which group is different. This is when post hoc tests are used, to find out which is the group that actually shows differences [15]. There are many different post hoc tests, each with different conditions to be met in order to consider them valid.

It is easy to think that in order to avoid applying post hoc tests, ANOVA can be applied on all pairs of the groups being studied. However, this approach will end up decreasing the confidence of the results, as each time that ANOVA or any test is applied to a pair, with $p \leq 0.05$, means that there is a 5% probability of error rejecting the null hypothesis (that there is no significative difference between the groups), just by chance, in other words, to run into a type I error [16].

If tests are applied three times, the chance to run into an error scales up to:

$$0.95^3 = 0.86$$

Which means that overall, p ≤ 0.14, which doesn't add much confidence on the results.

In the current work, we can end up with up to 15 pairs or combinations of groups:

$$0.95^{15} = 0.46$$

Which means p ≤ 0.54, which is not useful at all, as we have a 54% chance to run into a type I error.

In order to solve this problem when multiple groups are involved, several different post hoc tests have been created, some are more powerful than others and/or have different uses.

Tukey HSD test

The Tukey's Honest Significant Difference test, is a test that compares all groups pair combination of means, to find out which specific groups are different [17] [18]. This comparison is done through the calculation of HSD, which represents the distance between the groups.

In order to calculate the HSD test, the following formula is used:

$$HSD = \frac{M_i - M_j}{\sqrt{\frac{MS_w}{n_h}}}$$

Where $M_i - M_j$ is for the pair being calculated, the means difference, MS_w is the Mean Square Within, and n is the number of elements in the group.

There are a few assumptions for this test:

- Group's normal distribution
- Observations are independent
- Exist a homogeneity of variance

When n, the sample size for the groups, is unequal across the groups, then the estimated standard deviation is used for each pair comparison. This is the Tukey-Kramer post hoc test [17].

Bonferroni test (one step correction)

The Bonferroni test or Bonferroni procedure applies T test on each pair combination, controlling the error by dividing the expected p value by the number of pairs to be tested [19] [20] [21].

Thus, this test just makes sure that by pair to pair testing, the error is not increased as the number of combination increases. Just as ANOVA for 2 groups becomes a T test, applying Bonferroni to two groups, will also become a simple T test [22].

The less number of comparisons or group pairs involved, the more powerful Bonferroni's test is comparing to Tukey test. Vice versa is also valid, the larger number of pair tests involved, the more powerful Tukey's test become.

Bonferroni's test does not assume comparison independency.

Bonferroni's formula:

$$\alpha[PT] \approx \frac{\alpha[PF]}{C}$$

Where C is the number of tests, $\alpha[PF]$ is the desired α confidence value, and $\alpha[PT]$ is the α value that should be used for each test.

Holm Bonferroni test (step-down method using Bonferroni adjustments)

Sometimes, Bonferroni test is seen as a too conservative or strict test [21]. The Holm Bonferroni or Holm's Sequential Bonferroni procedure, is more statistically powerful than the single-step Bonferroni. The formula used to calculate the Holm Bonferroni is:

$$HB = \frac{\alpha}{n - r + 1}$$

Where α is the target confidence value pursued (usually 0.05), n is the number of paired test, r is the ranking of the current pair being calculated, ordered by ascending degree of significance.

Sidak test (one-step correction)

Sidak test is a pair multiple comparison test, based on a statistical t [23]. This test, corrects the significative level for multiple comparisons and yields narrower limits than Bonferroni's test. This is a more conservative test than Bonferroni's, in fact, there is a relationship between Sidak and Bonferroni tests, where Bonferroni is the first linear term of a Taylor expansion of the Sidak equation [24]:

$$\alpha[PT] = 1 - (1 - \alpha[PF])^{1/C}$$

Sidak's test do assume that each comparison is independent from the others.

Holm Sidak test (step down method using Sidak adjustments)

This is a test which has more power than Bonferroni test, but that doesn't compute confidence intervals for the median difference, for every comparison [25]. However, one important feature of this test, is that the data required to calculate it is a list of p values (any list of p values). This means, that this test could be done by itself and not as a follow up after ANOVA.

The test, verifies each p value from smallest to greatest, verifying, for a total number of m tests:

$$p_k > 1 - (1 - \alpha)^{1/m-1+1}$$

Whenever this becomes true, then that p value and all the other greater than it will indicate a non-statistical significance, and likewise all previous hypothesis from H_1 … H_{k-1} will reject the null hypothesis, hence have a statistically significative difference.

Simes Hochberg test (step-up method)

Also known as just the Hochberg test, this procedure starts by ordering the hypothesis (H_i), based on their p_i values, in ascendant order [26] [27].

Then, being m the total number of tests (or combinations to be validated), for a specific α value, the procedure calculates which is the highest p_k value where the following inequality holds true:

$$p_k \leq \frac{\alpha}{m+1-k}$$

Once this H_k hypothesis has been located, then all null hypothesis from H1 … Hk, must be rejected.

This procedure, is only valid when there is non-negative dependency between the samples.

Hommel test (closed method based on Simes tests)

This method makes use of the Simes test to control the FWER, applying it in the closure framework [28].

In order to follow Hommel procedure, start by ordering the m hypothesis based on ascending p values.

Then, find the j value, where the following inequality holds true:

$$p_{m-j+k} > {}^{k\alpha}/_{j}$$

For all k = 1 … j

If j can not be found, then all null hypothesis are rejected. Otherwise, all the Hi hypothesis where $p_i \leq {}^{\alpha}/_{j}$, must be rejected.

FDR Benjamini / Hochberg test

This post hoc test focuses on controlling the false discovery rate, which is the proportion of group comparisons detected as statistically significative different, which actually are not [29].

Simes mentioned back in 1986 a technique to control the false discovery rate [26], but it was Benjamini and Hochberg in 1995 who actually developed it [29].

To use the technique, p values for each test performed, the number of tests (the number of comparisons), and a false discovery rate Q must be known. The data must be sorted by p value in ascending order. To each comparison, once sorted, an *i* value is assigned from 1 up to m, and used those values, to each comparison the Benjamini-Hochberg critical value is calculated as $(i/m)Q$.

An example can be seen in the table 2, assuming a discovery rate of 0.05.

Comparison	p value	i	$(i/m)Q$
PCOS – WOMEN	0.003	1	$0.01\bar{6}$
PCOS – MEN	0.041	2	$0.0\bar{3}$
WOMEN – MEN	0.321	3	0.05

Table 2. Benjamini / Hochberg post hoc test example
Number of comparisons, m = 3

According to the example, the largest p values where it is still true that $p < (i/m)Q$, is for the PCOS – WOMEN comparison. So, the other two comparisons are not significative, despite to the fact that a p value < 0.05 is considered significative in the case of PCOS – MEN comparison.

Also, it could be that in some cases it is accepted that p values higher than the accepted 0.05 are significative. This may happen, based on the chosen discovery rate. In general terms FDR tests are more powerful but may increase type I errors [30].

Benjamini / Hochberg two step FDR correction test

This post hoc test, together with the Benjamini / Yekutieli and Benjamini / Krieger / Yekutieli, share the same procedure, changing only the formula to calculate the thresholds values [29].

The procedure is as follows:

1. The p values, are sorted from the highest to the lowest
2. Calculate the threshold for each p value, in order, and perform the validation:
$$p_i < Threshold(p_i)$$
3. When this validation holds true, then all the p values smaller than p_i, and including p_i will be considered as statistically significative

In the case of the two step Benjamini / Hochberg, the threshold is calculated, by interpolating the values between the largest and the smallest p values.

Being m the number of p values and q the desired percentage of false discovery rate (in decimal values, from 0 to 1), the threshold must be calculated as:

Smallest p value: q/m

Largest p value: q

Then, a linear interpolation is calculated between these 2 extremes.

FDR Benjamini / Yekutieli test

The key on the Benjamini – Yekutieli FDR test, is that it controls the false discovery rate under positive dependence assumptions [31].

As explained before, the method for calculation is the same than 2 step Benjamini / Hochberg, although the formula to calculate the thresholds changes.

Smallest p value: $q/[m \cdot (1 + 1/2 + 1/3 + \cdots + 1/m)]$

Largest p value: $q/(1 + 1/2 + 1/3 + \cdots + 1/m)$

Benjamini / Krieger / Yekutieli two step FDR correction test

In this case, an extra parameter needs to be provided, which is the estimated number of null hypothesis that are true (not rejected) [32].

With this, the threshold calculations are:

Smallest p value: $q/[(1 + q) \cdot Ntrue]$

Largest p value: $[q/(1 + q)] \cdot (N/Ntrue)$

FDR adaptive Gavrilov-Benjamini-Sarkar test

In addition to the two step tests (Benjamini / Hochberg and Benjamini / Krieger / Yekutieli), the Gravilov / Benjamini / Sarkar, are adaptive methods that estimate or adjust the number of null hypothesis. In order to do this adjustment, alpha value is recalculated based on this number of null hypothesis (m_0) [33].

In general terms, the new adjusted alpha is:

$$\alpha_{adj} = \frac{\alpha \cdot m}{m_0}$$

The two step Benjamini / Hochberg, will estimate the value of m0, by running a first stage of the normal Benjamini / Hochberg test. Afterwards, it will use the new adjusted alpha in a second run.

The two step Benjamini / Krieger / Yekutieli will use the normal Benjamini / Krieger / Yekutieli test to adjust alpha, adjusting it on each stage.

Finally, the Gravilov / Benjamini / Sarkar will estimate alpha but using only one stage.

SVMs: support vector machines

SVMs or support vector machines, are algorithms that aims to classify your data by using a kernel function, so that it can be labelled or categorize new elements, according to the different groups that your training data contains [34]. Generally speaking, SVMs can do both, classification and regression, and there are several steps involved on creating a SVM, as also, the kernel configuration will influence on how well your data is classified [35].

The kernel, will find a hyperplane that correctly classify your data. Depending on your kernel and the number of features involved this can be a very expensive algorithm in terms of computing resources [34] [35] [36].

SVM's are based on supervised learning, and in order to validate the vector machine, 2 steps are done: training and prediction, although a validation phase is usually added where the chosen kernel configuration can be validated.

First, in the training phase, a set of vectors and categories are supplied to the SVM algorithm, which will calculate a hyperplane that can separate all the different vectors into the different categories supplied.

The prediction phase, will take new vectors and attempt to classify them by using the following formula:

$$y = \sum_i \alpha_i H(s_i, X) + b$$

Where α_i, b and s_i are the support vectors calculated during the training stage, and *H()* is the kernel function used.

In the current work, a multiclass SVM is setup, by using a 'one against all' strategy [37], where a SVM is configured for each category in order to separate the vectors that belongs to that category from the other categories.

'One against all' strategy was first introduced by Vladimir Vapnik in 1995 [37], where a vector is classified on a category if and only if the SVM that classifies for that category accepts the vector and it was rejected by all the other SVMs categories.

The most common kernel functions used are [38] [39] [40]:

a. linear kernel [41]

$$H(x, x') = \langle x, x' \rangle$$

b. polynomial kernel [42]

$$H(x, x') = (\langle x, x' \rangle + r)^d$$

Where r is the coefficient and d the polynomial degree

c. rbf, which yields a gaussian shape [43]

$$H(x, x') = exp(-\gamma \|x - x'\|^2)$$

Where γ is a number greater than 0

d. Sigmoid [44]

$$H(x, x') = tanh(\gamma \langle x, x' \rangle + r)$$

Finally, the validation step, takes a set of data not used in the training and uses the SVM configuration found to classify the data. Then, the results from the classification is compared with the real category to which each vector belongs, in order to validate and find out which is the prediction rate.

Stratified sampling

In order to guarantee that the SVMs are trained with enough data that belongs to each category present in the initial data, stratified random sampling is used [45 pp. 141-145]. First, the data is separated on each different category (each one known as a stratuma), then it is randomized and finally, keeping the initial data ratios enough data is selected from each stratum to be used in the training.

2.2 Procedure

In order to meet the objectives, a strategy has to be developed, that allows to gather statistical information from a set of data. Moreover, from the fundamentals, and after evaluating the type and amount of data received from the medical researchers, it has been decided to use ANOVA

as the test procedure, to evaluate the different groups combinations, in order to find interesting variables that may be of importance in the metabolic disorder, polycystic ovary syndrome (PCOS) study, and in general terms also, for any other experiment being carried out that may use the generic tool.

The strategy used to find these variables follows these steps:

Step 1: Data cleansing

Removing non-existent data (NULL values), and prepare the data provided in order to be used. Values in some of the variables are not complete for all the individuals in the sample, and hence the data needs to be cleansed before it can be tested and analysed. This will be a fully automated step.

Step 2: Discrete variables detection

Detect the variables that have discrete values such as **GROUP**, **OBESE**, **AGE**, etc., in order to decide which groups to use in the statistical analysis. These groups are the ones that we will apply a statistical test on (ANOVA), in order to detect the variables that shows statistical differences and that are suggested by the test as interesting for further research.

The test will be applied so that one of the variables is the base group variable, or the variable that allows the categorization by the main study groups. In the current disorder being researched, this group is called **GROUP**.

Step 3: Grouping

The user, will have to choose a base group variable, however, for the current research, as stated before it will be the **GROUP** variable.

This step will create all the group combinations, between the base group variable and the other discrete variables. This means, that if **OBESE** is another discrete variable, we will create groups that result from the combination of **GROUP** and **OBESE**, in order to find variables with statistical differences.

An example of this, is that after grouping **GROUP** and **OBESE** we will have 6 groups:

a. OBESE PCOS

b. NON OBESE PCOS
c. OBESE women
d. NON OBESE women
e. OBESE men
f. NON OBESE men

Which will be tested in the next step.

Step 4: Testing

ANOVA test is applied for each variable, to all the groups detected in the step 2, as the statistical test that will reject or validate the null hypothesis ($p \geq 0.05$). ANOVA test, will be applied assuming that the researcher supplies data that holds ANOVA requirements.

One of these requirements, the normal distribution validation, is especially sensitive, as in many cases where there is a lack of subjects to get data from (just like the current research on the PCOS), will mean that satisfying such requirement is impossible.

