Emines School Of Industrial Management
Statistics Module

Exploring the effect of Sway Descriptors on Risk of Fall in Elderly

AIT AOMAR Anas
MOUFAD Badr
MOUSTATIA Zoubir

Emines, April 2021

Abstract

Falls have become a source of serious health complications for the elderly and can even cause death. As a result, medical researches are increasingly focusing on this subject mainly how to diagnosis stability disorder for elderies. Among the machines used to carry out this diagnostic we find the stabilograme. It is a device equipped with several sensors making it possible to perform various measurements ranging from the surface where the patient's center of mass moves to the frequency of movement and even its speed. These doctors allied to statisticians then try to set up a model which, thanks to these different measurements, can predict whether the patient is more prone to fall in the future.

This subject problematic lies in the fact that the parameters and measurements collected by the stabilograme are too numerous (approximately 160 parameters) and it would not be very efficient to use them all in a predictive model. So, in what follows in this report we will propose different methods to detect the variables that have the most impact on our predictive model. We will make sure to study not only the individual contribution of each variable but also different combinations of parameters whether they are linear combinations or not. To finally be able to offer a set of variables that allows us to obtain the best precision of the predictive model.

Acknowledgements

Before walking through the chapters of this article, we would like to express our sincere gratitude to our supervisor Mrs. **Ikram CHAIRI** for her instructive guidance, insightful remarks, and fruitful discussions. She was minutely engaged in supervising our work and constantly available for our inquiries. We thank her particularly for trusting and sharing with us her work and experience in the field. Without any exaggeration, it was the backbone of our work.

Our huge thanks to our statistics professors Mr. **Saad BENJELOUN** and Mr. **Khalil SAID** for the valuable knowledge they gave us throughout the semester. The introductive courses on statistics namely estimators, hypothesis testing, linear regression, and especially the case studies were tremendously useful in our work.

We extend our gratitude to our school **EMINES** – School of Industrial Management and our university **Mohammed VI Polytechnic University** for offering us this exclusive learning opportunity and providing us with the staff and infrastructure to conduct our work in the best conditions.

Contents

List of Tables

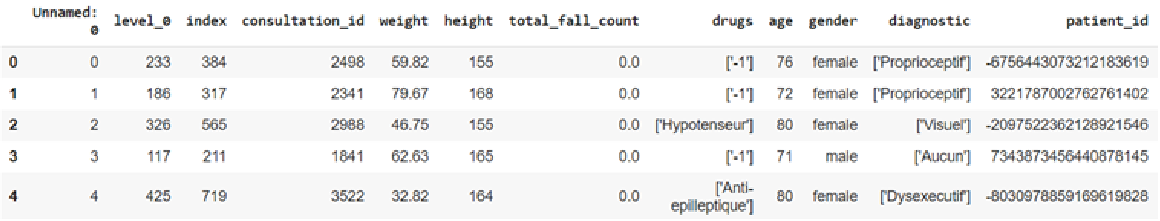
List of Figures

Chapter 1

Presentation of the data set

Our data set contains 100 observations and 158 columns and can be divided into two categories. A first concerning the patient's data and a second concerning the data coming from the stabilogram sensors.

