TecnoCampus
Escola Superior Politècnica

Centre adscrit a la
upf. Universitat Pompeu Fabra Barcelona

Degree in Computer Engineering of Management and Information Systems

Analysis of multisensor data on cervical injuries

Report

Eloi Rodríguez Gaxas
TUTORS: Xavier Font
Carles Paul

2017/2018

TecnoCampus
Mataró-Maresme

Abstract

Study of the data from 27 users, eleven of them with a cervical fracture, extracted from thermographic frames, electroencephalographic waves and inertial biomechanics mobility angels, to detect patterns and achieve a model from the dataset and give validity to the results obtained previously with the sensors. A path through all the steps of the statistical modeling to understand the data and achieve the results.

Resum

Estudi de les dades de 27 usuaris, onze d'ells amb lessió cervical, extreta de fotogrames termogràfics, les ones electroencefalogràfiques i els angles de mobilitat inercials, per detectar patrons i aconseguir un model del conjunt de dades i per així validar els resultats obtinguts anteriorment amb els sensors. Un camí a través de tots els passos de la modelització estadística per entendre les dades i assolir els resultats.

Resumen

Estudio de los datos de 27 usuarios, once de ellos con fractura cervical, extraída de fotogramas termográficos, las ondas electroencefalográficas y los ángulos de movilidad inercial, para detectar patrones y conseguir un modelo del conjunto de datos y dar validez a los resultados obtenidos anteriormente. Un camino a través de todos los pasos de la modelización estadística para entender los datos y alcanzar los resultado.

Table of Contents

Index of figures

Index of tables

1. Introduction

The project, using a dataset of patients where we have the biomechanical, EEG and thermographic sensors answer and, in addition, survey questions, extracts statistics, which determine the patient possibility of suffering a cervical injury.

The main idea behind the project is avoiding the necessity of radiographies in the majority of cases, saving resources and time to the professionals and their patients and to obtain a first vision of the possible injury.

It could avoid the necessity of reading and understand the output of the sensor to determine if the patients have or have not an injury.

Pretends to be the first phase of the patients to streamline the processes and be able to use more time for users with higher needs.

Another possibility is to have the knowledge of an injury status during the course of a physiotherapist recovery sessions and be able to determine whereas a patient has reached its maximum level of recovery or if it could be beneficial more meetings.

This project starts with data obtained from a previous development where a series of metrics where decided and applied to obtain the best qualitative values possible. The records correspond to the output of electroencephalographic, inertial biomechanics and thermographic camera sensors. With this data, it can be observed limits in the values representing the users with injury and no injury.

The main purpose is obtaining a model showing the correlations between the predictors and the target and achieve predictions based on training/test methodologies or supervised algorithms, demonstrating the reliability of the given data and the used techniques to have obtained the results.

2. Theoretical scheme

Cervical fractures are normally started with plain radiographies. Methods like tomography, CT and MRI are also being used *[1]*.

Radiography can identify most cervical fractures and ligamentous injuries. Plain radiographies are cheap, readily available and noninvasive. Approximately 85-90% of cervical spine injuries are evident in lateral radiographs.

Computed Tomography has supplanted plain radiography in many centers. It is capable of discovering injuries, which plain radiographies cannot.

Using different techniques to understand the patient status could help improve an area, which currently is slow and expensive. This is an innovative and creative approach to a stabilized sector that still uses the same methods.

This project is the continuation the previous work ‘Sistema de valoració biomecànica inercial, EEG I termografia aplicat a cervicals’, where the student extracted the sensors’ outputs from several patients and defined limits, which could identify the existence of a cervical injury.

2.1. The data

The techniques used were electroencephalograph, thermography and inertial biomechanics.

The project starts with data from the sensors:

- Electroencephalograph (EEG) *[2]*:

  Electroencephalography is the recording and evaluation of electrical pulses generated by the brain and obtained by electrodes located on the surface of the scalp.

  EEG systems amplify the signal generated by the oscillation of potential dendrites of neuronal populations. Depending on the number of sensors, it is simpler to identify the origin of the electric potentials generated by a group of neurons.

- Thermography:

  Infrared thermography is a technique that allows, at a distance and without any contact, to measure and visualize surface temperatures with precision. Physics allows converting infrared radiation measurements into temperature measurements, this is achieved by measuring the radiation emitted in the infrared portion of the electromagnetic spectrum from the surface of the object, converting these measurements into electrical signals *[3]*.

  Thermography, in the medical world, is currently used to detect skin problems, surface tumors, hematomas or hemorrhages. It is also used for the detection of muscular excitement and problems of blood irrigation *[3]*.

- Inertial biomechanics:

  Biomechanics is a scientific discipline that is dedicated to studying the activity of the human body, under different circumstances and conditions, analyzing the mechanical consequences that derive from the activity. To study the effects of this activity, Biomechanics uses the knowledge of mechanics, engineering, anatomy, physiology and other disciplines. Biomechanics is interested in the movement of the human body and the mechanical loads and energies that are produced by this movement *[4]*.

The project starts with data from 27 patients and 41 different fields, which includes survey questions made to the patients.

2.2. Statistical modeling

Statistical modeling is a simplified way to approximate reality and to make predictions from this approximation. Is a mathematical model that embodies a set of statistical assumptions concerning the generation of some sample data and similar data from a larger population *[5]*.

2.2.1. Dependent and explanatory variables

The dependent variable, or target variable, is the one we want to describe, explain or predict. In statistics, this variable is often represented it in the Y-axis.

In this project, the target variable is ‘LesionsCervicals’ as shown in the next chapters.

The explanatory variables, or predictor variables, are the rest of the data, which influence the target. Based on them the different models can predict and explain the target [5].

2.2.2. Choosing a model

Every statistical model answers to different questions. Depending on the number of targets or predictors, you may need one model or another.

The choice can also be guided by the shape of the relationships between the dependent and explanatory variables. *[5]*

The next table is a small guide to help you choose the model to use:

Dependent variable	**Explanatory variable(s)**	**Example**	**Parametric models**	**Other solutions**
One quantitative variable	One qualitative variable (= factor) with two levels	Effect of contamination (yes / no) on the concentration of a trace element in a plant	One-way ANOVA with two levels	Mann-Whitney test
	One qualitative variable with k levels	Effect of the site (4 factories) on the concentration of a trace element in a plant	One-way ANOVA	Kruskal-Wallis test
	Several qualitative variables with several levels	Combinatory effects of site (4 factories) and plant species on the concentration of a compound in plant tissue	Multi-way ANOVA (factorial designs)	
	One quantitative variable	Effect of temperature on the concentration of a protein	Simple linear regression; nonlinear models (depends on the shape of the relationship between the dependent / explanatory variable)	nonparametric regression(*);quantile regression; classification / regression trees(*); K Nearest Neighbors(*)

	Several quantitative variables	Effect of the concentration of several contaminants on plant biomass	Multiple linear regression; nonlinear models	PLS regression(*); K Nearest Neighbors(*)
	Mixture of qualitative / quantitative variables	Combinatory effects of sex and age on glycaemia associated to a type of diabetes	ANCOVA	PLS regression(*); quantile regression; classification / regression trees(*); K Nearest Neighbors(*)
Several quantitative variables	Qualitative &/or quantitative variable(s)	Effect an environmental variables matrix on the transcriptome	MANOVA	Redundancy analysis; PLS regression(*)
One qualitative variable	Qualitative &/or quantitative variable(s)	Dose effect on survival / death of mouse individuals	Logistic regression (binomial or ordinal or multinomial)	PLS-DA(*); Discriminant Analysis(*); classification / regression trees(*); K Nearest Neighbors(*)
One count variable (with many zero's)	Qualitative &/or quantitative variable(s)	Dose effect on the number of necroses in mice	Log-linear regression (Poisson)	

Table 2.2.2.1. Guide for choosing the model to use. Source: XLSTAT web

3. Objectives and scope

3.1. Objective

The main purpose of the project is developing a model, which presents statistics graphs and probability of cervical injuries.

This model must be capable of predicting the target based on the explanatory variables.

Also, find the correlations between the data to determine which variables affect the target.

The objective is to validate the data and demonstrate the reliability of this approach. The project is the continuation from the ‘Sistema de valoració biomecànica inercial, EEG i termografia aplicat a cervicals’ where the main idea was to be able to detect a cervical injury based on different sensors.

This project using the obtained data from the previous tries to validate the theory of detecting the cervical injuries using statistics and machine learning.

Another important objective, essential to obtain the main purpose, is the detection of the data without value. It may be because the variables do not affect the target or maybe it was not taken properly, leading to data mistakes.

3.2. Scope

The scope of the project is:

1. Data exploratory.
    1.1. Identify all the columns.
    1.2. Identify the type of the columns.
    1.3. Describe the data.
    1.4. Check for correlations.
    1.5. Identify the dependent and explanatory variables (target/predictors)
2. Develop a model.
    2.1. Decide which model to use.
    2.2. Prepare the data
    2.3. Divide the data in train/test.
    2.4. Fit the model (target and predictors).
    2.5. Extract the predictions.
    2.6. Analyze the results.
3. Use clustering methods.
    3.1. Decide the clustering method.
    3.2. Fit the data (target and predictors).
    3.3. Visualize the real scatter plot divided by the classes.
    3.4. Visualize the predicted scatter plot divided by the classes.
    3.5. Compare the outputs
4. Conclusions

4. Methodology

To find the necessary information a process has been defined.