Moreover, normal distribution tests, do require a minimum number of elements in a group (usually 20), and validating if a group belong to a normal distribution with less elements won't be feasible. It is necessary to take with extreme caution, the results obtained from data that do not fulfil the requirements for its analysis.

The tool developed does include an option to enforce normal distribution validation before applying ANOVA, to ensure that the results obtained are based on data that satisfy the procedure requirements.

Step 5: Post hoc validation

For those variables which ANOVA test indicates that are statistically significant, a set of post hoc tests will be applied, in order to find out which are the groups that show a difference statistically significant.

However, the fact that a group is statistically significant different from another, doesn't mean anything unless that information is set on the context of a research. This is beyond this work, as the work can only go up to facilitate data exploration and indicate interesting variables to further research. The researcher is the one who has to provide a decision tree that validates when a variable is interesting based on the results of the post hoc tests.

Step 6: Significative variables determination

With the post hoc test results, the user must decide which are the values the analysis must focus on. For instance, in the current research, we have three different categories on the **GROUP** variable:

- PCOS
- Women
- Men

Researchers could consider interesting parameters, those that show a significative difference for the group of women with the disorder (PCOS).

Furthermore, when more than one variable is selected on the step 5, then the value combinations is shown so that the tool can focus the analysis on a specific subgroup.

For instance, when **GROUP** and **OBESE** groups are selected, 6 pairs are obtained, as shown on the step 3.

Then, researchers could focus the analysis on the Obese and Non-obese women with the disorder (OBESE PCOS and NON OBESE PCOS).

Step 7: Classification machine

Finally, with the set of interesting variables found on the step 5, and based on the values set by the user (which determine when a variable must be considered interesting), in the step 6, a classification machine is created, which can automatically determine if a given patient belongs to one of the main groups of study.

The classification uses a one-against-all approach and the classification machine will do an automatic search through all the four kernels:

- linear kernel: $\langle x, x' \rangle$
- polynomial: $(\gamma \langle x, x' \rangle + r)^d$
- rbf: $exp(-\gamma \|x - x'\|^2)$
- sigmoid: $tanh(\gamma \langle x, x' \rangle + r)$

The system will automatically train the classification machine, preserving 10% of the data for validation, applying stratification random sampling, to guarantee that there are samples belonging to each group.

The variables chosen to be included on this automatic SVM configuration, are those found on the base group analysis.

Moreover, the user will have the ability to choose how many post hoc test does the variable must satisfy, in order to be included in the SVM training.

The results of the training and the validation are shown to the user, where it can be seen whether the chosen kernel is appropriate for the data and variables being analysed.

Figure 3, illustrates the 7 steps explained in the procedure, starting with the raw data input, and up to the SVM generation.

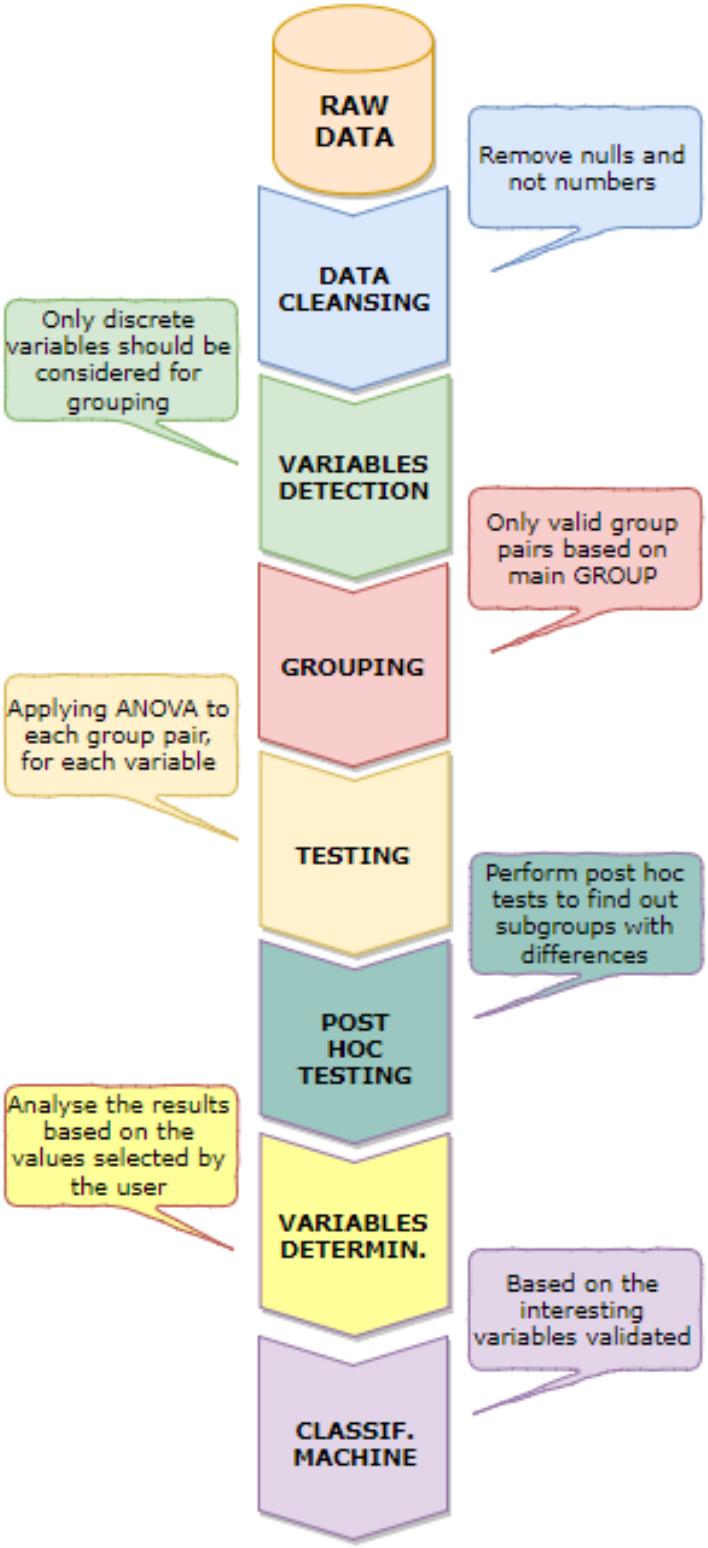


Figure 3. Procedure workflow

Procedure example

A full run will illustrate the application of the methodology for the base group. The first 2 steps are generic and will be applied before the proper statistical analysis.

1. Data cleansing: is done by removing any null or non-numeric values from the data.
   In the code (python language [46]), this cleansing is done by casting the data, applying a *to_numeric* [47] function to each value:

```
datos_base[i] = datos_base[i].apply(pd.to_numeric, errors='coerce')
```

   The coerce parameter will set the value to *NaN*.

2. The Variable Detection step will find which variables have discrete values. After running this on the example data we find the following list of discrete variables:
   - GROUP
   - SEX
   - AFHA
   - AFOB
   - AFDM
   - AFDL
   - AFCVD
   - AFHTA
   - OBESE

   With this variables list, group combinations are created based on the discrete variables selected for study. For this example, though, only the **GROUP** variable is selected to be used for combinations.

3. The system then creates as many value combinations based on the number of variables selected and the values of these variables.
   For the **GROUP** variable three analysis alternatives are shown to the user:
   - PCOS
   - Control women
   - Control men

   The user will choose which of these alternatives will be further analysed by the tool. If more variables are selected by the user, then more variables values options and groups will be created.
   In this example, only the “women with the disorder: PCOS alternative”, is chosen for analysis.

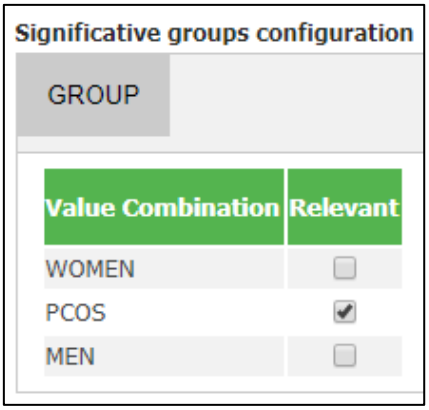


Figure 4. GROUP variable value selection

The user can also select which post hoc tests to apply in the analysis. All of them are chosen on this example.

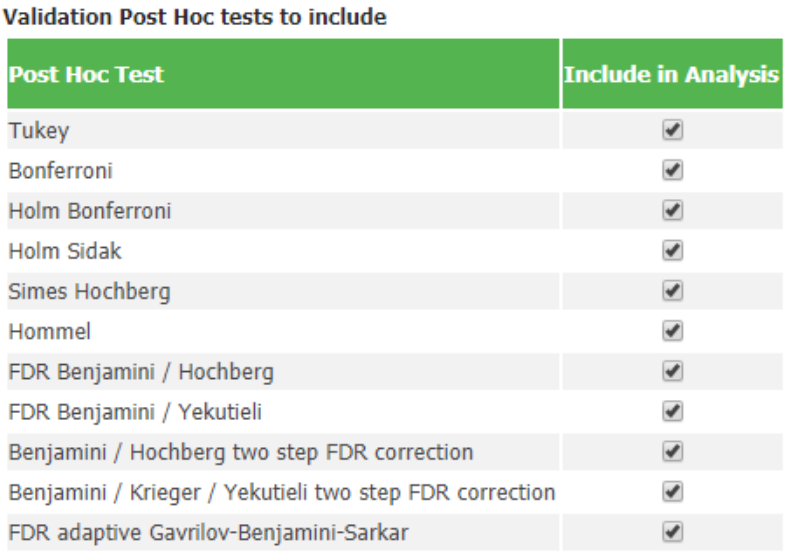


Figure 5. Post hoc test selection

4. This step is purely statistical. For each variable combination selected on the previous step and for each data variable, ANOVA is calculated. In the table 3, it is shown the results f and p, for the **GROUP** variable ANOVA test, for some data variables:

Variable	f	p
AUCsin_FGF21_L	1.137540	0.328763
normAUCsin_FGF21_L	1.137663	0.328724
Androstendiona_0	**12.499722**	**0.000040**
FGF21_P0	1.245038	0.296702
FGF21_P60	0.806341	0.452209
FGF21_P120	0.301499	0.741044
IL6_P0	**5.160457**	**0.009176**
normAUCsin_FGF21_P	1.684481	0.195899

Table 3. Extract of ANOVA results applied on the base group
In bold, statistically significative variables ($p < 0.05$)

5. After applying ANOVA, it is proceeded to its validation, through the use of post hoc tests. In this case, all the post hoc tests have been selected for testing, to all the variables that ANOVA indicated to be interesting, i.e. $p \leq 0.05$ (67 variables found).
   In the Table 4, an extract of this variables can be seen.

Variable	f	p
GalectinL240	3.209578	0.04881863
AUCsinGalectinL	3.331846	0.04381557
AdiponecG60	3.807608	0.02889484
GIPL120	4.062447	0.02318440
GIPL240	6.167487	0.04036041
AUCsinGIPL	4.646777	0.01409493
normAUCsinGIPL	4.646723	0.01409557
GIPP120	3.802939	0.02901216
IL6G0	3.364561	0.04256948

Table 4. Extract list of statistically significative variables found for the base group

Once these significative variables have been found, the data analysis can start.

Based on the values selected in the step 3 (only the women with the disorder group was selected), and also on the post hoc selected, these variables will be checked further by each post hoc, figuring out if they do show that the variable marked by ANOVA as interesting, do actually indicate a significative statistically difference between the specified value(s) (figure 4 and procedure step 6), against any other value.

This means, if a variable is marked as different, why is it different, and if the value for that variable is interesting for the researcher (according to what he or she specified). The results are shown on the figures 6 and 7.