1.1 Patient's data



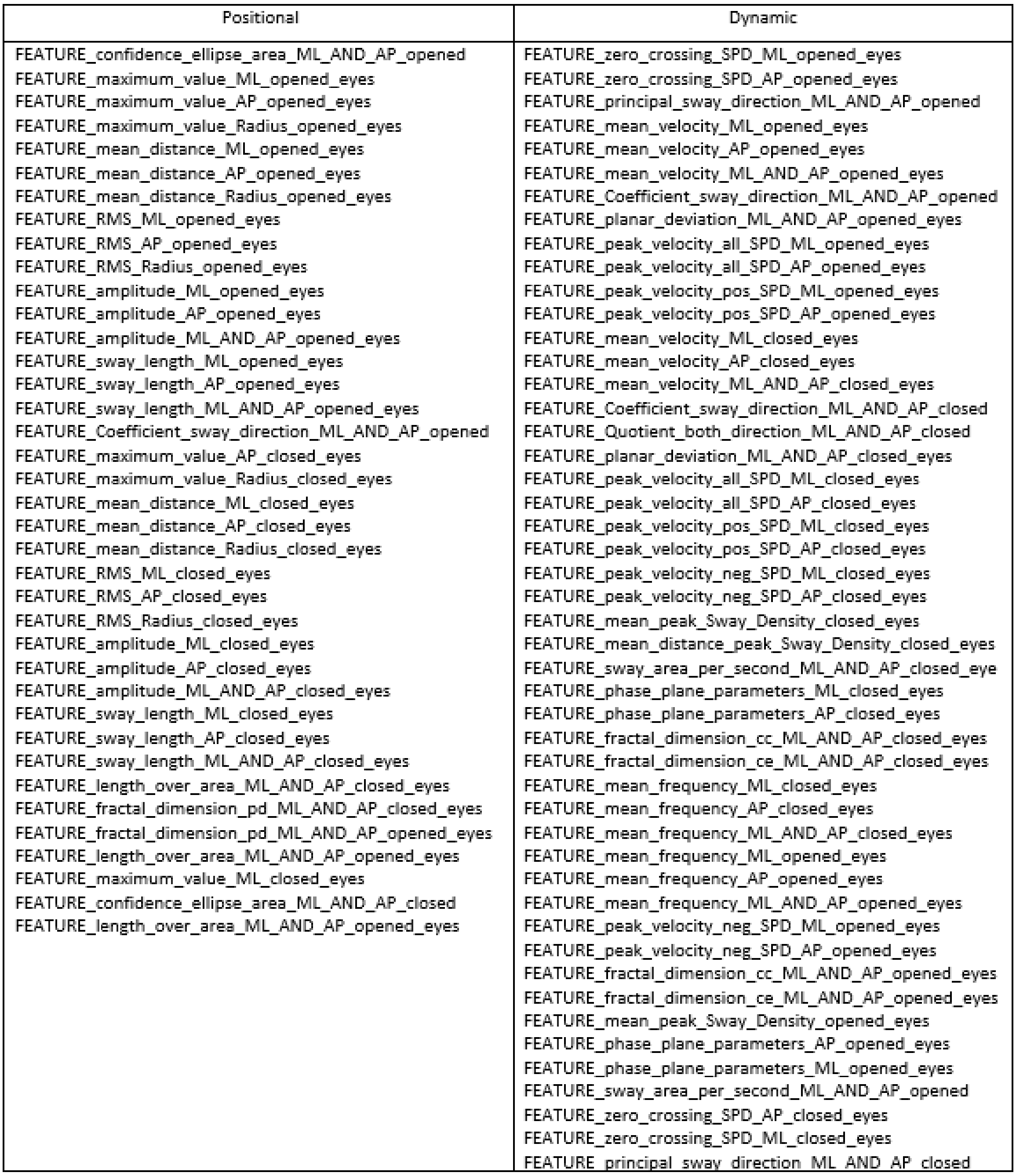
	Unnamed: 0	level_0	index	consultation_id	weight	height	total_fall_count	drugs	age	gender	diagnostic	patient_id
0	0	233	384	2498	59.82	155	0.0	['-1']	76	female	['Proprioceptif']	-6756443073212183619
1	1	186	317	2341	79.67	168	0.0	['-1']	72	female	['Proprioceptif']	3221787002762761402
2	2	326	565	2988	46.75	155	0.0	['Hypotenseur']	80	female	['Visuel']	-2097522362128921546
3	3	117	211	1841	62.63	165	0.0	['-1']	71	male	['Aucun']	7343873456440878145
4	4	425	719	3522	32.82	164	0.0	['Anti-epilleptique']	80	female	['Dysexecutif']	-8030978859169619828

Figure 1.1: Head of the data set containing data related to the patient.

This data set is made of continuous variables such as age, weight, height, and total fall count which is our target variable that represents the number of times that patient fell, and categorical variables like gender, drugs which symbolizes the medical prescription followed by the patient and diagnostic that represents the medical history of that patient. We also have other columns such as patient id and consultation id that won't be useful in our study because in our case each patient has only one consultation.