The search engine has been basically Google Scholar. Google Scholar is a searcher from academic publishers, professional societies, online repositories, universities and other websites.

This gives you the assurance that the information is probably reliable. However, information was sought primarily from universities or academic journals.

The used words were: cervical fracture, computed cervical fracture data, electroencephalography, thermography, inertial biomechanics, languages for computing data, python data analysis, R data analysis, …

In the first phase of the project, the research is the most important part. The methodology used in this section has been saving every web or article, talking about the topic or something related, then doing a selection of all the information based on the font that published it.

The second phase is based on the agile techniques *[6]*. The agile part is on the recursive. It will not be used a scrum methodology because there is only one developer for doing the data science and there is not a whole team to work as it should be.

The used part of the agile technique is the recursive cycle. Each week a new progress is done, test it, analyzed with the tutors and refactor it in case of need.

With the weekly meetings, periodic revisions can be done and changes in the calendar are easier, making it more flexible to handle the possible errors.

Given that the main purpose of the project is to test the data reliability, we need a methodology to follow and achieve our goals.

The project is divided into four different tactics to work with the data, as it will be explained in the development chapter. In each approach, the same procedure will be applied to be able to compare the results.

The process for every strategy will be:

- Prepare the data
    - Transform the data to be able to use it with the different statistical models.
- Searching for correlations
    - With each set of data, we will search for possible linear correlations between the predictors and the target.
- Testing the models
    - Obtain predictions from a set of different models
- Results analysis
    - Analysis the archived results from each model

5. Functional and technological requirements

5.1. Functional Requirements

The functional requirements will be explained using the Give When Then *[7]* technique used in the Domain Driven Design *[8]*.

This are the requirements for the project:

Give When Then – Scenario 1	
Scenario	Use data exploratory method to understand the dataset.
Given	A dataset
And	A series of columns from the dataset
When	An explanation from the different variables is asked
Then	We identify all the variables
And	The total length from the dataset
And	Identify the target / predictors

Table 5.1.1. Give When Then – Scenario 1.

Give When Then – Scenario 2	
Scenario	Transform the values to be able to work with them.
Given	A dataset
And	A series of columns from the dataset
When	The variables are not suitable to be used by a statistical model.
Then	Identify the variables types
And	Identify how the model needs the data
And	Prepare the records to be suitable for the model.

Table 5.1.2. Give When Then – Scenario 2.

Give When Then – Scenario 3	
Scenario	Divide the data in training / testing sets.
Given	A dataset
And	A series of columns from the dataset
When	The dataset wants to be tested with a model
And	The dataset must be dived for training and testing the model
Then	Identify the needed length from the new training and testing sets.
And	Randomize the dataset split

Table 5.1.3. Give When Then – Scenario 3.

Give When Then – Scenario 4	
Scenario	Obtain predictions from the different statistical models.
Given	A dataset
And	The targets and predictors
And	A statistical model
When	We need to obtain predictions from the model
And	Indicate the targets and predictors to the model
And	Obtain the training and testing datasets
And	Train the model to learn about the records
And	Indicate the testing dataset to the model.
Then	The model returns a correct prediction average.

Table 5.1.4. Give When Then – Scenario 4.

Give When Then – Scenario 5	
Scenario	Obtain multiple predictions from different models directly.
Given	A dataset
And	The targets and predictors
And	A set of statistical models
When	We need to obtain predictions directly from all the models
And	Indicate the targets and predictors to the model
And	Obtain the training and testing datasets
And	Train the model to learn about the records
And	Indicate the testing dataset to each model.
Then	The method returns a set of predictions for each model

Table 5.1.5. Give When Then – Scenario 5.

Give When Then – Scenario 6	
Scenario	Comparative between the different models results.
Given	A set of different statistical models
And	The predictions average for each model
When	
And	Indicate the targets and predictors to the model
And	Obtain the training and testing datasets
And	Train the model to learn about the records
And	Indicate the testing dataset to each model.
Then	The method returns a set of predictions for each model

Table 5.1.6. Give When Then – Scenario 6.

Give When Then – Scenario 7	
Scenario	Cluster the data to compare the results between the real classification and the predictions
Given	A dataset
And	The different possible class
And	A clustering algorithm
When	We indicate the method to the clustering model
And	We indicate the data to the model
And	We indicate the possible classes
Then	The method returns a plot with dots distributed in the x, y axis differentiated by colors

Table 5.1.7. Give When Then – Scenario 7.

5.2. Technological Requirements

In this chapter, we will discuss and explore the different alternatives to do data analysis that exists in the current market.

The list of programming languages has incresead with the time and currently you can do data science with almost any language program.

This is list of possibilities, but there is more *[9]*:

- R
    - It is a powerful language that excels at a huge variety of statistical and data visualization applications. Open source program with an active community of contributors.
- Python
    - General purpose programming language but with a large amount of packages for data science and machine learning applications
- SQL
    - More useful as a data processing language rather than as an advance analytical tool.
- Java
    - Useful for the ability to integrate data science production code directly into existing codebase. However, it does not have a large number of specialized packages and it can be quite limiting.
- Matlab
    - Very useful for applications with sophisticated mathematical requirements.

From the list of program languages, we will choose between R *[10]* and Python *[11]*.

R	Python
Pros	
A package for almost every quantitative and statistical application. Including neural networks, non-linear regression, …	General purpose programming language with an extensive range of purpose-built models and community support.
The base installation comes with very comprehensive, in-built statistical functions and methods.	Easy language to learn and use making it easier for new programmers.
Really good data visualization with libraries like ggplot2	Great number of packages that makes python a solid option for advanced machine learning applications
Cons	
The language can suffer from performance	It is a dynamically typed language. It means that can appear errors when running, for example passing a string to a method when it expects an integer.
Great for statistics and data science but with problems for general purpose programming.	For specific statistical and data analysis purposes it can be left behind compared to other statistical programming languages.
Unusual features, which defers from typical programming languages.	

Table 5.2.1. Pros and cons from R and Python, comparative table.

As there are no notable differences I am choosing Python basically because when having a developer background, it feels more comfortable whereas having an analyst background R will likely be more *[12]*.

After the comparative, the project uses Python as the programing language

The use of Python and not R, as typically would be when talking of data analysis, is that both languages are actively being develop, have a large number of tools and libraries for collecting, managing and visualizing data. As the project does not need specific features, we will use Python for being easier to develop with and multipurpose.

To achieve the objectives, the technological requirements are:

- A computer with internet access to download all the necessary packages.
    - Python is not a very demanding language and a computer with 2 GB of RAM is enough to code and run the programs.
- Python 3.X.

Given that the project already starts with the data and the main purpose is to analyze it, we do not have more technological requirements. In fact, it is not necessary the use of a graphic interface, the project could be done only using a console.

6. Development

In this chapter, we will work with the data and search the different possible approaches.

As explained before in ‘4. Methodology’, in this part of the project will be done using agile techniques.

The topics review here are:

- Data Exploratory
- The different models
- Sensors outputs
- Survey outputs
- Computed sensors outputs
- Sensors and survey outputs
- Clustering

In the chapters Sensors outputs, Survey outputs, Computed sensors outputs and Sensors and survey outputs, we will prepare the data, search for possible correlations, test the different models and analyze the results.

To analyze the effectiveness from a model, we normally use techniques as confusion matrix, where you can see what have predict correctly and incorrectly the models and the predictions % correct prediction average from the different statistical models.

6.1. Data Exploratory

Exploratory Data Analysis (EDA) *[13]* is the first step in every data analysis process. Here, you make sense of the data you have and then figure out what questions you want to ask and how to frame them, as well as how to manipulate your available data sources to get the answers you need.

In this chapter, we will:

- Understand the different columns from the data and what meaning they have, to be able to use the best approaches from the dataset and to group them in categories.
- Find the different type of values, programmatically, contains the data. With this we will be able to transform it accurately.
- Discover the length of the dataset and the quantity of rows.
- Find the limits from the sensors values that determine the value from the target.

The main idea is to understand the data and see what can it offers. It is the search for possible correlations between variables.

6.1.1. Data description

The first step is to identify all the columns and its meaning. Doing the research, we can identify our target, the variable we want to predict, and the explanatory variables, the predictors that describe our target.

The following columns compose the collected data:

Usuari, Edat, Sexe, Professio, Esport, Alcohol, DormirAvui, DormirNormalment, AccidentAntic, AccidentActual, Sequeles, LesionsCervicals, Postura, Usordinador, EstresActual, DolorsAssociatsEstres, TipusMalCap, UsTecnicaXcarregarPesos, Molesties, MolestiesEEG, MovimentsLimitats, MovimentsComplets, Dif.Temp.Hab., Per.Iner.SI, Per.Iner.NOSE, Per.Iner.NO, Dif.Dreta, Dif.Esquena, Dif.Esquerra, ExtFlex.Max.Des., Incl.Max.Des., Rotac.Max.Des., Desv.SI.Max., Desv.NOSE.Max., Desv.NO.Max., EEG.zona.cortical.Desv., EEG.zona.cortical.Desv.Interquartil, EEG.zona.temporal.Desv., EEG.zona.temporal.Desv.Interquartil, EEG.zona.parietal.Desv., EEG.zona.parietal.Desv.Interquartil.