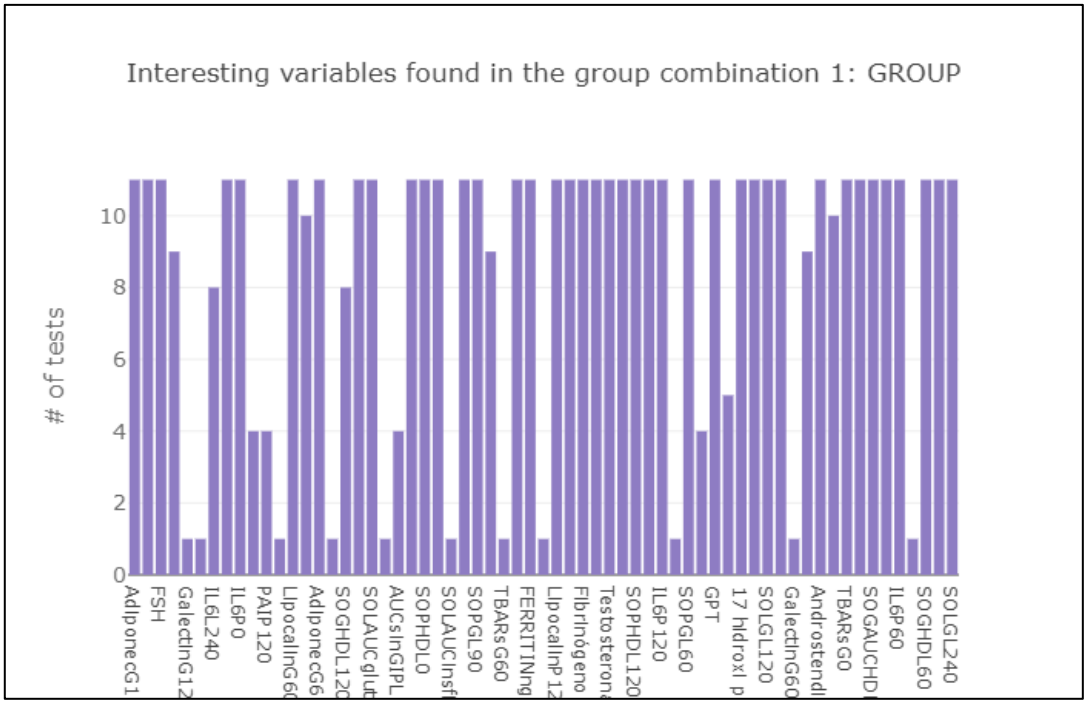


Figure 6. Interesting variables found vs # of post hoc tests passed

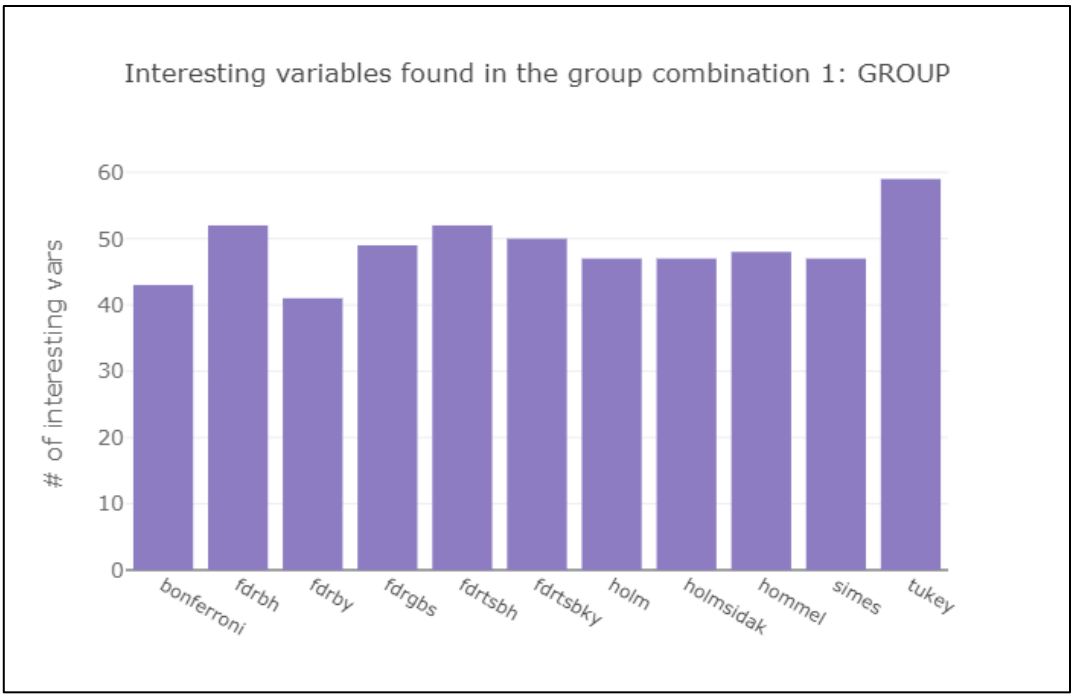


Figure 7. # of variables found by each post hoc test

We can easily see how Tukey test may be too loose on it conclusions, and it could indicate that is not a good fit for the data provided, however, the researcher must keep in mind that type II errors are also possible.

6. Considering all the post hoc tests, out of 67 variables marked by ANOVA as interesting, only 43 are validate as variables with significative statistical differences regarding the women with PCOS group value. This is the variable determination, where the results from the different post hoc tests are validated, by selecting only those variables that show to have statistical significative differences on the values previously selected by the researcher.

   The tool does offer a "more information", option, where the researcher can check for a specific variable, the results from the selected post hoc tests.

   An example of this is shown for the **AUCsinGIPL** variable, in the Figure 8.

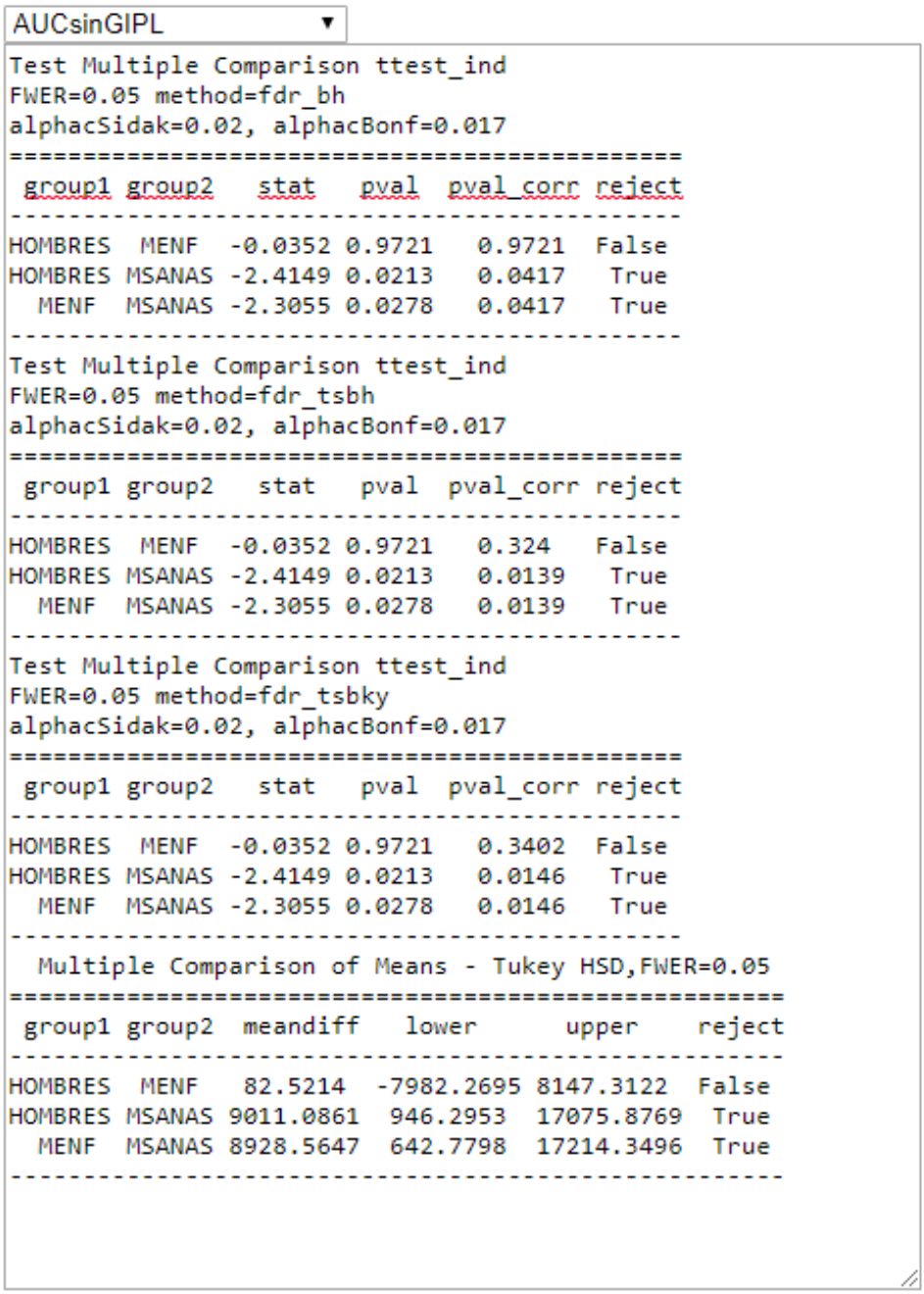


Figure 8. AUCsinGIPL variable post hoc analysis

It is interesting to note how 4 different post hoc tests (Tukey, fdr_bh, fdr_tsbh and fdr_tsbky), detected significative statistical differences between the group values *men* and *healthy women*, and also between *healthy women* and *women with the disorder*, but no differences between *men* and *women with the disorder*. This may indicate that the values for the *women with the disorder* are similar to *men*, as well as differences are found between *healthy women* and the other two groups. Also, this is an expected result according to previous studies of the metabolic disorder.

7. Once we have a final set of variables, all is left is to build a classification machine that can classify a patient according to these variables.
   After running the automated tool selecting the requisite that the variables to be used must have been validated by all 11 post hoc tests, the tool found a degree 4 polynomial kernel, with a C value of 1 as the best suited for the classification.
   The result can be seen in the figure 9.

```
-----------------------------------------------------
Best parameters set found on development set:
-----------------------------------------------------
{'kernel': 'poly', 'C': 1, 'degree': 4}
-----------------------------------------------------
Detailed classification report:
-----------------------------------------------------
The model is trained on the full development set
-----------------------------------------------------
             precision    recall  f1-score   support

    HOMBRES       1.00      1.00      1.00        12
       MENF       1.00      1.00      1.00        12
     MSANAS       1.00      1.00      1.00        12

avg / total       1.00      1.00      1.00        36

-----------------------------------------------------
The scores are computed on the full evaluation set
-----------------------------------------------------
             precision    recall  f1-score   support

    HOMBRES       1.00      1.00      1.00         2
       MENF       0.50      0.50      0.50         2
     MSANAS       0.00      0.00      0.00         1

avg / total       0.60      0.60      0.60         5

-----------------------------------------------------
```

Figure 9. Results after SVM training on those variables that are vaidated by all 11 tests

RESULTS

An important assumption was done while doing the data analysis. It was assumed that the data comes from a normal distribution population. This is a requirement for ANOVA, but due to the fact that we have few patients on each subgroup, the normal distribution validation cannot be trusted (it requires at least 20 elements in the sample). Hence, the application developed offers an option, where the user can enforce this normal distribution check (in the case of future analysis, when the researchers get more data from patients), or by not setting that checkbox, assume that the data is normal and normal distribution validation is not performed.
The results shown do not apply a normal distribution check.
The researchers from the Ramón y Cajal Hospital, have a special interest in the analysis of the results that comes from joining the **GROUP** and **OBESE** variables. This is because, being a metabolic disorder, polycystic ovary syndrome, PCOS, the condition being studied, it is natural to analyse the OBESE status in the different base groups.
Nevertheless, more discrete variables were found during the basic data analysis and are shown in the current results.

Discrete variables

This is the list of discrete variables found in the data:

- GROUP
- AFHA
- AFOB
- AFDM
- AFDL
- AFCVD
- AFHTA
- OBESE

The base group variable used to get the results shown in this document is **GROUP**. This variable represents 3 basic groups, as it has been explained before.
Using this base variable group, and following the guidance from the researchers, the **OBESE** variable was selected to be included in the full analysis. This, creates the following combinations for analysis (note that **GROUP** is added also as a combination alone):

- GROUP
- GROUP - OBESE

Afterwards, for each combination to be analysed, only the variable values that included the women with the disorder group, are included in the analysis (see Figure 4 and 10):

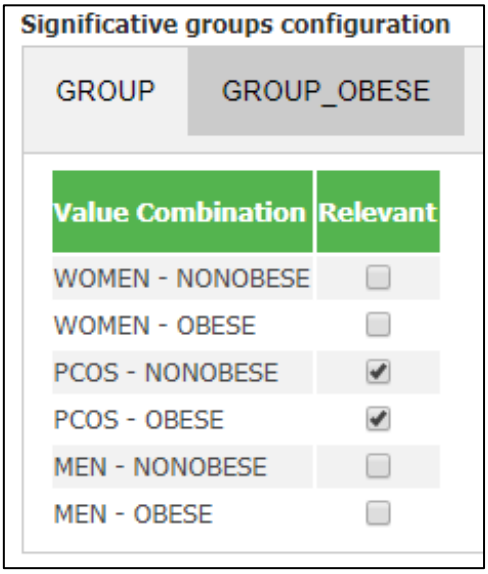


Figure 10. GROUP - OBESE significative groups configuration

To finish the analysis configuration, all post hoc test were selected.

Interesting Variables

After running the analysis, table 5 show the number of interesting variables found for each variable combination.

Combination	# of interesting variables found
GROUP	63
GROUP-OBESE	96

Table 5. General results: total variables found

However, by looking at the number of variables considered as interesting by each post hoc, it can be seen that Tukey test is way too permissive (figure 7 for **GROUP** analysis and 11 for **GROUP-OBESE** analysis), therefore, another analysis is run, where Tukey tests is not included, which means that then number of variables detected has changed (table 6).

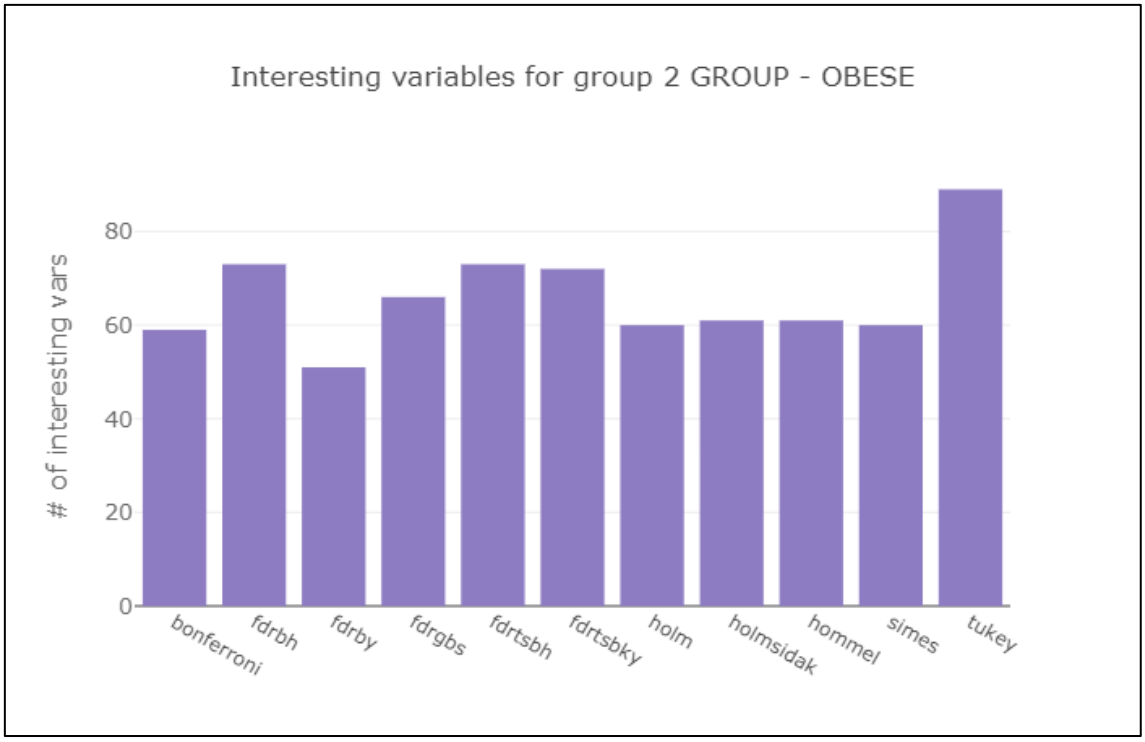


Figure 11. # of variables found by each post hoc test in GROUP-OBESE combination analysis

Combination	# of interesting variables found
GROUP	52
GROUP-OBESE	74

Table 6. General results: total variables found without Tukey test

Now, it is interesting to check which variables have been set as interesting by the most post hoc test now that Tukey is not being included (figures 12 and 13). And in the figures 14 and 15, a graph can be seen about how many variables where detected vs the number of post passed for each group combination.