We can separate the data into two different sections:

- Survey inputs:

  The data contains several columns referencing to questions made to the participating users. This data contains questions such as: the user sex, hours of sleep, …

  This data can be really useful combined with the sensors outputs, to give us a different approach.

  The variable '*LesionsCervicals'* is our target. This variable dictates if the user has or has not a cervical fracture. This column is not just binomial, it also determines the injury grade.

  The column *'LesionsCervicals'* can be 'NO', 'Lleu', 'Moderat', 'Fort'. This variable is the target.

- Sensor outputs

  The rest of the columns represent the results obtained from the tests with the EEG, thermography and inertial biomechanics.

These variables represent the output from the sensors. Based on their results the model can predict the target.

The thermography camera results represents a photogram from each part of the neck, showing the temperature difference between before the exercise and after it.

The Electroencephalography results consist of the most representative outputs from the cortical, temporal and parietal areas.

The inertial biomechanics data consists of the standard deviation, maximum, minimum and the % of deviation from the normal state.

After the past experiments, they arrived the next conclusion.

- Thermographic data:

The patients with cervical injuries presents a difference of temperature, between before exercise and after, from 0.88ºC.

In the other hand, we have that the users, which do not present a cervical injury, not normally represents a major difference from 0.4ºC.

- Electroencephalography data:

People with cervical injuries show a degree of mobility and a standard deviation much lower than normal.

In addition, people in advanced age also demonstrate relatively lower degree of mobility and standard deviation.

- Inertial biomechanical data:

People with relatively important injuries indicate values lower than 75% mobility. The patients with advanced ages, demonstrates a 90% of normal mobility.

6.2. The different models

In this chapter, we will analyze the different statistical models and understand how they work.

We can divide the models in two different categories: Decision Tree Classifier and Linear Classifiers.

The Decision Tree classifiers constructs a tree with all the different predictors and tries to determine the value of the target.

The Linear classifiers search for a direct correlation from one or more predictors with the target. Does not have in count all the values.

When talking about machine learning algorithms we have two different type:

- Supervised Machine Learning *[14]*
    - This methods use training and testing datasets where they need to know which are the predictors and targets. Using the training records, they learn about the variables and how they influence the output.

- Unsupervised Machine Learning *[14]*
    - Unsupervised learning is where you only have input data (X) and no corresponding output variables. These are called unsupervised learning because unlike supervised learning, there is no correct answers and there is no learning before. Algorithms are left to their own devises to discover and present the interesting structure in the data.

The models used in this project are Logistic Regression, Support Vector Machines, Random Forest Classifier, Gradient Boosting Classifier and K-Means Clustering.

6.2.1. Logistic Regression

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest and a set of independent variables *[15]*.

6.2.2. Support Vector Machines

In machine learning, support vector machines *[16]* are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

6.2.3. Random Forest Classifier

The Random Forest Classifier *[17]* is an ensemble classification algorithm. This classifier instead of using only one explanatory variable to predict the target, it uses multiple classifiers.

The classifiers are the randomly created decision tree. Each decision tree is a single classifier and the target prediction is based on the majority voting method. Every classifier vote for to one target class and the most voted is the winner for that input.

6.2.4. Gradient Boosting Classifier

The gradient boosting algorithm *[18]* take decisions from the tree sequential and not in parallel as it does the random forest classifier.

The gradient boosting algorithm repetitively leverage the patterns in residuals, strengthen a model with weak predictions, and make it better. Once we reach a stage that residuals do not have any pattern that could be modeled, we can stop modeling residuals (otherwise it might lead to overfitting). Algorithmically, we are minimizing our loss function, such that test loss reach its minima.

6.2.5. K-Means Classifier

K-Means *[19]* is an unsupervised algorithm, used to label your data. The objective from this classifier is to find groups in the data. The number of groups is represented with the K, which you indicate to the model.

Data points are clustered based on feature similarity. The results of the *K*-means clustering algorithm are:

1. The centroids of the *K* clusters, which can be used to label new data.
2. Labels for the training data (each data point is assigned to a single cluster).

6.3. Sensors Outputs

This chapter works only with the values from the sensors. It will use this data to test the different models and discover how well a model can predict the target using only the sensors.

The columns corresponding to the sensors are 19 having each of one a different meaning. To be able to work with them and obtain some results we will need the transform it, using aggregations, maximum or minimum values, ….

As we discovered in the chapter ’6.1.1. Data description’, this data consists of the three different parts of the neck: left, right and back.

As we do not have enough big dataset, we will use only the maximum or minimum of each set of three columns.

6.3.1. Prepare the data

In this chapter, the main idea is to see all the columns and decide if they are as we need and drop the columns that we do not want in our model, as well as creating new columns if needed based on other variables.

To continue all the survey inputs, except *'LesionsCervicals'*, will be extracted from the rest, to just work with the relevant variables.

The result is a total of 20 columns and 27 entries.

COLUMN NAME	COUNT	TYPE
LESIONSCERVICALS	27 non-null	object
DIF.TEMP.HAB.	27 non-null	float64
PER.INER.SI	27 non-null	int64
PER.INER.NOSE	27 non-null	int64
PER.INER.NO	27 non-null	int64
DIF.DRETA	27 non-null	float64
DIF.ESQUENA	27 non-null	float64
DIF.ESQUERRA	27 non-null	float64
EXTFLEX.MAX.DES.	27 non-null	float64
INCL.MAX.DES.	27 non-null	float64
ROTAC.MAX.DES.	27 non-null	float64
DESV.SI.MAX.	27 non-null	float64
DESV.NOSE.MAX.	27 non-null	float64
DESV.NO.MAX.	27 non-null	float64
EEG.ZONA.CORTICAL.DESV.	27 non-null	float64
EEG.ZONA.CORTICAL.DESV.INTERQUARTIL	27 non-null	float64
EEG.ZONA.TEMPORAL.DESV.	27 non-null	float64
EEG.ZONA.TEMPORAL.DESV.INTERQUARTIL	27 non-null	float64
EEG.ZONA.PARIETAL.DESV.	27 non-null	float64
EEG.ZONA.PARIETAL.DESV.INTERQUARTIL	27 non-null	float64

Table 6.3.1.1. Description of the columns from the data.

As we can see, we have one object, corresponding to 'NO', 'Lleu', 'Moderat' and 'Fort', three integers and 16 inputs that are decimals. All the variables have each row informed, meaning that there are no null entries.

The only object, which also is our target, should be transformed into an int64 for an easier work later.

The rows are transformed as:

- ‘NO’ : ‘0’
- ‘Lleu’ : ‘1’
- ‘Moderat’ : ‘1’
- ‘Fort’ : ‘2’

We have already removed all the survey columns in the section ‘6.1.1 Data description’ because this chapter is based on the sensors results as explained before.

The sensors results are separated into three columns each, representing the left, right and back.

Based on the conclusions from the previous project and by looking at the data, we can observe the limits between having a cervical fracture or not.

These limits are:

- Thermography:

The users with cervical fractures show a difference of 0.8ºC approximately, while the others are around 0.4ºC.

- EGG:

The users with cervical fractures indicate a much lower degree of mobility and a lower standard deviation.

- Inertial:

The users with important cervical fractures have less than a 75% of the mobility.

Having those limits in the count and considering the small data set, we can create new columns, which get the bigger or lower result, depending on the variable, from the three columns and not take in count right now in which part is the injury.

After the transformation, the resulting set is:

LesionsCervicals	***Max_Dif_Temp***	***Max_Dif_Des***	***Max_Dif_EGG_Desv***	***Max_Dif_EGG_Interquartil***
0	0.69	53.377778	17.47	45.64
1	1.23	68.577778	44.24	41.54
2	1.65	49.944444	32.58	34.36
0	0.48	63.711111	22.66	31.79
2	0.65	16.877778	69.04	43.08
…	…	…	…	…

Table 6.3.1.2. The last five rows from table after cleaning the data.

6.3.2. Possible correlations

For searching correlations, we will be using the Heatmap method from the library Seaborn.

To find the correlations it uses the Pearson method by default.

The Pearson product-moment correlation coefficient *[20]* or Pearson correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of zero indicates that there is no association between the two variables. A value greater than zero indicates a positive association and a value less than zero indicates a negative association.

The Heatmap shows, in the first row, the target and the others are the predictors, forming a square.

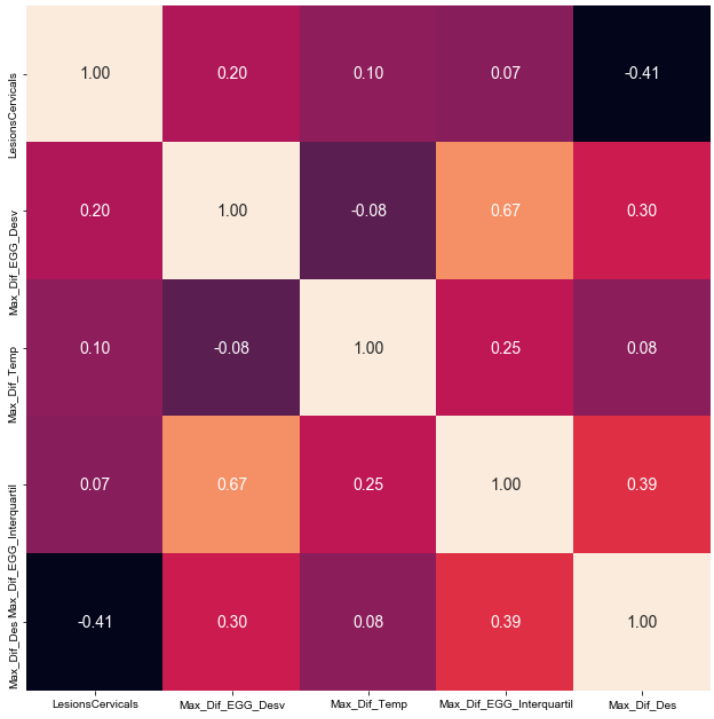

Fig. 6.3.2.1. Correlation Heatmap between the different variables.

As we can see, it does not seem that *'LesionsCervicals'* has a lot of correlated variables. The most correlated column is *'Max_Dif_EEG_Desv'*, the other values are depreciable.

6.3.3. Developing a model

As we explained before, developing a model is an essential part to approximate the reality and demonstrate the validity of the results. If our model is capable of predicting with a good accuracy, it means that the explanatory variables, in our case that are the result of the sensors, were collected correctly and that has an effect on the target *[9]*.

Modeling is based on three main parts:

1. Define and design:

   Design the study you want to do. Understand the data and choose the explanatory variables and dependent variables.

   Determine the level of measurement of each response and predictor variable.

2. Prepare and explore:

   Collect, code, enter and clean data.

   The collected data may not be exactly as you need it, so it is necessary to clean the data, transform it and even create new variables from the other columns.

   Check the distributions of the variables you intend to use. Search bivariate relationships between the data, to decide what to put into the model.

3. Refine the model

   Check the model fit, test it, resolve data issues and interpret the results.

6.3.3.1. Choosing the model

As seen in the chapter '6.3.2. Possible correlations' we do not have strong bivariate variables, meaning a strong linear regression to search.

Therefore, the first chosen model is Random Forest Tree. This model is a classifier, which means, that using the predictors, the model tries to predict the target result. Even though, we have also tried the other models to use the model with better predictions rate.

To work with classifiers, we need the split the data into training and testing. The model uses the training set to understand the data and be able to make predictions about the target output of the testing data set.

After preparing the data and running the algorithms, we obtain:

- Model Class: LogisticRegression
  Cross Validation: 0.51 (+/- 0.19)

- Model Class: LinearSVC
  Cross Validation: 0.47 (+/- 0.31)

- Model Class: RandomForestClassifier
  Cross Validation: 0.66 (+/- 0.26)

- Model Class: GradientBoostingClassifier
  Cross Validation: 0.55 (+/- 0.30)

6.3.3.2. Random Forest Classifier

The Random Forest Classifier obtained the best results.

The first step has been to divide the set into train/test. The 80% of the test is for training our model and the 20% for testing it.

In this attempt, the model got an accuracy of 66%. This may be different in other trees and the accuracy can be higher or lower.

Using a confusion matrix, which is a specific table layout that allows visualization of the performance of an algorithm, we are able to observe what the classifier has predicted correctly.

	Predicted 0	**Predicted 1**	**Predicted 2**
Actual 0	4	0	0
Actual 1	1	0	0
Actual 2	0	1	0

Table 6.3.3.2.1. Confusion Matrix based on the Random Forest Classifier predictions.

The model has only predicted correctly the cases where the user had no cervical fracture. This may be because 15 of the users had no injury and the rest, 12 users, have an injury in different grades.

In the next bars graph we can see the importance of each predictor in taking the decision to the target.

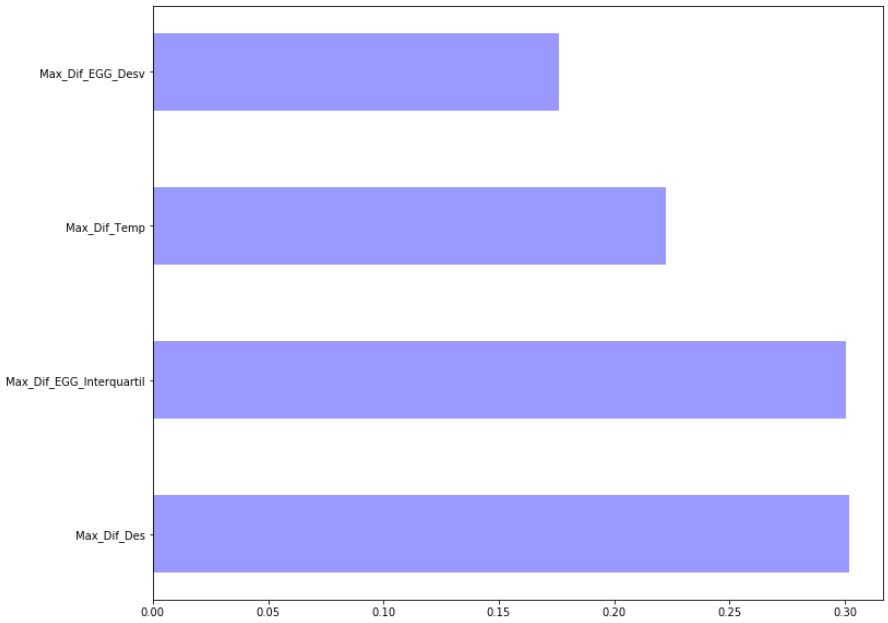

Fig 6.3.3.2.2. Bar graph of the importance of each predictor used in the model.

The most useful predictors for the model were ‘Max_Dif_EGG_Interquartil’ that represents the interquartile deviations allowing identifying the number of peaks susceptible to being manifestations of pain and ‘Max_Dif_Des’ that represents the percentage of mobility.

6.3.4. Sensors analysis

As we can see, the achieved results are not completely satisfactory. They are quite low but they give us the output.

These results are indicating the lack of data. The approach could achieve better predictions but the training set has been only of 21 rows and furthermore, some patients have inconsistent values, making it harder for the model to predict accurately.

Even though the results were not as good as expected, still are relevant and can be a template of how to achieve positive predictions.

6.4. Survey Outputs

The records contain some information about the participants. This information was extracted using a survey before the tests started.

The data contain:

- Basic info, which is the subject sex and age.
- User lifestyle, referring to if practices sport with regularity, if drinks alcohol regularly, how has sleep today, how does normally sleep, if the user has an old accident or an actual one, if has squeals from the accident, how does consider its corporal position, the time spent with the computer and if loads objects with the proper technique.
- Test aspects, where we have the actual stress, pain associate it with the stress, the level of headache, if had discomfort in any of the tests made, if it was comfortable with the EEG headset, if felt that the movements were limited by something and if the movements were completed.

6.4.1. Prepare the data

The survey info is not ready to be used. All this information is classified as categorical data, as they are all Strings. For working with the models, we need to transform it into numbers.

The data consist of different type of answers. We will deal with the columns containing the next outputs: "SI", "NO", "F", "M", "-", "Lleu", "Moderat", "Fort", "Dolenta", "Normal" and "Bona".

Columns like sleep today, actual stress, … Are given with two digits, meaning, for example, a stress level of 3/5. For this columns, we will use the mean between both numbers.

6.4.2. Possible correlations

In this chapter, we are searching for possible correlations between the survey values and the target.

For this purpose, the idea is to use a Heatmap. This technique uses the Pearson method, which searches for a linear association, meaning that the target is directly affected by a predictor.

The next Heatmap represents the ten most correlated variables.