Lastly, a table with the 37 variables that were detected as interesting in both group combinations, and the number of tests passed in each combination, is shown in table 7.

All of these different graphs and tables will be further analysed below.

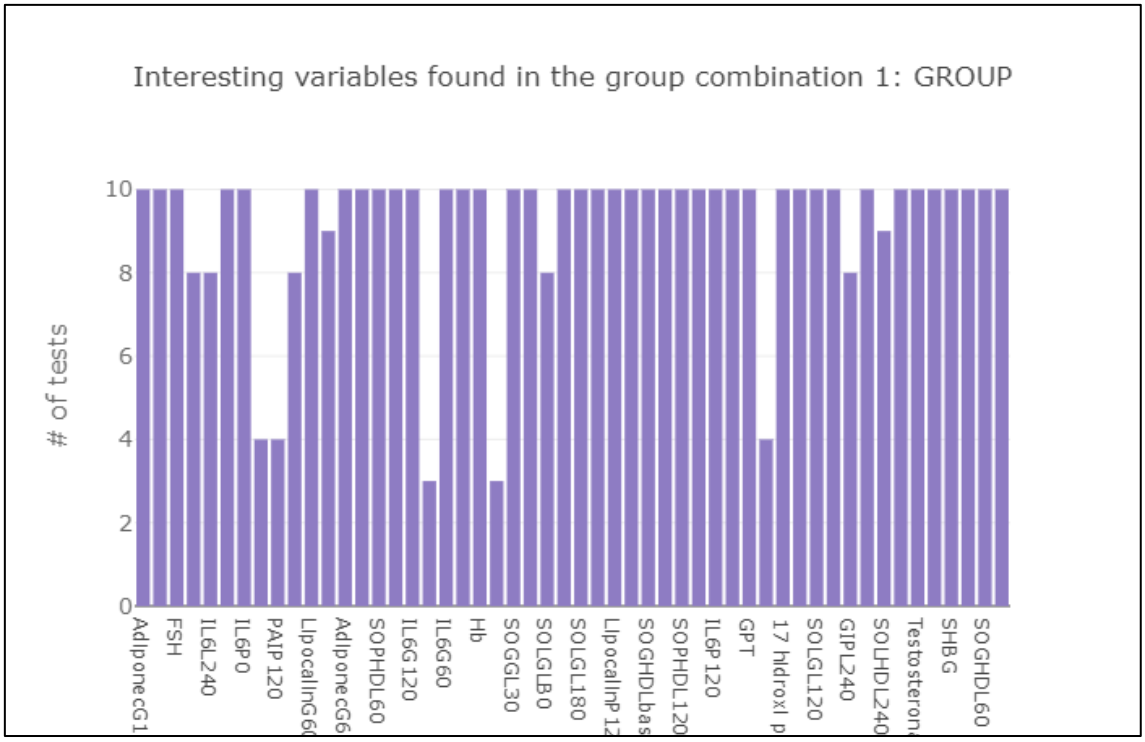


Figure 12. Detected interesting variables combination GROUP

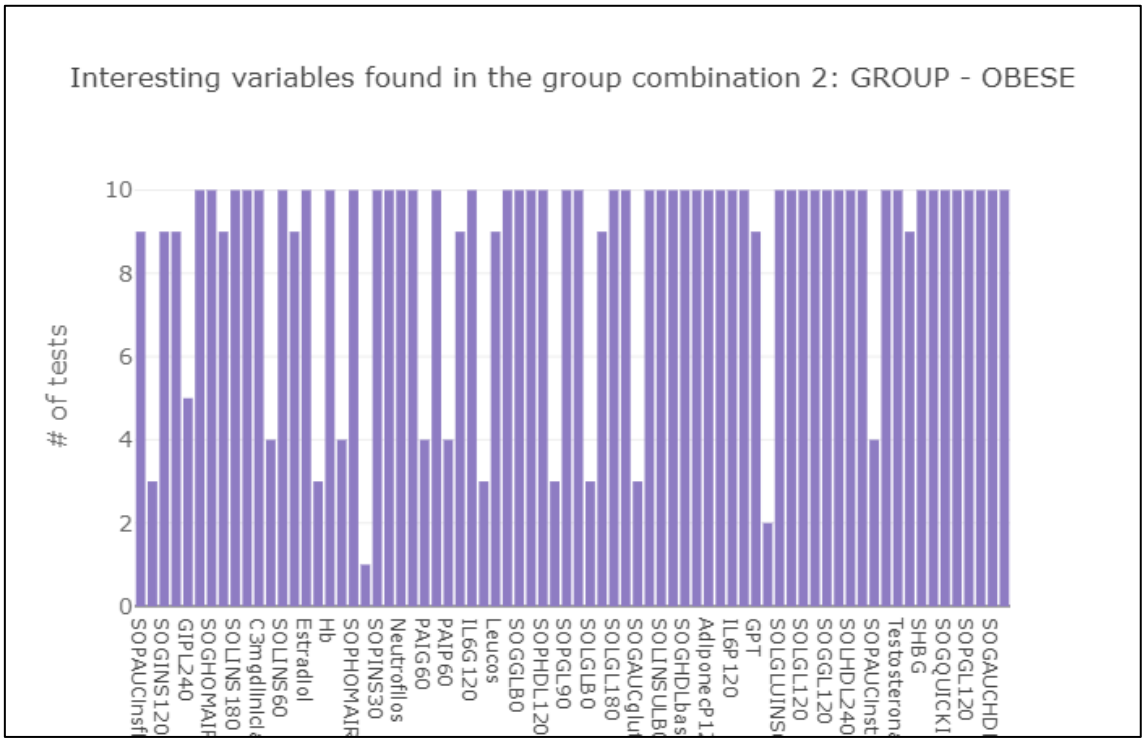


Figure 13. Detected interesting variables combination GROUP - OBESE

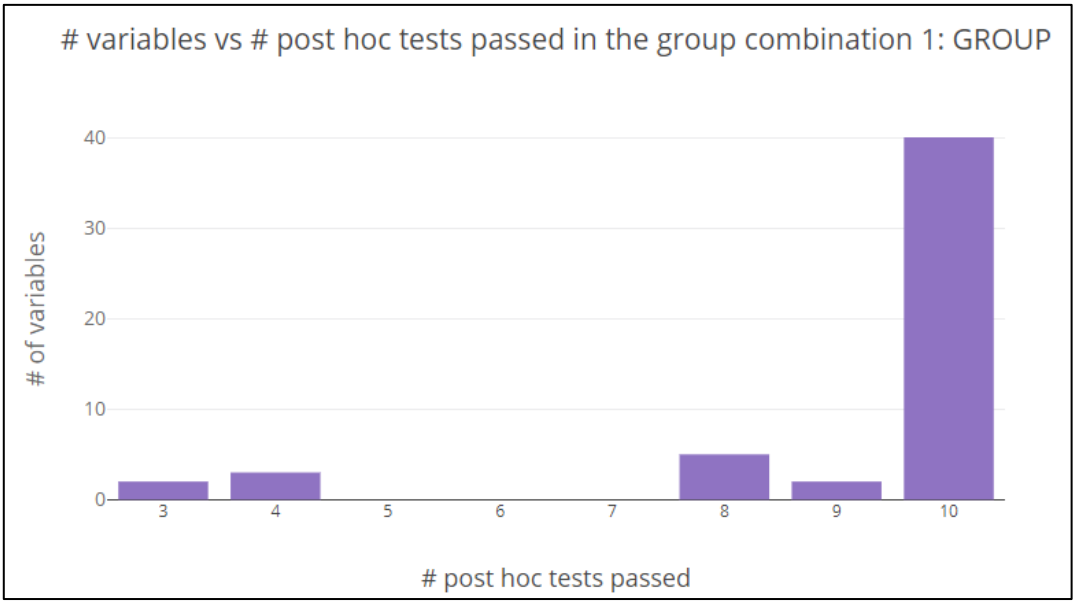


Figure 14. Total interesting variables vs number of post hoc tests passed, combination GROUP

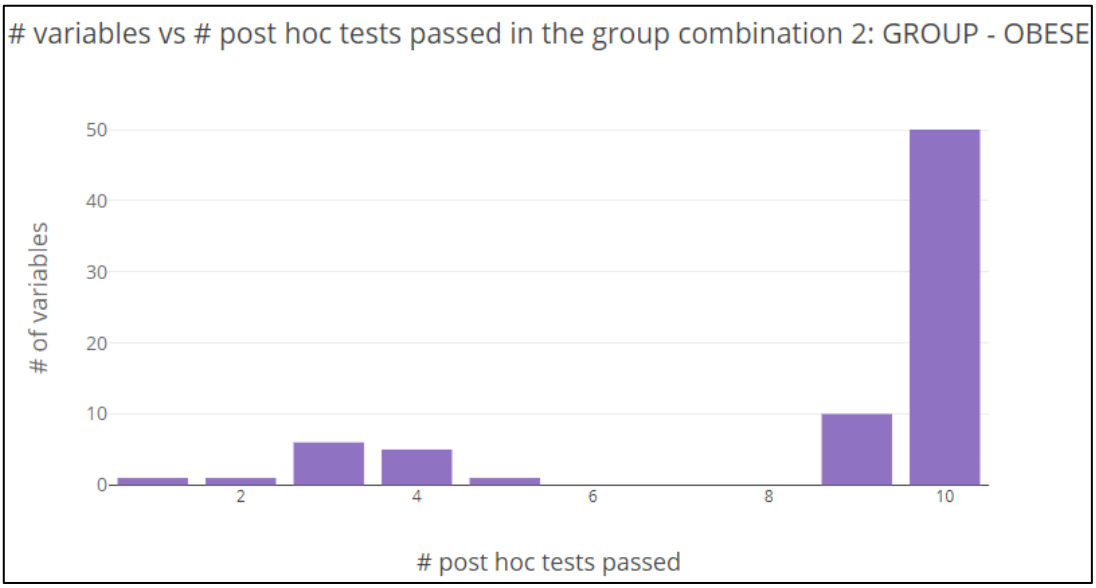


Figure 15. Total interesting variables vs number of post hoc tests passed, combination GROUP - OBESE

Variable	GROUP	GROUP - OBESE
Androstendiona0	10	10
Estradiol	4	10
Fibrinógeno	10	10
GIPL240	8	5
GPT	10	9
Hb	10	10
IL6G120	10	10
IL6G60	10	10

IL6P0	10	9
IL6P120	10	10
IL6P60	10	10
PAIP120	4	9
SHBG	10	10
SOGAUCHDLtotal	10	10
SOGGL30	10	3
SOGGLB0	10	10
SOGHDL120	8	4
SOGHDL60	10	10
SOGHDLbasal0	10	10
SOLAUCglutotal	10	10
SOLGL120	10	10
SOLGL180	10	10
SOLGL60	10	10
SOLGLB0	8	3
SOLHDL240	9	10
SOPAUCglutotal	10	10
SOPAUCHDLtotal	10	10
SOPGL120	10	10
SOPGL30	9	1
SOPGL60	10	10
SOPGL90	10	10
SOPHDL0	10	10
SOPHDL120	10	10
SOPHDL60	10	10
TBARsG0	10	10
Testosterona libre	10	10
Testosterona total	10	10

Table 7. List of variables and number of post hoc tests passed detected in both analysis: GROUP and GROUP - OBESE

Further Analysis

With the data obtained so far, researches can't figure out why a variable is statistically significative.

This, leads to the deep analysis of the different group differences. For instance, in the case of the GROUP combination, which has 3 possible values, as it has been explained before, there is up to 3 pair values comparisons (PCOS – WOMEN, PCOS – MEN and WOMEN - MEN), which means up to 8 different cases when significative variables differences are included (see table 8).

Case	PCOS / WOMEN	PCOS / MEN	WOMEN / MEN
1	True	True	True
2	True	True	False
3	True	False	True
4	True	False	False
5	False	True	True
6	False	True	False
7	False	False	True
8	False	False	False

Table 8. Cases based on GROUP analysis.
True: Significative difference
False: No significative difference

Note: however, that due to the fact that in the current analysis a variable is considered interesting only when the PCOS group shows a significative difference with other value, then only the cases 1 up to 6 are relevant.

If this classification is followed and the variables are categorized based on these cases, the number of variables obtained on each case are shown on table 9.

Case	PCOS / WOMEN	PCOS / MEN	WOMEN / MEN	# Variables detected
1	True	True	True	3
2	True	True	False	3
3	True	False	True	2

4	True	False	False	1
5	False	True	True	25
6	False	True	False	18

Table 9. Number of variables per case on GROUP analysis

However, the case where PCOS and women comparison shows no statistical significative difference, but both PCOS – men and women – men, do show a significative difference, could be seen as normal cases (case number 5).

This last analysis, yields to only 27 variables detected as interesting, after a deep analysis for the combination GROUP.

These 27 variables are shown on table 10.