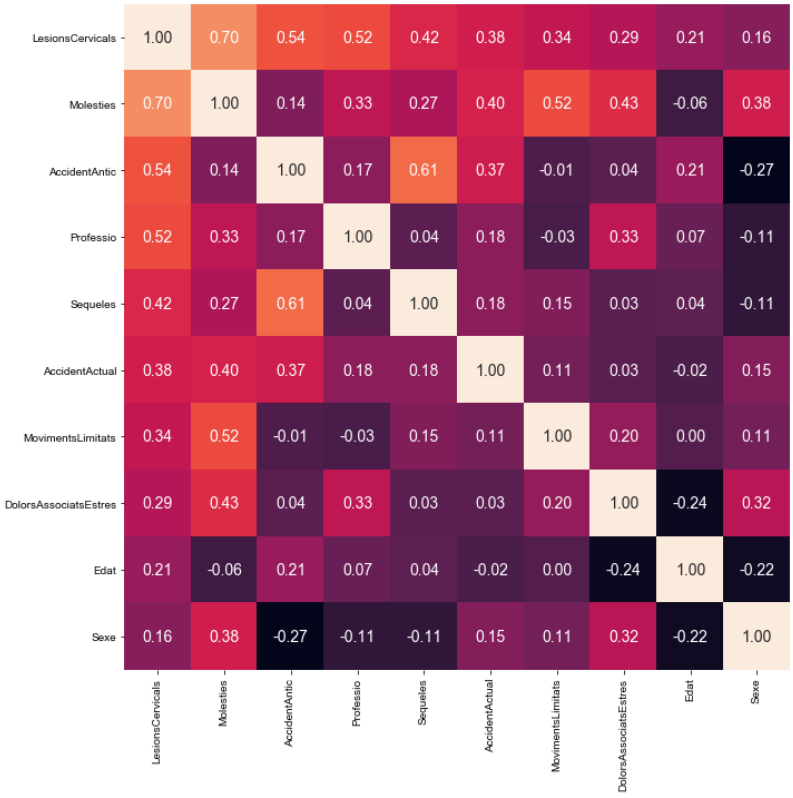

Fig. 6.4.2.1. Correlation Heatmap between the categorical predictors and target.

The obtained results demonstrate that the categorical data has a direct influence on the target. The variable ‘LesionsCervicals’ is highly correlated with ‘Molesties’, ‘AccidentAntic’ and ‘Professio’, the rest of the columns are depreciable.

Furthermore, this output indicates that the models with better results could be the linear regression models rather than the classifiers.

To help the models perform better, we will remove all the data except “Molesties”, “AccidentAntic”, “Professio”, “Sequeles” and “AccidentActual”, which have obtained the best results. The rest of the data could generate noisy and make the models obtain lower result predictions.

6.4.3. Testing the models

In this chapter, after having prepared the data, as explained at ‘6.4.1. Prepare the data’, we will construct a python class to test a series of models.

Using this technique, we are able to test the most used algorithm and choose the best one for the data.

The tested models are:

- LogisticRegression
- LinearSVC
- RandomForestClassifier
- GradientBoostingClassifier

In order to achieve more accurate and reliable results, we will use a cross-validation method.

This method what it does is given a dataset, indicating which are the predictors and the target, separates the data in a training/testing sets and tests the accuracy of the model. This process is done sequentially the times you indicate.

Once the method has all the results, it returns the mean from all the predictions and the max and min difference from them.

To work with classifiers, we need the split the data into training and testing. The model uses the training set to understand the data and be able to make predictions about the target output of the testing data set.

The acquired predictions are:

- Model Class: LogisticRegression
  Cross Validation: 0.71 (+/- 0.07)

- Model Class: LinearSVC
  Cross Validation: 0.63 (+/- 0.08)

- Model Class: RandomForestClassifier
  Cross Validation: 0.64 (+/- 0.29)

- Model Class: GradientBoostingClassifier
  Cross Validation: 0.60 (+/- 0.19)

All the models have similar outcomes but with some differences.

The model with a higher score is the Logistic Regression. It has obtained a mean of 70% of correct predictions. A part from having the highest score, also have the lowest deviation, meaning that the model is reliable.

The next better one is the Random Forest Classifier. It has a score of 64% but its deviation is 29%. This is a big difference, meaning that the model is not completely reliable.

The Gradient Boosting Classifier has a similar problem as the Random Forest. It has a smaller deviation but also smaller accuracy.

Finally, the Linear SVC has a normal percentage of 63%, however, the difference between the tests it is the smaller one with only an 8%. It does not give the best results but is more stable than the others.

6.4.4. Survey outputs analysis

We have achieved good and considerable results. The linear regressions models have obtained better results than the classifiers, as it can be seen in the chapter ‘6.2.3. Testing the models’.

As we expected in the chapter '6.4.2. Possible correlations, the most reliable models are not the classifiers.

With the obtained predictions averages, this categorical data must be combined with the sensors data, to achieve better predictions.

Every medical test has the part of the sensors and on the other hand questions made to the patient. Based on both parts, doctors can achieve better and more reliable results, having this in mind we will use the sensors outputs and the survey questions to achieve better predictions.

6.5. Computed sensors output

In this topic, the idea is to test the sensors data with a different approach than in the chapter '6.3. Sensors outputs'.

Before, we got each column from the results, which are separated in three parts: right, back and left, depending on the part of the neck, and used the min or max value depending on which sensors were. With the thermographic camera, we used the maximum value and with the inertial biomechanics and the EEG headset, the minimum results.

The idea is to create new columns based on the difference between the max and min values.

Taking a look at the records we can observe that normally if a patient has an injury in the right, back or left, the sensor output is much different than the other two parts, while on the other hand, if the patient has no injuries, the three parts are more or less in the same range, the difference is smaller.

6.5.1. Prepare the data

As explained in the past chapter, we will use the difference between columns from the same sensor, to achieve better predictions.

After applying the method to the data, we obtain:

LesionsCervicals	Max_Dif_Temp	Max_Dif_Des	Max_Dif_EEG_Desv	Max_Dif_EEG_Interquartil
0	1.25	24.558523	70.14	24.62
0	0.14	19.336932	23.55	26.15
0	1.14	7.218182	40.18	19.62
0	0.69	20.862626	9.04	19.48
0	1.43	17.627273	211.14	65.64
0	0.59	6.838194	100.51	58.98
0	1.97	8.492929	56.12	38.46
0	0.61	13.952273	135.90	116.79
0	0.23	23.442361	49.05	7.69
1	0.79	25.460417	205.66	169.23
1	0.17	14.638889	22.84	17.95
0	4.22	9.750505	175.51	101.02
0	2.07	4.193939	38.51	8.71
0	0.33	21.168750	47.04	60.00
2	0.80	7.122222	48.52	115.39
2	0.83	12.662500	71.85	172.31
1	2.27	4.829861	27.30	13.84
1	4.63	10.212121	4.73	106.15
2	1.02	17.406818	60.27	51.28
0	0.85	23.615152	77.59	201.02
0	1.22	28.944886	77.81	274.36
1	1.48	16.835354	21.32	14.36
0	1.01	32.472222	91.27	40.51
1	0.68	15.240404	12.83	10.25
2	1.34	20.774306	9.83	14.36
0	0.78	6.326389	23.89	43.08
2	0.27	18.347222	8.80	6.15

Table 6.5.1.1. Table of the final processed data with the difference between columns.

6.5.2. Possible correlations

As done in past chapters, we are searching for linear correlations. Using the computed data, explore if they can affect directly to the target 'LesionsCervicals'.

The next Heatmap represents the predictors and the target in color map, showing the correlations between the records.

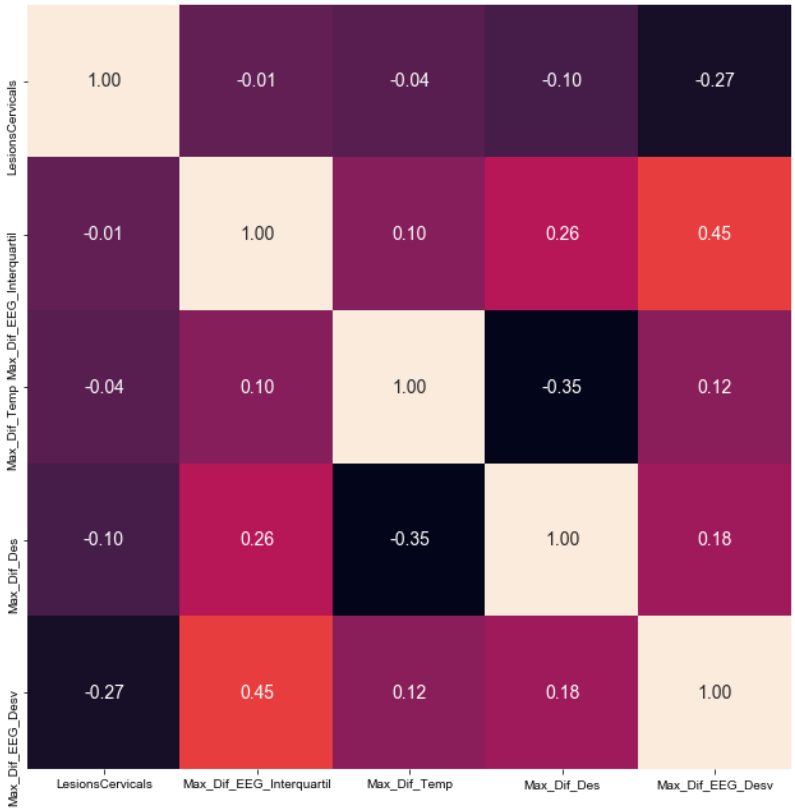

Fig. 6.5.2.1. Correlation Heatmap between the computed sensors predictors and target.

The results demonstrate that there is no direct correlation. The first row, which is the target with all the predictors, only have negative outputs. This means that most probably the linear models will not give good predictions.

6.5.3. Testing the models

As in the chapter ‘6.4.3. Testing the models’ we will use a python class to test different models faster and easier.

The tested models are:

- LogisticRegression
- LinearSVC
- RandomForestClassifier
- GradientBoostingClassifier

To work with classifiers, we need the split the data into training and testing. The model uses the training set to understand the data and be able to make predictions about the target output of the testing data set.

After running the class using the cross validation method, to achieved more reliable results, the acquired predictions are:

- Model Class: LogisticRegression
  Cross Validation: 0.30 (+/- 0.25)

- Model Class: LinearSVC
  Cross Validation: 0.41 (+/- 0.42)

- Model Class: RandomForestClassifier
  Cross Validation: 0.53 (+/- 0.27)

- Model Class: GradientBoostingClassifier
  Cross Validation: 0.33 (+/- 0.20)

The predictions have small averages and they can not be used. The best model is the Random Forest Classifier but only with a 53%.

As we predicted in the chapter ‘6.3.2. Possible correlations’, the models that use linear regression, have obtained low results because there are no correlations between the predictors and the target. The only with some positive prediction has been a classifier.

6.5.4. Computed sensors output analysis

We cannot use or have in mind this approach to work with the data as the results are excessively low.

Taking a look to data most of the patients corroborate the idea where if they have an injury the difference is bit whereas if they do not have an injury the difference is smaller, but then we have some patients that do not have the expected outputs.

This is probably because the data was not taking precisely enough with these users. This kind of data must be obtained in perfect environments, where the room temperature is exactly at one level and the exercise must be perfectly executed to avoid errors in the data.

6.6. Sensors and survey outputs

In this chapter, we will use the learned before and use all the data to achieve better predictions.

Here we will use the technique applied in the chapter ‘6.3.2. Prepare the data’ and ‘6.4.1. Prepare the data’. This means that we will use the minimum outputs the EEG and inertial data and the maximum values for the thermographic camera tests. In the other hand, the survey outputs will be transformed from categorical to numeric, to be able to work with them.

Besides the past chapters where we searched for a correlation, in this final approach is not necessary as the Heatmap search for a linear correlation, meaning that the predictor influences directly to the target.

The results would be the same than the past computed Heatmaps. The sensors outputs do not have a strong correlation, whereas the survey results have stronger relationships.

6.6.1. Testing the models

We will use the same technique as in all the other 'Testing the models' chapters. With the help of a python class, we will test multiple models directly and validate them using the cross-validation method.

The tested models are Logistic Regression, Linear SVC, Random Forest Classifier and Gradient Boosting Classifier.

To work with classifiers, we need the split the data into training and testing. The model uses the training set to understand the data and be able to make predictions about the target output of the testing data set.

The obtained predictions are:

- Model Class: LogisticRegression
  Cross Validation: 0.55 (+/- 0.30)

- Model Class: LinearSVC
  Cross Validation: 0.59 (+/- 0.25)

- Model Class: RandomForestClassifier
  Cross Validation: 0.59 (+/- 0.02)

- Model Class: GradientBoostingClassifier
  Cross Validation: 0.63 (+/- 0.17)

The results have not been as expected. The correct prediction average is small and not completely reliable.

The classifiers have obtained better results. The Gradient Boosting Classifier has a scoring average of 63% and a deviation of 17%. The random Forest Classifier has obtained a lower score but also with a lower deviation, meaning that is more stable.

6.6.2. Sensors and survey outputs analysis

This approach tried to achieve better results using the different columns. The obtained results are not as good as expected.

The main reason to explain this is that we have a total of 40 columns and only 27 records, after preparing the data, we have a total of 25 columns. Even though, we still have too many values and not enough patients.

This situation makes that the models have more problems to give correct predictions as the training set has not enough rows for them to learn. In addition to not having a sample enough big, some data have inconsistent values, where it does not have the normal outputs.

6.7. Clustering

Clustering is grouping a set of objects based on their characteristics. In our case, based on the predictors' values the method will assign them into one of the target classes.

Is an exploratory analysis that tries to identify structures within the data, homogenous groups of cases.

The clustering method used in our case is the K-means Clustering. It is a type of unsupervised learning. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

It assigns the classes based on the proximity of the values and the similarities between them.

6.7.1. Prepare the data

The data for clustering is modified exactly as we did before at the section ‘6.2.2. Prepare the data’.

We have four new columns based on the sensors results and the possible limits of them, between having a cervical fracture or not.

The limits are we have seen before, are:

- Thermography:

  The users with cervical fractures show a difference of 0.8ºC approximately, while the others are around 0.4ºC.

- EGG:

  The users with cervical fractures indicate a much lower degree of mobility and a lower standard deviation.

- Inertial:

  The users with important cervical fractures have less than a 75% of the mobility.

The rest of the columns have been removed from the data.

6.7.2. K-means Clustering

The method used to achieve the clustering is from the library Sklearn.

The idea is to visualize the results. For doing this, we need to indicate the number of clusters the data have. As our target is 'LesionsCervicals' we know there are just three clusters, corresponding to zero, one and two, as we decided in the chapter '6.1.1. Data description'.

The k-means model knowing there are three classes, will divide the data into those possible outputs.

Once the number of clusters is decided, to visualize the data, we do a scatter plot, dividing the classes in the colors red, lime and black.

For this cluster are used the predictors 'Max_Dif_EGG_Interquartil' and 'Max_Dif_Des', because they were the most important variables in the decision from the Random Forest Classifier as we saw it at the chapter '6.3.3. Random Forest Classifier'.

This first graphic is the real classification using the target values.

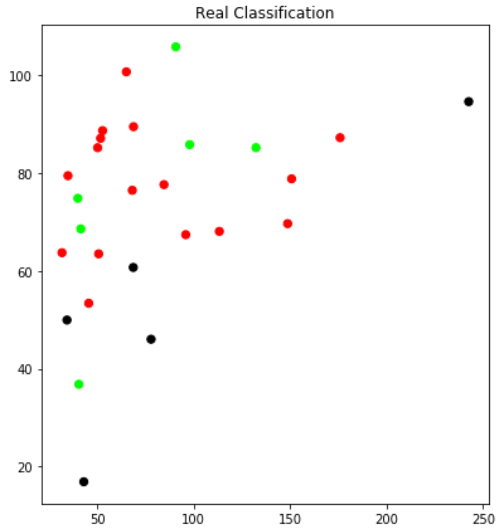

Fig 6.7.2.1. Scatter plot based on 'Max_Dif_EEG_Interquartil' and 'Max_Dif_Des'.

As we can see, the target is quite dispersed and does not seem to follow any specific pattern. This will make difficult for the method K-Means to obtain the correct classes.

In the next graphic, we have the real classification scatter plot and the K-Means classification plot.

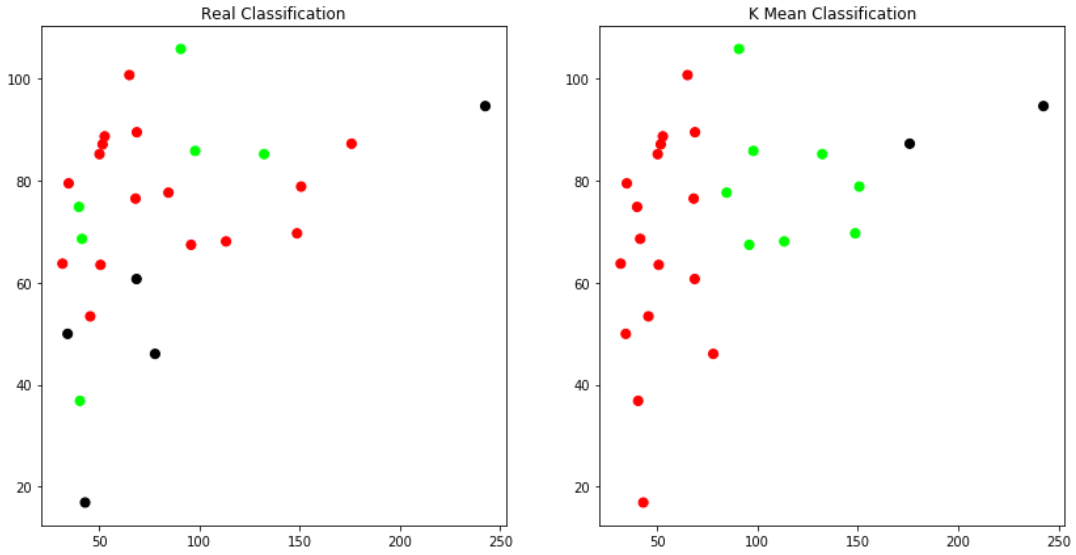

Fig 6.7.2.2. The real scatter plot classification and the K-Means scatter plot classification.

As we expected the K-Means have not classified correctly the different rows.

The black dots represent the value 2 from ‘LesionsCervicals’, meaning that those users have a higher degree of fracture. These users should all have the lowest values in the graphic, but for example, there is one user with the values 242.56 and 94.6 for ‘Max_Dif_EEG_Interquartil’ and ‘Max_Dif_Des’ respectively.

The same happens with the green ones. Those are the users with a low or medium fracture degree. They should have low values, but there a couple of users with unexpected results.

As the above results were not as expected, the next plot is another clustering using the method K-Means but adding another predictor, the ‘Max_Dif_Temp’. This creates a three dimensional scatter plot.