Variables	Case
AdiponecG120	6
AdiponecG60	6
Fibrinógeno	6
IL6P0	6
IL6P120	6
LipocalinG60	6
normAUCsinGIPL	6
SOGGL30	6
SOGGLB0	6
SOLAUCglutotal	6
SOLGL120	6
SOLGL60	6
SOLGLB0	6
SOPAUCHDLtotal	6
SOPGL120	6
SOPGL30	6
SOPGL60	6
SOPGL90	6
AdipsinG0	4
AUCsinGIPL	2
GIPL240	2

Androstendiona0	3
GIPP120	3
TBARsG0	3
IL6P60	1
Testosterona libre	1
Testosterona total	1

Table 10. List of variables and case which they satisfy on deep analysis, combination GROUP

Similarly, for the GROUP – OBESE combination, although the number of combinations for 15 values pairs (table 11), can go up to 32768 (2^{15}), due to the fact that only 74 were found, is easier to first show in the appendix I the number of variables per different case type.

PCOS – NONOBESE	PCOS – OBESE
PCOS – NONOBESE	WOMEN_NONOBESE
PCOS – NONOBESE	WOMEN_OBESE
PCOS – NONOBESE	MEN_NONOBESE
PCOS – NONOBESE	MEN_OBESE
PCOS – OBESE	WOMEN_NONOBESE
PCOS – OBESE	WOMEN_OBESE
PCOS – OBESE	MEN_NONOBESE
PCOS – OBESE	MEN_OBESE
WOMEN_NONOBESE	WOMEN_OBESE
WOMEN_NONOBESE	MEN_NONOBESE
WOMEN_NONOBESE	MEN_OBESE
WOMEN_OBESE	MEN_NONOBESE
WOMEN_OBESE	MEN_OBESE
MEN_NONOBESE	MEN_OBESE

Table 11. Value pairs for combination GROUP - OBESE

After searching for the different cases in these 74 variables, 42 different cases were identified. The results in this case, are shown in the appendix I, together with the full list of variables and which case each variable belongs to.

This last analysis, shows how difficult can it be to identify interesting cases for the researchers, although once identified, it's easy to find the variables that holds for each case, hence guiding the researcher into confirming or finding new relationships between parameters and the study being done.

A final note should be highlighted: after applying the full analysis, including all the discrete variables but enforcing a normal distribution verification, only the GROUP combination

(without including any other discrete variable), passed the normal distribution validation. Moreover, only 5 variables were found.
In the table 12, the list of variables and the case to which each belong is shown.

Variables	Case
17 hidroxi progesterona0	5
AdiponecG60	6
AdiponecG120	6
AdiponecP60	5
GPT	5

Table 12. List of variables and case which they satisfy on deep analysis, enforcing normal distribution, combination GROUP

Just as before, case 5 could be considered as normal, hence only variables **AdiponecG60** and **AdiponecG120**, could be selected as interesting after passing all the possible validations (normal distributions, ANOVA, post hoc tests, interesting case classification).
The graphs regarding the post hoc tests for this case is shown on figure 16.

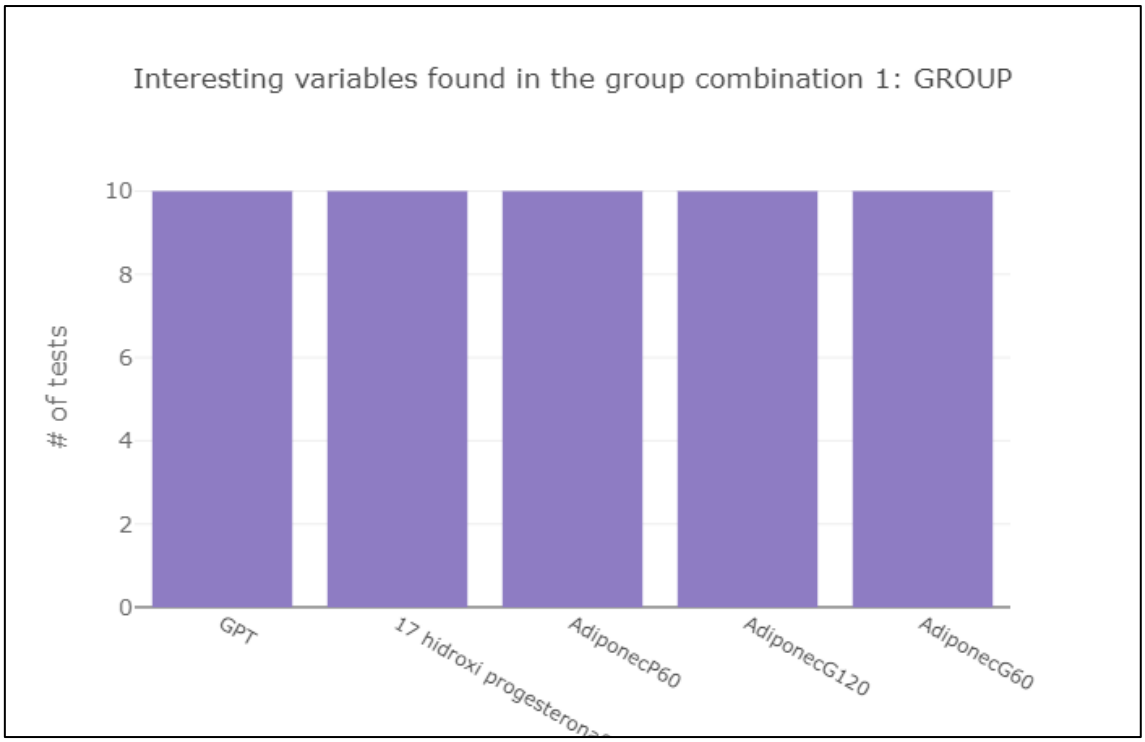


Figure 16. Number of tests that verify each variable, GROUP combination enforcing normal distribution

Clearly, the lack of more statistically significative variables is caused by the small number of patients, which prevent the correct validation of normal distributions, although after the validation, the results are much more significative.

CONCLUSIONS

After analysing the results, the following statements have been concluded.

- The metabolic disorder research is quite interesting, and even though the lack of patients in the initial data do affect the results and medical conclusions that can be drawn, we continued with the work, building a tool that can help the doctors by running a full statistical procedure even when the data doesn't fulfil the requirements for ANOVA. The initial results from the procedure, are along the lines of the medical research. This can already help the doctors by giving them an initial validation, that the data seems to be in line with their studies and that can be validated later, once the researchers have more data to run the analysis again.

- Also, the fact that there is no other clear discrete variable to do the analysis, except for the **GROUP** and **OBESE** variables, may add combinations that are not relevant for the metabolic disorder. For example, if the researcher decides to carry on an analysis on other detected discrete variables like AFHA, could show results that are not relevant, specially when the lack of data may affect the results.
  Researchers must be aware of this, and work with data sets that do exclude those discrete variables which are not potentially relevant for the study. The application by itself, can not do this medical assumption. Again, extreme caution is recommended when working with data that do not fulfil the conditions for ANOVA tests.

- A classification machine based on non-conclusive variables may lead to false classification. It is needed then, to first validate the results before trusting the classification machine. Also, part of the classification machine validation, includes training and testing, which are better achieved when there is enough data that represents as much as possible the population.
  In the metabolic disorder scenario, a lack of patients may yield to a SVM configuration that may not be as accurate as expected.
  However, when the doctors get more data to feed the system, the tool will be ready to handle and analyse the data and produce a more accurate SVM system.

- The potential of the study lies on the fact that a researcher trying to find out if its data matches a specific conclusion (there is a statistical technique supporting the conclusion), and whether an automated machine can classify incoming data based on that conclusion (by choosing the variables that seems to support the conclusion). The system will do just that, making it easier for a medical research to first, validate its data, second, specify

study values, and third, make it easier to determine a machine that will do the classification fully automated for any new data.

- María Insenser, researcher from the Ramón y Cajal Hospital, added the following conclusion, after evaluating the results obtained from the procedure and tool:

  The variables detected as interesting in the GROUP and GROUP - OBESE analysis support the usefulness of the tool from researcher's point of view. As expected from the study design, the sexual steroids (testosterone libre and testosterone total) and SHBG levels (sex hormone binding globulin) showed a significant difference between groups. Moreover, variables related to inflammation as IL-6 (Interleukine-6) or oxidative stress as TBARS (Thiobarbituric acid reactive substances) could be implicated in physio-pathological mechanisms of the disorder. The potential relevant variables identified with this tool are a valuable resource of knowledge to understand the etiology of this metabolic disorder.

REFERENCES

1. Polycystic ovary syndrome: a complex condition with psychological, reproductive and metabolic manifestations that impacts on health across the lifespan. H, Teede, A, Deeks and L, Moran. 2010, BMC Med., Vol. 8, p. 41.

2. About Polycystic Ovary Syndrome. [Online] 2013. [Cited: 04 02 2018.] https://www.nichd.nih.gov/health/topics/pcos/conditioninfo.

3. HK, Ramakrishna. *Medical Statistics.* Shivamogga : Springer, 2017.

4. Roth-Johnson, Liz. Introduction to Descriptive Statistics: Using Mean, Median, and Standard Deviation. [Online] 2015. https://www.visionlearning.com/en/library/Math-in-Science/62/Introduction-to-Descriptive-Statistics/218/reading.

5. *Critical review of the state of medicine during the last ten years.* Black, A. and C. 1817, Edinburgh Medical and Surgical Journal, pp. 1-68.

6. Jayadevan, Rajeev. onmanorama.com. [Online] 2016. http://english.manoramaonline.com/wellness/health/everyday-health-going-for-a-lab-test-you-need-to-read-this-first.html.

7. Huang, H. https://www.stat.berkeley.edu/~hhuang/. *Multiple Hypothesis Testing and False Discovery Rate.* [Online] https://www.stat.berkeley.edu/~hhuang/STAT141/Lecture-FDR.pdf.

8. Weisstein, Eric W. MathWorld--A Wolfram Web Resource. *Fisher's Exact Test.* [Online] http://mathworld.wolfram.com/FishersExactTest.html.

9. *On the interpretation of χ2 from contingency tables, and the calculation of P.* Fisher, R. A. 1922, Journal of the Royal Statistical Society 85, pp. 87-94.

10. Fisher, R.A. *Statistical Methods for Research Workers.* s.l. : Oliver and Boyd, 1954.

11. Stephanie. Statistics How To. [Online] 2018. http://www.statisticshowto.com/probability-and-statistics/t-test/.

12. *On a test of whether one of two random variables is stochastically larger than the other.* Mann, H. B. and Whitney, D. R. 1947, Ann. Math. Statist. 18, pp. 50-60.

13. Wilcoxon, Frank. Individual comparisons by ranking methods. *Biometrics 1.* 1945, pp. 80-83.

14. Peck, Roxy, Olsen, Chris and Devore, Jay L. *Introduction to Statistics & Data Analysis.* s.l. : Cengage Learning, 2014.

15. Baker, Greg. CMPT 318. *Statistical Tests.* [Online] 2017. http://www.cs.sfu.ca/~ggbaker/data-science/content/stats-tests.html.

16. *Statistical notes for clinical researchers: post-hoc multiple comparisons.* Kim, Hae-Young. 2015, RDE Restorative Dentistry & Endodontics, pp. 172-176.

17. Stephanie. Statistics How To. [Online] 2017. http://www.statisticshowto.com/tukey-test-honest-significant-difference/.

18. Stevens, Joseph J. *Post Hoc Tests in ANOVA.* Eugene, Oregon, United States : s.n., 1999.

19. *Teoria statistica delle classi e calcolo delle probabilità.* Bonferroni, C. E. Firenze : s.n., 1936, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

20. *Multiple Comparisons Among Means.* Dunn, Olive Jean. 293, 1961, Journal of the American Statistical Association, Vol. 56, págs. 52-64.

21. *A simple sequential rejective multiple test procedure.* Holm, S. 1979, Scandinavian Journal of Statistics 6, pp. 65-70.

22. *Multiple significance tests: the Bonferroni method.* Bland, J Martin y Altman, Douglas G. London : s.n., 1995, Vol. 310.

23. *Rectangular Confidence Regions for the Means of Multivariate Normal Distributions.* Šidák, Zbyněk. 318, 1967, Journal of the American Statistical Association, Vol. 62, págs. 626-633.

24. *Holm's Sequential Bonferroni Procedure.* Abdi, Hervé. 2010, Encyclopedia of Research Design.

25. GraphPad Software. The Holm-Šídák approach to multiple comparisons. [En línea] https://www.graphpad.com/guides/prism/7/statistics/index.htm?stat_holms_multiple_comparison_test.htm.

26. *An improved Bonferroni procedure for multiple tests of significance.* Simes, R. J. 1986, Biometrika 73, pp. 751-754.

27. *A Sharper Bonferroni Procedure for Multiple Tests of Significance.* Hochberg, Yosef. 4, 1988, Biometrika, Vol. 75, pp. 800-802.

28. *A stagewise rejective multiple test procedure based on a modified Bonferroni test.* Hommel, G. 2, Mainz : s.n., 1988, Biometrika, Vol. 75, pp. 383-386.

29. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Benjamini, Yoav and Hochberg, Yosef. 1995, Journal of the Royal Statistical Society. Series B (Methodological), pp. 289-300.

30. *Multiple hypothesis testing.* J.P., Shaffer. 1995, Annual Review of Psychology, Vol. 46, pp. 561-584.

31. *The control of the false discovery rate in multiple testing under dependency.* Benjamini, Yoav and Yekutieli, Daniel. 2001, The Annals of Statistics, pp. 1165-1188.

32. *Adaptive linear step-up procedures that control the false.* Benjamini, Yoav, Krieger, Abba M. and Yekutieli, Daniel. 2006, Biometrika, pp. 491-507.

33. *An adaptive step-down procedure with proven FDR control under independence.* Gavrilov, Yulia, Benjamini, Yoav and Sarkar, Sanat K. 2009, The Annals of Statistics Vol. 37, No. 2, pp. 619-629.

34. *A tutorial on support vector machines for pattern recognition.* Burges, Christopher J. C. 1998, Data Mining and Knowledge Discovery, Vol. 2, pp. 121-167.