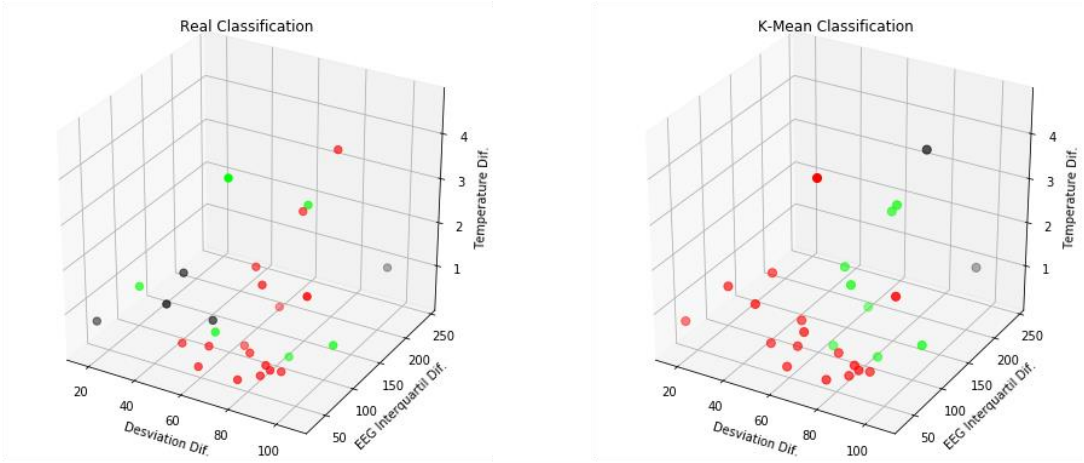

Fig 6.7.2.3. The real scatter plot classification and the K-Means scatter plot classification using three predictors.

The same problem happened like before. In this case, we also have some users with no injury, the red dots, with quite high differences of temperature, when they should not exceed the 0.8ºC, as that is the limit we studied in the chapter ‘6.2.2. Prepare the data’.

7. Analysis

In this chapter the results of the chapter ‘6. Development’ section is discussed and analyzed. We want to know why we got those values and how the data have affected them.

As we discovered in the section ‘6.1. Data Exploratory’ and more precisely at the subsection ‘6.1.1. Data Description’, the starting records were formed by 41 columns, 21 of them are survey questions and 27 rows.

This is a small dataset. The models do not have enough training data and the predictions have low scorings.

Later in the chapter ‘6.3.3 Developing a model’, where we used the Random Forest Classifier, as there were small correlations between the predictors and the target, we got a prediction of 66%. This is not a low accuracy but as we used the total of rows is small we could only have twenty-one users for training and six users for testing, making the decision tree small and difficult to predict the target correctly.

The categorical data approach obtained similar results, but different from the sensors outputs, this data has stronger linear correlations and the classifier models did not perform as good as before.

The third used technique, computed sensors outputs, could have performed really good but the predictions were negligible. That chapter showed us that the data was not perfectly captured.

For example, if the limits from the thermographic camera are 0.8ºC, we should not have a patient with no injuries and a difference from more than 1ºC. The same happens with the EEG and the biomechanical sensors.

The final method, the chapter ‘6.6. Sensors and survey outputs’, obtained small averages and the models did not perform as expected. The main reason is the dataset extension. The models only had 20 rows for training and achieving a prediction based on a total of 24 columns.

In the chapter ‘6.7. Clustering’ we tried to achieve a classification using the method K-Means. The classification did not match completely with the real one. The main problem

is that the output of the sensors from some users was not as expected, having incoherencies with respect to the other users' values.

Taking a closer look the figures ‘Fig 6.7.2.2.’ and ‘Fig 6.7.2.3.’ from the clustering chapter, we can observe that the real classification differs from the classification proposed by the K-Means algorithm, but it is normal as in the plot there is no sign of order and we have a lot of users causing noise in the dataset.

It is difficult to achieve a correct model, as the data set is small and some the users have wrong results from the sensors, leading to mistakes.

Possible extensions could be the obtaining of a larger dataset in an environment where the degree of error would be minimal.

8. Objectives achievement

The project started with a list of objectives. These objectives have been changed during the development of the project, depending on time or results.

At the beginning of the project, we defined a series of requirements. We will check each of them and see if we have accomplished the objectives or not.

- Use data exploratory method to understand the dataset

  This requirement has been accomplished.

  In the chapter '6.1. Data Exploratory', we took a look at the data to discover the different values, their type and the length of the dataset.

  It has been an easy requirement but at the same time fundamental for the rest of the development. Thanks to it, we understood the data, how to group it into different sections and how to transform the values

- Transform the values to be able to work with them

  This requirement has been accomplished.

  In each chapter from the development part, we prepared the data to be able to work with the models.

- Divide the data into training/testing sets

  This requirement has been accomplished.

  We have used this technique in different moments during the development. Given the fact that we have used supervised models, this requirements has been essential to obtain the desired results.

- Obtain predictions from the different statistical models

  This requirement has been accomplished.

  In the chapters ‘6.3. Sensors Outputs’, ‘6.4. Survey Outputs’, ‘6.5. Computed Sensors Outputs’ and ‘6.6. Sensors and Survey Outputs’ we used the predictions from the different models to analyze their correct predictions average and compare the different approaches results to detect which models combined with the different possibilities obtain better predictions.

- Obtain multiple predictions from different models directly

  This requirement has been accomplished.

  To test the different models easier and faster we developed a python class where indicating the dataset and the different models it returns the average predictions from the models.

- Comparative between the different models' results

  This requirement has been accomplished.

  At the end of each chapter from the development, we have an analysis of the results obtained and at the end a comparison between all the different approaches.

- Cluster the data to compare the results between the real classification and the predictions

  This requirement has been accomplished.

  In the chapter ‘6.7. Clustering’ we used the K-Means algorithm to compare the real classification with the predicted classification.

Apart from the agreed requirements, this project consists of basically one main objective: validate the reliability of the use of multisensor with cervical injuries.

The objective has not been accomplished.

Despite the fact we achieved predictions and the models seemed to understand the data and be able to offer correct results the target, the data is not big enough.

As this data is not representative of the population we cannot corroborate the theory.

Even though this data is not a representative dataset, this is the beginning of a possible future project. We could use the achieved knowledge to applied directly to larger datasets and test the models in a more real situation. It can be seen as a template of how to transform this type of values and apply it to the different models.

9. Conclusion

This chapter concludes the whole project. It explains all those conclusions and learning, that have arisen and have been generated throughout the realization of it.

It is important to remember that the conclusions drawn in this whole process can only be assured that they are valid, within the same project. Therefore, it should be taken into account that the sample used was not representative, and therefore the results obtained cannot be considered as a possible extrapolate from the population in general.

Even though the results have not been as expected, the idea of detecting cervical fractures with this clean, environment-friendly methods, can be studied in a more complex project and obtain positive outcomes.

The knowledge would allow doing constant and representative follow-ups of the evolution of those affected, during the rehabilitation periods.

As explained in the chapters ‘7. Analysis’ and ‘8. Objectives Achievement’ the main objective, giving reliability to the proposed theory, has not been achieved. Even though we obtained positive predictions and working models with the dataset, the total length from the records does not allow realistic outputs. We trained the model with 21 rows and tested with six patients, that is not enough to determine the correctness from the dataset.

This project has been based on a starting sample, alpha, dataset. The length of the table do not allow the construction of good predictive models and the data its self contains errors. With that said, this project cannot elaborate solid conclusions, based on the starting records, about the reliability of the technique to discover cervical injuries proposed at the beginning.

10. Bibliography

[1] Upper Cervical Spine Trauma Imaging [online] [search: December 19, 2017]. Available at https://emedicine.medscape.com/article/397563-overFri.w

[2] Electroencephalograph [online] [search: December 19, 2017]. Available at *https://medical-dictionary.thefreedictionary.com/electroencephalograph*

[3] Electrónica Fácil. [online] [search: December 20, 2017] Available at *https://www.electronicafacil.net/foros/PNphpBB2-Fri.wtopic-t-2568.html*

[4] Qué es. Para qué sirve. Ámbitos de actuación [online] [search: December 20, 2017]. Available at http://www.mibienestar.es/salud/2-general/2-biomecanica.html

[5] What is statistical modeling [online] [search: *February* 10, 2018]. Available at *https://help.xlstat.com/customer/en/portal/articles/2062460-what-is-statistical-modeing-?b_id=9283*

[6] Agile Methodology [online] [search: *February* 10, 2018]. Available at *http://agilemethodology.org/*

[7] GiveWhenThen [online] [search: *February* 12, 2018]. Available at *https://martinfowler.com/bliki/GivenWhenThen.html*

[8] Introducción a Domain Driven Design (DDD): Parte 1 [online] [search: *February* 12, 2018]. Available at https://devexperto.com/domain-driven-design-1/

[9] Which Languages Shoould Tou Learn For Data Science [online] [search: December 22, 2017]. Available at https://medium.freecodecamp.org/which-languages-should-you-learn-for-data-science-e806ba55a81f