35. *An Improved Training Algorithm for Support Vector Machines.* Osuna, Edgar, Freund, Robert and Girosi, Federico. 1997, pp. 276-285.

36. *Training Support Vector Machines: an Application to Face Detection.* Osuna, Edgar, Freund, Robert and Girosi, Federico. 1997, pp. 130-136.

37. Vapnik, Vladimir N. *The Nature of Statistical Learning Theory.* London : Springer-Verlag,, 1995.

38. Vert, Jean-Philippe, Tsuda, Koji and Schölkopf, Bernhard. *Kernel Methods in Computational Biology.* s.l. : MIT Press, 2004. pp. 35-70.

39. *A tutorial on support vector regression.* Smola, Alex J. and Bernhard Schölkopf. 2004, Statistics and Computing, Vol. 14, pp. 199–222.

40. Chang, Chih-Chung and Lin, Chih-Jen. *LIBSVM: A Library for Support Vector Machines.* Taipei : s.n., 2013.

41. *SVM Classification an Approach on Detecting Abnormality in Brain MRI Images.* Kumari, R. 2013, International Journal of Engineering Research and Applications, pp. 1686-1690.

42. *Protein subcellular localization of fluorescence microscopy images: Employing new statistical and Texton based image features and SVM based ensemble classification.* Tahir, M. and Khan, A. 2016, InformationSciences, Vol. 345, pp. 65-80.

43. *Probability estimates for multi-class classification.* Wu, T.-F., Lin, C.-J. and Weng, R. C. 2004, . Journal of Machine Learning Research, Vol. 5, pp. 975–1005.

44. Schölkopf, B. *Support Vector Learning.* 1997.

45. Thompson, Steven K. *Sampling.* s.l. : Wiley, 2012.

46. Python Software Foundation. Python. [Online] https://www.python.org/.

47. Pandas. Pandas documentation. [Online] https://pandas.pydata.org/pandas-docs/stable/generated/pandas.to_numeric.html.

48. Ronacher, Armin. Flask. [Online] http://flask.pocoo.org/.

49. Python Software Foundation. itertools — Functions creating iterators for efficient looping. [Online] https://docs.python.org/2/library/itertools.html.

50. McKinney, Wes. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference.* 2010. pp. 51-56.

51. E, Jones, et al. SciPy: Open Source Scientific Tools for Python. [Online] 2001. [Cited: 04 02 2018.] https://www.scipy.org/.

52. Perktold, Josef, Seabold, Skipper and Taylor, Jonathan. Statsmodels: statistics in python. [Online] 2009. http://www.statsmodels.org/stable/index.html.

53. *Scikit-learn: Machine Learning in Python.* Pedregosa and al, et. 2011, JMLR 12, pp. 2825-2830.

APPENDIX I

GROUP AND GROUP-OBESE ANALYSIS RESULTS

Value pairs with statistical significative difference		# Variables detected	Variables detected
PCOS – NONOBESE	PCOS – OBESE	1	SOLINS60
PCOS – OBESE	MEN_NONOBESE		
PCOS – NONOBESE	PCOS – OBESE	1	SOLINS120
PCOS – OBESE	MEN_NONOBESE		
WOMEN_OBESE	MEN_NONOBESE		
MEN_NONOBESE	MEN_OBESE		
PCOS – NONOBESE	PCOS – OBESE	1	C3mgdlinicial
PCOS – OBESE	WOMEN_NONOBESE		
PCOS – OBESE	MEN_NONOBESE		
PCOS – NONOBESE	PCOS – OBESE	1	SOLAUCinstotal
PCOS – OBESE	WOMEN_NONOBESE		
PCOS – OBESE	MEN_NONOBESE		
MEN_NONOBESE	MEN_OBESE		
PCOS – NONOBESE	PCOS – OBESE	1	SOLHOMAIR
PCOS – OBESE	WOMEN_NONOBESE		
PCOS – OBESE	MEN_NONOBESE		
WOMEN_OBESE	MEN_NONOBESE		
PCOS – NONOBESE	PCOS – OBESE	1	SOLINSULB0
PCOS – OBESE	WOMEN_NONOBESE		
PCOS – OBESE	WOMEN_OBESE		
PCOS – OBESE	MEN_NONOBESE		
WOMEN_OBESE	MEN_NONOBESE		

PCOS – NONOBESE	PCOS – OBESE	1	SOGTG120
PCOS – NONOBESE	MEN_OBESE		
PCOS – OBESE	WOMEN_NONOBESE		
WOMEN_NONOBESE	MEN_OBESE		
MEN_NONOBESE	MEN_OBESE		
PCOS – NONOBESE	PCOS – OBESE	1	SOPISIcomp
PCOS – NONOBESE	WOMEN_OBESE		
PCOS – NONOBESE	MEN_OBESE		
PCOS – OBESE	WOMEN_NONOBESE		
PCOS – OBESE	MEN_NONOBESE		
WOMEN_NONOBESE	WOMEN_OBESE		
WOMEN_NONOBESE	MEN_OBESE		
WOMEN_OBESE	MEN_NONOBESE		
MEN_NONOBESE	MEN_OBESE		
PCOS – NONOBESE	PCOS – OBESE	1	SOGGL120
PCOS – NONOBESE	WOMEN_OBESE		
PCOS – NONOBESE	MEN_NONOBESE		
PCOS – NONOBESE	MEN_OBESE		
PCOS – OBESE	WOMEN_NONOBESE		
PCOS – NONOBESE	WOMEN_NONOBESE	1	GIPL240
PCOS – OBESE	WOMEN_NONOBESE		
WOMEN_NONOBESE	MEN_NONOBESE		
WOMEN_NONOBESE	MEN_OBESE		
PCOS – NONOBESE	WOMEN_NONOBESE	2	Testosterona libre Testosterona total
PCOS – NONOBESE	MEN_NONOBESE		
PCOS – NONOBESE	MEN_OBESE		
PCOS – OBESE	MEN_NONOBESE		
PCOS – OBESE	MEN_OBESE		
WOMEN_NONOBESE	MEN_NONOBESE		
WOMEN_NONOBESE	MEN_OBESE		
WOMEN_OBESE	MEN_NONOBESE		
WOMEN_OBESE	MEN_OBESE		
PCOS – NONOBESE	WOMEN_OBESE	1	SOGGLB0
WOMEN_NONOBESE	WOMEN_OBESE		

PCOS – NONOBESE	WOMEN_OBESE	2	PAIL120 PAIP60
WOMEN_NONOBESE	WOMEN_OBESE		
WOMEN_OBESE	MEN_NONOBESE		
PCOS – NONOBESE	WOMEN_OBESE	1	SOPINS30
WOMEN_NONOBESE	WOMEN_OBESE		
WOMEN_OBESE	MEN_NONOBESE		
WOMEN_OBESE	MEN_OBESE		
PCOS – NONOBESE	WOMEN_OBESE	1	SOGQUICKI
PCOS – OBESE	MEN_NONOBESE		
WOMEN_OBESE	MEN_NONOBESE		
MEN_NONOBESE	MEN_OBESE		
PCOS – NONOBESE	WOMEN_OBESE	1	SOPAUCinstotal
PCOS – OBESE	WOMEN_NONOBESE		
PCOS – OBESE	MEN_NONOBESE		
WOMEN_NONOBESE	WOMEN_OBESE		
WOMEN_NONOBESE	MEN_OBESE		
WOMEN_OBESE	MEN_NONOBESE		
WOMEN_OBESE	MEN_OBESE		
PCOS – NONOBESE	WOMEN_OBESE	1	SOLGLB0
PCOS – NONOBESE	MEN_OBESE		
PCOS – NONOBESE	MEN_NONOBESE	2	AdiponecP120 Androstendiona0
PCOS – NONOBESE	MEN_NONOBESE	1	PAIG60
PCOS – OBESE	MEN_NONOBESE		
WOMEN_NONOBESE	WOMEN_OBESE		
WOMEN_OBESE	MEN_NONOBESE		
PCOS – NONOBESE	MEN_NONOBESE	1	Hb
PCOS – NONOBESE	MEN_OBESE		
PCOS – OBESE	MEN_NONOBESE		
PCOS – OBESE	MEN_OBESE		
WOMEN_NONOBESE	MEN_NONOBESE		
WOMEN_NONOBESE	MEN_OBESE		
WOMEN_OBESE	MEN_OBESE		

PCOS – NONOBESE MEN_OBESE	5	SOGGL60 SOPAUCglutotal SOPGL30 SOPGL90 SOPGLB0
PCOS – NONOBESE MEN_OBESE WOMEN_OBESE MEN_OBESE	1	SHBG
PCOS – NONOBESE MEN_OBESE WOMEN_NONOBESE MEN_OBESE	9	SOGAUCglutotal SOLAUCglutotal SOLGL120 SOLGL180 SOLGL60 SOPAUCHDLtotal SOPHDL120 SOPHDL60 TBARsG60
PCOS – NONOBESE MEN_OBESE WOMEN_NONOBESE MEN_OBESE MEN_NONOBESE MEN_OBESE	2	GPT SOGHDLbasal0
PCOS – NONOBESE MEN_OBESE WOMEN_NONOBESE MEN_OBESE WOMEN_OBESE MEN_OBESE	2	SOGHDL60 SOPHDL0
PCOS – NONOBESE MEN_OBESE WOMEN_NONOBESE MEN_OBESE WOMEN_OBESE MEN_OBESE MEN_NONOBESE MEN_OBESE	2	SOGAUCHDLtotal SOLHDL240
PCOS – NONOBESE MEN_OBESE PCOS – OBESE MEN_OBESE	1	SOPGL120
PCOS – NONOBESE MEN_OBESE PCOS – OBESE MEN_OBESE MEN_NONOBESE MEN_OBESE	1	TBARsG0
PCOS – NONOBESE MEN_OBESE PCOS – OBESE MEN_OBESE WOMEN_NONOBESE MEN_OBESE MEN_NONOBESE MEN_OBESE	1	SOPGL60

PCOS – NONOBESE	MEN_OBESE	1	SOGHDL120
PCOS – OBESE	MEN_OBESE		
WOMEN_NONOBESE	MEN_OBESE		
WOMEN_OBESE	MEN_OBESE		
MEN_NONOBESE	MEN_OBESE		
PCOS – OBESE	WOMEN_NONOBESE	1	IL6P0
PCOS – OBESE	MEN_OBESE		
PCOS – OBESE	WOMEN_NONOBESE	3	Fibrinógeno IL6P120 IL6P60
PCOS – OBESE	MEN_NONOBESE		
PCOS – OBESE	MEN_OBESE		
PCOS – OBESE	WOMEN_NONOBESE	1	IL6G120
PCOS – OBESE	MEN_NONOBESE		
PCOS – OBESE	MEN_OBESE		
WOMEN_OBESE	MEN_NONOBESE		
PCOS – OBESE	WOMEN_OBESE	1	PAIG120
WOMEN_NONOBESE	WOMEN_OBESE		
WOMEN_OBESE	MEN_NONOBESE		
PCOS – OBESE	MEN_NONOBESE	6	AdipsinP120 Leucos Neutrofilos SOGINS120 SOLGLUINS0 SOLINS240
PCOS – OBESE	MEN_NONOBESE	2	SOLINS180 SOPINS90
MEN_NONOBESE	MEN_OBESE		
PCOS – OBESE	MEN_NONOBESE	3	PAIP120 SOGISIcomp SOLISIcomp
WOMEN_OBESE	MEN_NONOBESE		
PCOS – OBESE	MEN_NONOBESE	5	SOGHOMAIR SOGINSULB0 SOPHOMAIR SOPINSULB0 TAD
WOMEN_OBESE	MEN_NONOBESE		
MEN_NONOBESE	MEN_OBESE		
PCOS – OBESE	MEN_NONOBESE	1	Estradiol
WOMEN_NONOBESE	MEN_NONOBESE		

PCOS – OBESE	MEN_NONOBESE	1	IL6G60
PCOS – OBESE	MEN_OBESE		
PCOS – OBESE	MEN_OBESE	1	SOGGL30
WOMEN_NONOBESE	WOMEN_OBESE		
WOMEN_NONOBESE	MEN_OBESE		
WOMEN_NONOBESE	WOMEN_OBESE	1	SOPAUCinsfinal
WOMEN_OBESE	MEN_NONOBESE		
WOMEN_OBESE	MEN_OBESE		

Table 13. Full results case analysis for GROUP - OBESE combination

APPENDIX II

WEB TOOL, USER GUIDE

The tool can be found on the following URL:

http://vjctfm.ddns.net:5000/

The landing page, serves as a small user guide, although detailed steps are explained on this appendix.

1. First, make sure to click on the ANALYSIS option, in order to start a new analysis.

2. Next, load your experiment data, using a CSV file format. Make sure that this file has headers, as they will denote the variable names. Each column in the CSV file is variable either discrete (that can be used to group the data), or continuous, which represents experiment results.
   After selecting the file from your filesystem, decide if the analysis will include a normal distribution validation (enforcing that ANOVA test is applied only if requirements are met), or if the analysis will be applied as a preliminary (and taking the results with extreme caution), and the requirements for ANOVA won't be enforced.
   Press the upload button, to load the file to the server, see figure 17.