[10] *The R Project for Statistical Computing* [online] [search: December 20, 2017]. Available at https://www.r-project.org/

[11] *Python* [online] [search: December 20, 2017]. Available at https://www.python.org/

[12] *R vs Python for Data Science: Summary of Modern Advances* [online] [search: December 20, 2017]. Available at https://elitedatascience.com/r-vs-python-for-data-science

[13] What is exploratory data analysis? [online] [search: *March* 15, 2018]. Available at *https://www.sisense.com/blog/exploratory-data-analysis/*

[14] Supervised and Unsupervised Machine Learning Algorithms [online] [search: *March* 20, 2018]. Available at https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

[15] Logistic Regression for Machine Learning [online] [search: *March* 24, 2018]. Available at https://machinelearningmastery.com/logistic-regression-for-machine-learning/

[16] Support Vector Machines (SVM) Introductory Overview [online] [search: *March* 24, 2018]. Available at http://www.statsoft.com/Textbook/Support-Vector-Machines

[17] Building Random Forest Classifiers in Python [online] [search: April 2, 2018]. Available at http://dataaspirant.com/2017/06/26/random-forest-classifier-python-scikit-learn/

[18] Gradient Boosting from Scratch [online] [search: April 5, 2018]. Available at https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d

[19] Introduction to K-means Clustering [online] [search: April 5, 2018]. Available at https://www.datascience.com/blog/k-means-clustering

[20] Pearson Product-Moment Correlation [online] [search: April 6, 2018]. Available at https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

TecnoCampus
Escola Superior Politècnica

Centre adscrit a la
upf. Universitat Pompeu Fabra Barcelona

Degree in Computing Engineering of Management and Information Systems

Analysis of multisensor data on cervical injuries

Analysis of viability

Eloi Rodríguez Gaxas

TUTOR: Xavier Font

Carles Paul

2017/2018

TecnoCampus
Mataró-Maresme

Table of Contents

Index of tables

1. Initial Planning

The project is divided into these stages:

- Project’s plan
- Analysis and design
- Encoding, testing and deployment
- Project’s closing

Using the work break down technique the following activities have been described.

The activities programmed in each stage are:

Activity Name	Duration	Start	End
Project's definition	**80 days?**	**Mon 06/11/17**	**Fri 29/12/17**
Project's object	6 days	Mon 06/11/17	Wed 08/11/17
Previous study	11 days	Thu 09/11/17	Thu 16/11/17
Objectives and scope	9 days	Thu 16/11/17	Wed 22/11/17
Methodology	8 days?	Thu 23/11/17	Tue 28/11/17
Functional and technological requirements	8 days	Wed 29/11/17	Mon 04/12/17
Study of project's viability	8 days?	Tue 05/12/17	Fri 08/12/17
Stages definition	6 days?	Mon 11/12/17	Wed 13/12/17
Work break down	10 days?	Thu 14/12/17	Wed 20/12/17
Project's calendar	7 days	Thu 21/12/17	Tue 26/12/17
Software methodology	7 days	Tue 26/12/17	Fri 29/12/17
Analysis and Design	**6 days**	**Mon 01/01/18**	**Wed 03/01/18**
GWT from requirements	6 days	Mon 01/01/18	Wed 03/01/18
Encoding, testing and deployment	**124 days?**	**Thu 04/01/18**	**Fri 30/03/18**
Data Exploratory	30 days?	Thu 04/01/18	Wed 24/01/18
Identify all the columns	10 days?	Thu 25/01/18	Wed 31/01/18
Describe the data	14 days?	Thu 01/02/18	Fri 09/02/18
Search correlations	10 days?	Mon 12/02/18	Fri 16/02/18
Identify the dependent and explanatory variables	8 days?	Mon 19/02/18	Thu 22/02/18
Decide between the models	10 days?	Fri 23/02/18	Thu 01/03/18
Prepare the data	4 days	Fri 02/03/18	Mon 05/03/18
Divide the data in training/testing	5 days	Tue 06/03/18	Thu 08/03/18
Fit the model	3 days	Thu 08/03/18	Fri 09/03/18
Extract predictions	3 days	Mon 12/03/18	Tue 13/03/18
Analyze the results	7 days	Tue 13/03/18	Fri 16/03/18
Decide the clustering method	5 days	Mon 19/03/18	Wed 21/03/18
Fit the classifier	3 days	Wed 21/03/18	Thu 22/03/18
Visualize the real scatter plot	3 days	Fri 23/03/18	Mon 26/03/18

Visualize the predicted scatter plot	4 days	Mon 26/03/18	Wed 28/03/18
Compare the outputs	5 days	Wed 28/03/18	Fri 30/03/18
Project's closing	**38 days?**	**Mon 02/04/18**	**Thu 26/04/18**
Project conclusions	10 days?	Mon 02/04/18	Fri 06/04/18
Preparation of the defense-presentation of the project	28 days?	Mon 09/04/18	Thu 26/04/18

Table 1.1. Calendar of the project.

In the encoding, testing and deployment have been used an agile technique.

This stage has been deployed in sprints, where the test of each module, correction and presentation and acceptance, are done in each sprint to evaluate every activity.

The team is composed of just one member. This implies that all activities rely on the previous one. All the path is critic.

In the case of having a bigger team, some activities could have been done at the same time.

The team members would be:

- Project manager
- Analyst
- Developer
- Tester

2. Budget

For the project, we only have one developer but doing all the different roles.

Assuming a standard salary of 1.200€ per month for a junior developer, it means that adding all the taxes the company should pay approximately 1.560€ per month.

With an approximate duration of six months, the company should pay approximately 11.000€.

Adding the taxes from the employer's quota of the Social Security, the cost to the company would be around 14.300€ for a duration of 6 months.

The only thing needed is a computer, it has a total cost of 600€, it means that the cost of the technological requirements are 60€, for a duration of six months.

3. Viability analysis

The traumatology world is an extended and settled world with reliable techniques, which gives accurate results.

This project does not intend to replace the current techniques but give a first process, which can avoid making radiographies in specific cases.

These techniques are already being used in the veterinarian area to search anomalies in the animals.

It can save money as there could be patients which where the radiographies are not necessary.

Another important use for these techniques is to check the patient evolution during a recovery process. With the use of the sensors, we could determine if the user is doing progress, meaning that more sessions with the physiotherapist can be positive, or that the person has arrived to its maximum recovery. It may not have returned to its normal state but after an injury not all the patients are able to do it. In the actuality, as we do not have a properly method to do this, we do not know if the sessions are necessary or not.

With this technique we would save resources, the physiotherapist, and the doctors could have a solid criteria.

4. Technical viability analysis

Abstracting the data from the sensors to compute it and get reliable results has not been done. However, computing the data and getting plots it is not a problem.

With python there are a lot of packages and tutorials showing how to do it. The first part of the project, computing the data, is not a problem.

The problematic part is obtaining a reliable result from the data treated by the script.

The major problem is the lack of data. With only 28 rows and 27 columns can be a problem on obtaining reliable outputs.

In the other hand, we have the use of the sensors, which need to be used following a methodology and a set of metrics, to obtain reliable results. After having prepared and formed the operators, the tests are easy to develop.

5. Economic viability analysis

The project can be implemented in traumatology consults.

It is a large filed with lots of possibilities and the results could change the way things are done nowadays.

With less costs as the sensors are cheaper and there would be no necessity to print all the radiographies and use of CDs.

Radiographies are expensive processes. As it is the only process capable to determine if there is an injury or not, they are done to lot of patients where maybe there is no need to. With this project a first phase can be done and use the radiography with the patients that has the need.

The economical expenses from this project are the sensors. This electronics are expensive but it would be only one first investment and the subsequent maintenance.

It may be difficult to start implementing this process but with the past of time, it would recover the investment based on the lack of necessity to do radiographies to all the patients and using them only for the users with higher needs.

6. Environmental viability analysis

With this project's technique the environmental waste would be much smaller since there would be no need to print everything and use the technique with all the patients.

The environmental footprint for creating and using the sensors is minimal.

Neither the thermographic camera, nor any of the two types of sensors, produces any kind of residue.

Regarding the social effect of the study and all the generated knowledge, regarding the relationships between data that can be obtained with inertial sensors, thermographic camera and EEG sensors; It has been verified and assessed that the effect would be positive, since the ability to diagnose injuries of all types and in various areas of the body was better.

For this project all the major residues comes from the printing of the project, the paper used, the CD and the needed electricity for using the computer and all the electronics.

7. Legal aspects

For the reason that this project uses sensitive personal information, it is included in the data protection law.

Data protection:

- Organic Law 15/1999, of December 13, on the protection of personal data.

TecnoCampus
Escola Superior Politècnica

Centre adscrit a la
upf. Universitat Pompeu Fabra Barcelona

Degree in Computing Engineering of Management and Information Systems

Analysis of multisensor data on cervical injuries

Annexes

ELOI RODRÍGUEZ GAXAS
TUTORS: XAVIER FONT
CARLES PAUL

2017/18

TecnoCampus
Mataró-Maresme

Table of Contents

Annex I. CD content

- Project report
- Source code and executable code
- PDF with instructions for the executable code
- CSV with the used dataset