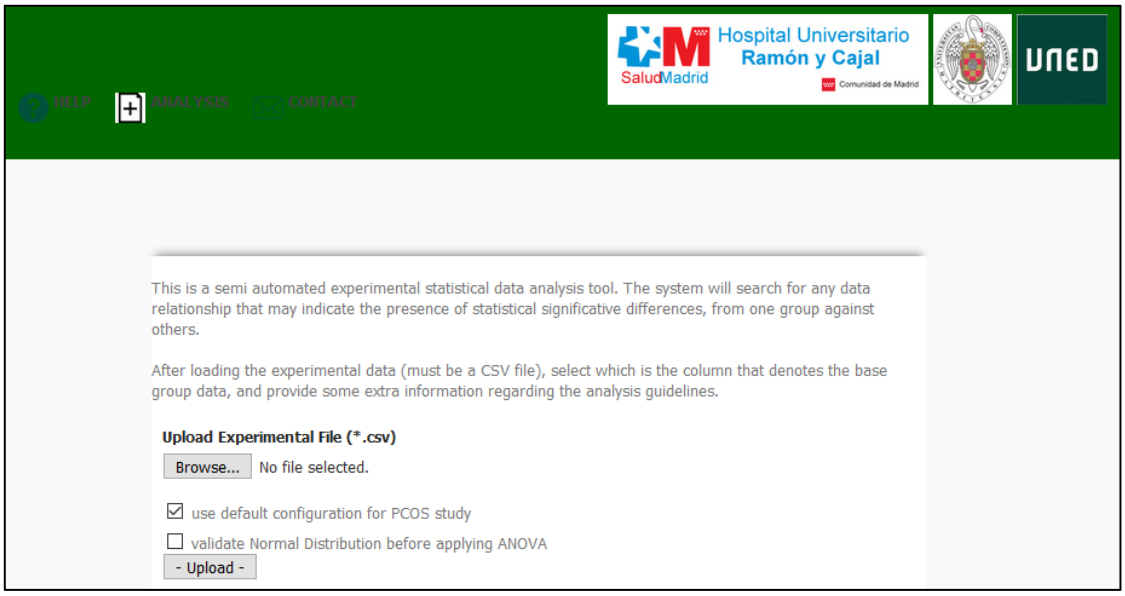


Figure 17. Data load and initial configuration screenshot

3. After the file is uploaded, data will be cleansed and discrete variables will be detected (figure 18). Now, select the base group variable and those variables that must be included in the analysis (from the procedure, this is step number two).

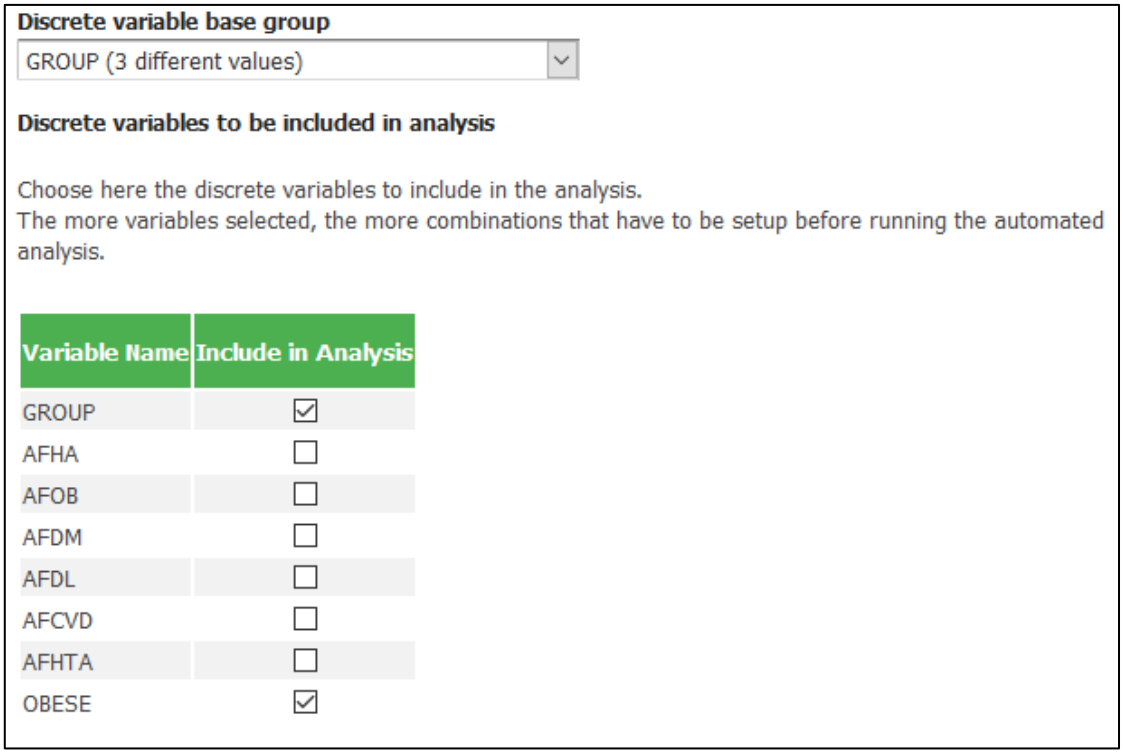


Figure 18. Discrete variables configuration

4. Based on the variables selected on the previous step, the group combinations will be added. Configure each combination, by using the "Significative combinations, values configuration" option. Each combination will be shown as a tab in a toolbar, and the value combinations will appear below so that only those relevant for the analysis can be selected. On figure 19 an example can be seen.

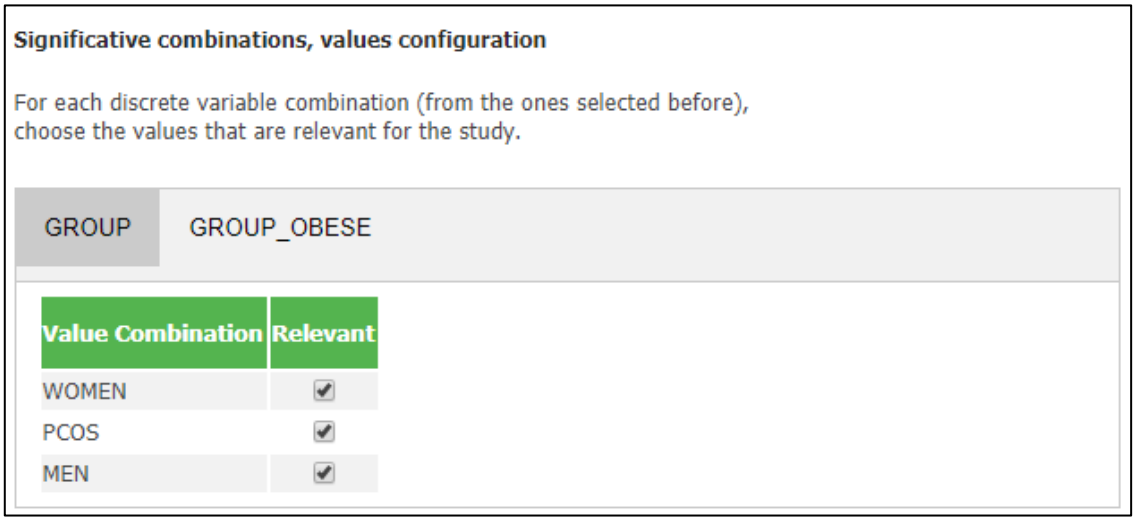


Figure 19. Significative combinations, values configuration

5. Finally, before starting the analysis, select those post hoc tests that will be applied to validate ANOVA results. The results from these tests, together with the values selected as relevant on the previous step, will determine the variables that will be flagged as interesting. On figure 20, this test selection screen can be seen.

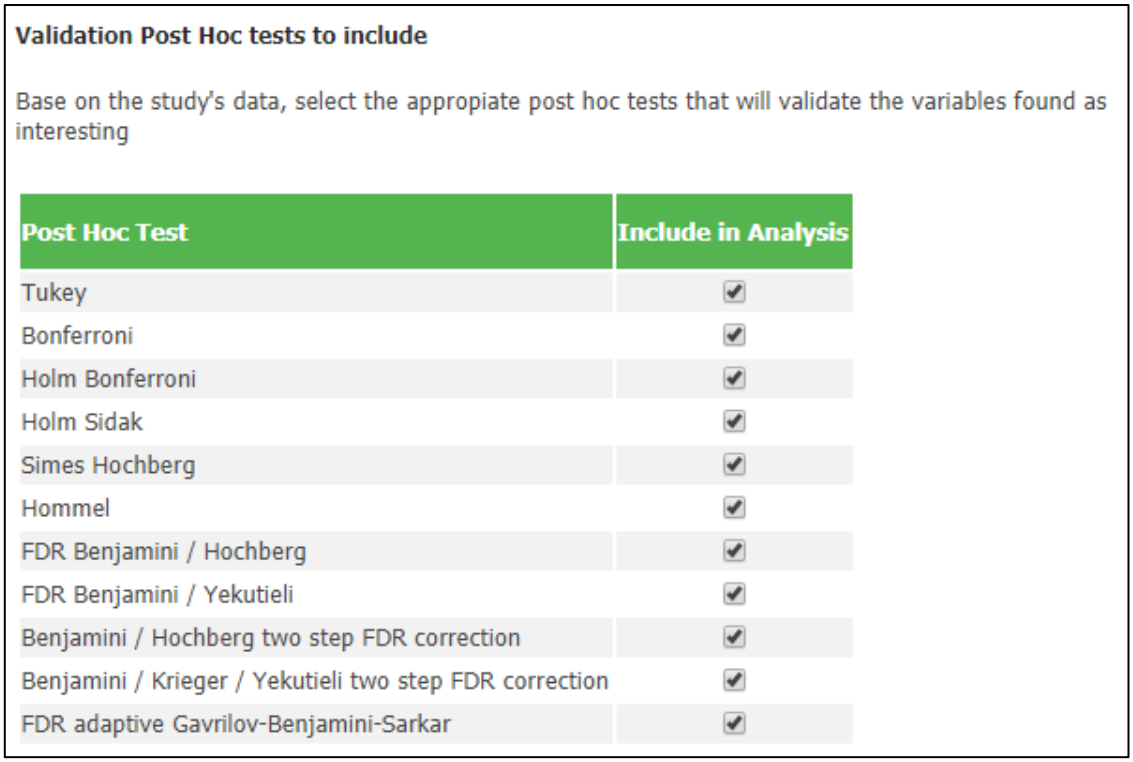


Figure 20. Post hoc test selection

6. Press the Analysis button, to start the analysis. A progress bar, together with a status text is shown. When the analysis is done, a list with the interesting variables detected for each combination, will appear on the results text area (figure 21).
   A detailed analysis result can be seen on the variable analysis screen, where for each analysed combination, each variable detailed result is displayed (figure 22).

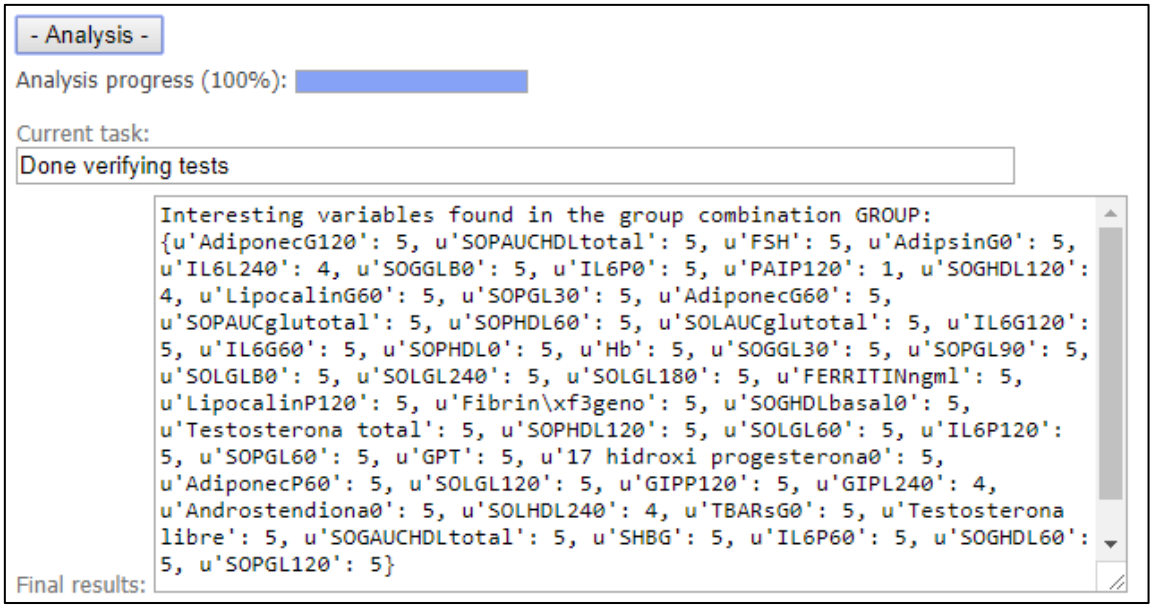


Figure 21. Analysis and results display

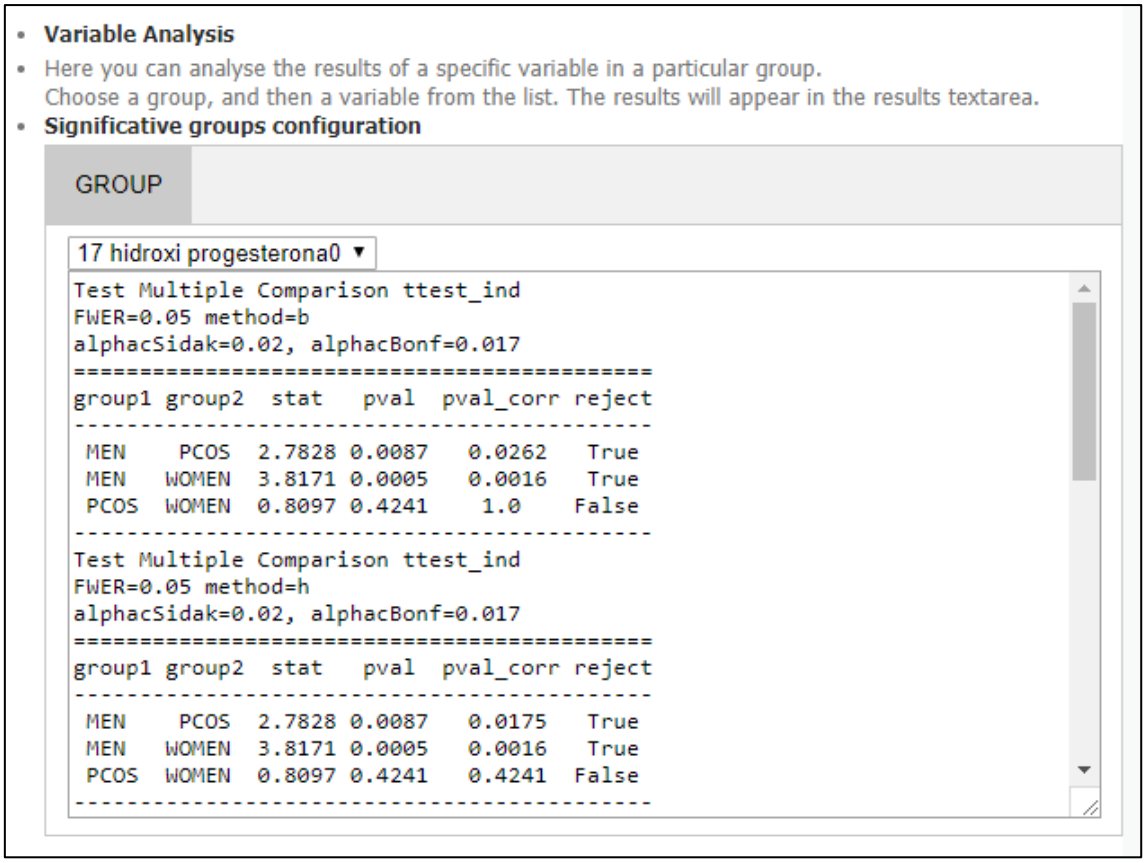


Figure 22. Variable's analysis result screen

7. The final step, the SVM training, is done after the analysis is finished. Before the training, select the number of tests that a detected variable must pass in order to be included on the training (figure 23).

   If you want to include all the detected variables, select 1 test as the minimum, or, if only the variables that passed all the tests must be considered, select as many tests as the number of tests selected on the step 5 (figure 20).

   Once this configuration step is done, press the Train SVM button, and check the results displayed on the results text area.

   Results will show not only the best SVM configuration found, but also how this configuration performed on the evaluation set.

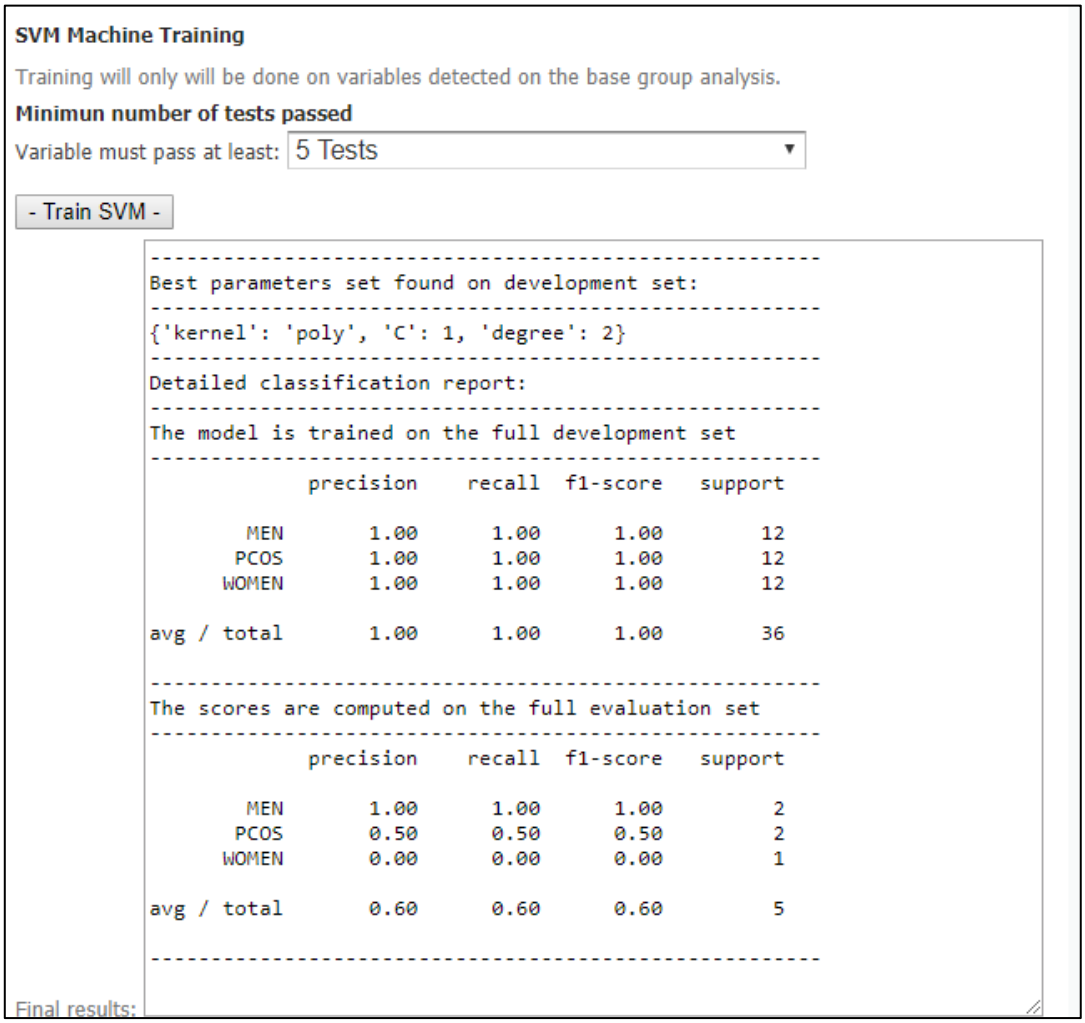


Figure 23. SVM training and results

APPENDIX III

TOOL DEVELOPMENT

The web tool has been built by using python language [46], together with the following libraries:

1. Flask: web server library [48]
2. Itertools: part of python libraries, used to calculate numerical combinations in a efficient manner [49]
3. Pandas: main library for data treatment [50]
4. Scipy: for statistical tests (ANOVA) [51]
5. Statsmodels: post hoc tests validation [52]
6. Sklearn: SVM algorithms [53]

The tool relies on a set of functions that reproduce the procedure steps, by using each of these functions on the website calls.

The data load and cleansing ifs perform by the *carga_datos* and *obtener_perdidos* functions. These two functions will read the CSV file and clean up the data from empty values.

The *reemplazar_valores* is used to set the default value names in the case of PCOS study (MEN, WOMEN, PCOS, OBESE, etc.).

Anova2 function, will perform ANOVA analysis on the data set, enforcing normal distribution as a requirement if asked for.

Post_hoc function, will take the results from ANOVA and will analyse it, based on the configuration parameters passed.

For the interesting variables detection, *var_interesantes* and *obtener_datos* functions are used.

Finally, to built and evaluate the SVM machine (finding the best machine given the parameters), *construir_SVM* function is used.

Below, the source code for the main functions is shown.

```
# -*- coding: utf-8 -*-
"""
Created on Fri Mar 03 19:10:17 2017

@author: Victor Cerquera
"""

import pandas as pd
import numpy as np
```

```
import matplotlib.pyplot as pl
from scipy import stats
import os.path
import itertools
from sklearn import svm, preprocessing
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import classification_report
from sklearn.utils import shuffle
from statsmodels.stats.multicomp import (pairwise_tukeyhsd,
                                         MultiComparison)

MUJERES_SANAS = 0
MUJERES_NO_SANAS = 1
HOMBRES = 2
PORC_VERIF = 0.05

DEBUG=0

def carga_datos(nombre_archivo, separador=';', codif='latin_1', valores_na = " "):
    if DEBUG:
        print("Cargando datos...\n")
    raw_data = []
    #Cargamos los datos ##########################################################
    try:
        raw_data =\
        pd.read_table(nombre_archivo, sep=separador, encoding=codif, na_values=valores_na)
        raw_data.columns = raw_data.columns.str.replace('_', '')
        raw_data.columns = raw_data.columns.str.replace('-', '')
    except:
        print("Error procesando el archivo %s" % (nombre_archivo))
    return raw_data

def obtener_perdidos(datos_base, columnas_sust = [], mapeo = {}):
    if DEBUG:
        print("Revisando datos para obtener perdidos...\n")
    datos_perdidos = {}
    cabeceras = datos_base.columns
    for i in cabeceras:
        if datos_base[i].dtype == 'O':
            datos_base[i] = datos_base[i].apply(pd.to_numeric, errors='coerce')
        datos_perdidos[i] = datos_base[i].size-datos_base[i].count()
        if i in columnas_sust:
            datos_base[i] = datos_base[i].map(mapeo[i])
    return datos_perdidos, datos_base

def reemplazar_valores(datos_trabajo, lista_etiqs=[]):
    if DEBUG:
```

```
        print("Reemplazando valores en variables discretas...\n")
    lista_valida_grupos = []
    for elem in lista_etiqs:
        if(datos_trabajo[elem].dtype == np.float64 or datos_trabajo[elem].dtype == np.int64):
            if not np.isnan(datos_trabajo[elem]).any():
                lista_valida_grupos.append(elem)
                datos_trabajo[elem] = datos_trabajo[elem].astype(int).astype(str).map({x:
'VALUE' + x for x in datos_trabajo[elem].astype(int).astype(str)})
        else:
            lista_valida_grupos.append(elem)
    return datos_trabajo, lista_valida_grupos

def anova2(datos_base, datos = [], columnas_a_excluir = [], validate_normal = False):
    cabeceras = datos_base.columns
    resultado_global = []
    for grupo in datos:
        grkeys = list(grupo.groups.keys())
        indice = []
        indice.append("f")
        indice.append("p")
        resultado = pd.DataFrame(index=indice)
        print 'GRKeys: ' + str(grkeys)
        print 'Aplicando anova a ' + str(grkeys[0]) + ' - ' + str(grkeys[1]) + ' - ' +
str(grkeys[2])
        for col in cabeceras:
            if col in columnas_a_excluir:
                continue
            try:
                if validate_normal:
                    no_normal = False
                    for x in grkeys:
                        if
stats.normaltest(grupo.get_group(x)[col][grupo.get_group(x)[col].notnull()])[1] <= 0.05:
                            continue
                        no_normal = True
                        t = (99999, 99999)
                        break
                    if not no_normal:
                        t =
stats.f_oneway(*[grupo.get_group(x)[col][grupo.get_group(x)[col].notnull()] for x in grkeys])
                else:
                    t =
stats.f_oneway(*[grupo.get_group(x)[col][grupo.get_group(x)[col].notnull()] for x in grkeys])
            except:
                t = [99999, 99999]
            resultado[col] = t
        resultado_global.append(resultado)
    return resultado_global
```

```
def post_hoc(anova_datos_ok, datos, columnas_a_excluir = []):
    cabeceras = anova_datos_ok.index.values
    resultado_global = {}
    etiquetas = datos.groups.keys()

    for col in cabeceras:
        if col in columnas_a_excluir:
            continue
        try:
            #Construimos los datos para Multicomparison
            res_et = []
            res = []
            for etiq in etiquetas:
                values =
list(datos.get_group(etiq)[col][datos.get_group(etiq)[col].notnull()])
                for v in values:
                    res.append(v)
                    if type(etiq) is str:
                        res_et.append(etiq)
                    else:
                        res_et.append('_'.join(etiq))
            t = MultiComparison(res, res_et)
        except:
            continue
        resultado_global[col] = t
    return resultado_global

def var_interesantes(test, vars = {}):
    for etiq in test.keys():
        if vars.has_key(etiq):
            vars[etiq] = vars[etiq] + 1
        else:
            vars[etiq] = 1
    return vars

def obtener_datos(grupo, variables, datos_trabajo):
    datos = []
    etiquetas = datos_trabajo[grupo]

    for individuo in datos_trabajo[grupo]:
        datos.append([])

    for caracteristica in variables:
        actual = 0
        for individuo in datos_trabajo[caracteristica]:
            datos[actual].append(individuo)
            actual = actual + 1
```

```
    return datos, list(etiquetas)

def construir_SVM(grupo, variables, datos_trabajo, kernel_svm = "linear", C_svm = 1.0, param = 0.5):
    # Con los datos de trabajo, construimos nuestros arrays de grupo y variables, que representan nuestros datos.
    datos, etiquetas = obtener_datos(grupo, variables, datos_trabajo)

    X_train, X_test, y_train, y_test = train_test_split(datos, etiquetas, test_size=0.1, stratify=etiquetas)
    parameters = [{'kernel': ['rbf'],
                'gamma': [1e-4, 1e-3, 0.01, 0.1, 0.2, 0.5, 0.75, 0.9, 1],
                 'C': [1, 10, 100, 1000]},
                 {'kernel': ['rbf'],
                 'C': [1, 10, 100, 1000]},
                 {'kernel': ['poly'],
                'degree': [2,3,4],
                 'C': [1, 10, 100, 1000]},
                 {'kernel': ['sigmoid'],
                 'C': [1, 10, 100, 1000]},
                {'kernel': ['linear'], 'C': [1, 10, 100, 1000]}]

    clf = GridSearchCV(svm.SVC(decision_function_shape='ovr'), parameters, cv=3, verbose=200)
    clf.fit(X_train, y_train)
    means = clf.cv_results_['mean_test_score']
    stds = clf.cv_results_['std_test_score']
    y_truet, y_predt = y_train, clf.predict(X_train)
    y_true, y_pred = y_test, clf.predict(X_test)


    resultados = "-----------------------------------------------------\n"
    resultados = resultados + "Best parameters set found on development set:\n"
    resultados = resultados + "-----------------------------------------------------\n"
    resultados = resultados + str(clf.best_params_) + "\n"
    resultados = resultados + "-----------------------------------------------------\n"
    resultados = resultados + "Detailed classification report:\n"
    resultados = resultados + "-----------------------------------------------------\n"
    resultados = resultados + "The model is trained on the full development set\n"
    resultados = resultados + "-----------------------------------------------------\n"
    resultados = resultados + str(classification_report(y_truet, y_predt)) + "\n"
    resultados = resultados + "-----------------------------------------------------\n"
    resultados = resultados + "The scores are computed on the full evaluation set\n"
    resultados = resultados + "-----------------------------------------------------\n"
    resultados = resultados + str(classification_report(y_true, y_pred)) + "\n"
    resultados = resultados + "-----------------------------------------------------\n"
    return resultados
```