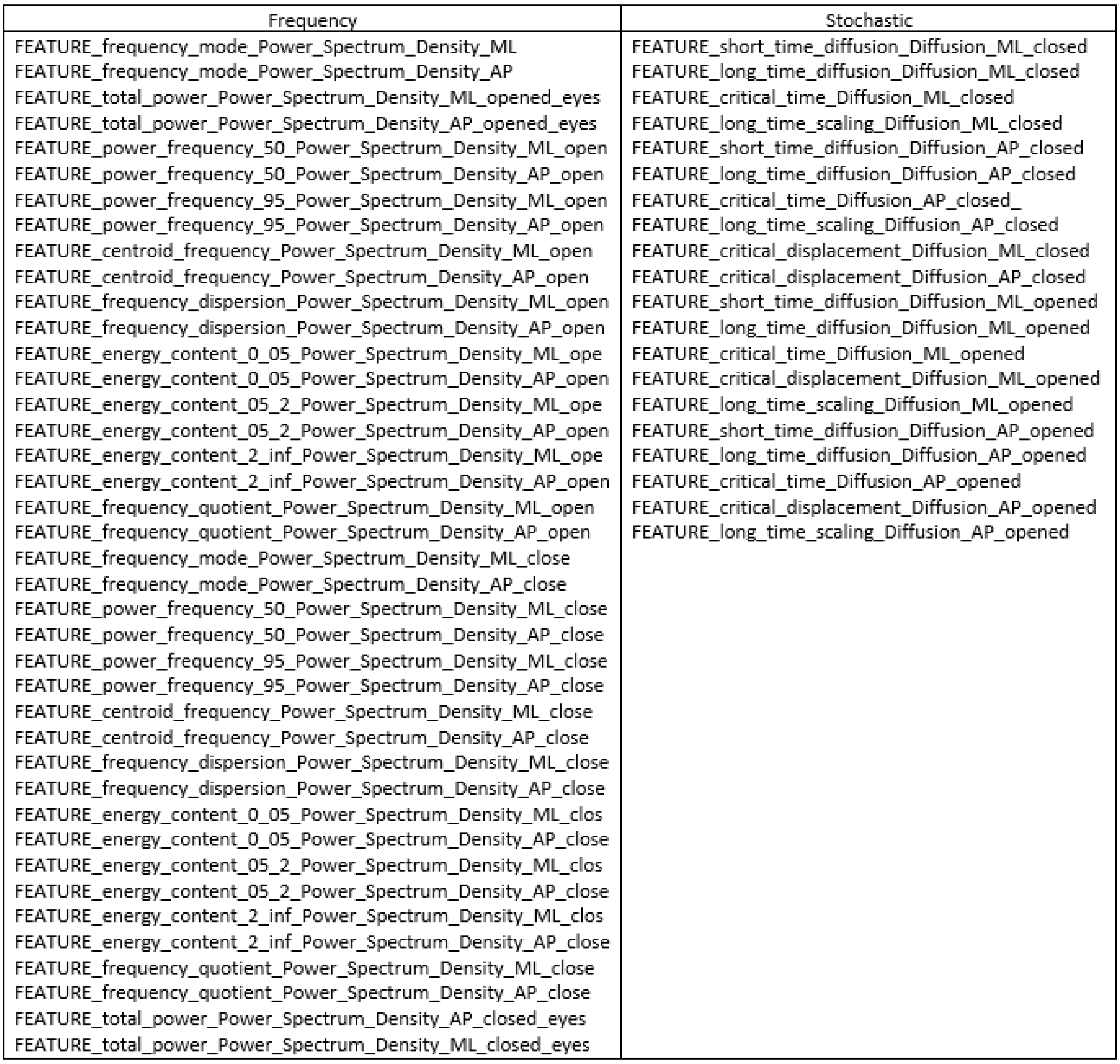
1.2 Stabilogram's data

The parameters coming from the stabilogram are 146 columns in number. It is quite complicated to distinguish between these different parameters. However, some literature research has shown that these can be subdivided into 4 different clusters. In all that follows we will be using this clustering to carry out our studies. The cluster distribution is as follows:



Positional	Dynamic
FEATURE_confidence_ellipse_area_ML_AND_AP_opened	FEATURE_zero_crossing_SPD_ML_opened_eyes
FEATURE_maximum_value_ML_opened_eyes	FEATURE_zero_crossing_SPD_AP_opened_eyes
FEATURE_maximum_value_AP_opened_eyes	FEATURE_principal_sway_direction_ML_AND_AP_opened
FEATURE_maximum_value_Radius_opened_eyes	FEATURE_mean_velocity_ML_opened_eyes
FEATURE_mean_distance_ML_opened_eyes	FEATURE_mean_velocity_AP_opened_eyes
FEATURE_mean_distance_AP_opened_eyes	FEATURE_mean_velocity_ML_AND_AP_opened_eyes
FEATURE_mean_distance_Radius_opened_eyes	FEATURE_Coefficient_sway_direction_ML_AND_AP_opened
FEATURE_RMS_ML_opened_eyes	FEATURE_planar_deviation_ML_AND_AP_opened_eyes
FEATURE_RMS_AP_opened_eyes	FEATURE_peak_velocity_all_SPD_ML_opened_eyes
FEATURE_RMS_Radius_opened_eyes	FEATURE_peak_velocity_all_SPD_AP_opened_eyes
FEATURE_amplitude_ML_opened_eyes	FEATURE_peak_velocity_pos_SPD_ML_opened_eyes
FEATURE_amplitude_AP_opened_eyes	FEATURE_peak_velocity_pos_SPD_AP_opened_eyes
FEATURE_amplitude_ML_AND_AP_opened_eyes	FEATURE_mean_velocity_ML_closed_eyes
FEATURE_sway_length_ML_opened_eyes	FEATURE_mean_velocity_AP_closed_eyes
FEATURE_sway_length_AP_opened_eyes	FEATURE_mean_velocity_ML_AND_AP_closed_eyes
FEATURE_sway_length_ML_AND_AP_opened_eyes	FEATURE_Coefficient_sway_direction_ML_AND_AP_closed
FEATURE_Coefficient_sway_direction_ML_AND_AP_opened	FEATURE_Quotient_both_direction_ML_AND_AP_closed
FEATURE_maximum_value_AP_closed_eyes	FEATURE_planar_deviation_ML_AND_AP_closed_eyes
FEATURE_maximum_value_Radius_closed_eyes	FEATURE_peak_velocity_all_SPD_ML_closed_eyes
FEATURE_mean_distance_ML_closed_eyes	FEATURE_peak_velocity_all_SPD_AP_closed_eyes
FEATURE_mean_distance_AP_closed_eyes	FEATURE_peak_velocity_pos_SPD_ML_closed_eyes
FEATURE_mean_distance_Radius_closed_eyes	FEATURE_peak_velocity_pos_SPD_AP_closed_eyes
FEATURE_RMS_ML_closed_eyes	FEATURE_peak_velocity_neg_SPD_ML_closed_eyes
FEATURE_RMS_AP_closed_eyes	FEATURE_peak_velocity_neg_SPD_AP_closed_eyes
FEATURE_RMS_Radius_closed_eyes	FEATURE_mean_peak_Sway_Density_closed_eyes
FEATURE_amplitude_ML_closed_eyes	FEATURE_mean_distance_peak_Sway_Density_closed_eyes
FEATURE_amplitude_AP_closed_eyes	FEATURE_sway_area_per_second_ML_AND_AP_closed_eye
FEATURE_amplitude_ML_AND_AP_closed_eyes	FEATURE_phase_plane_parameters_ML_closed_eyes
FEATURE_sway_length_ML_closed_eyes	FEATURE_phase_plane_parameters_AP_closed_eyes
FEATURE_sway_length_AP_closed_eyes	FEATURE_fractal_dimension_cc_ML_AND_AP_closed_eyes
FEATURE_sway_length_ML_AND_AP_closed_eyes	FEATURE_fractal_dimension_ce_ML_AND_AP_closed_eyes
FEATURE_length_over_area_ML_AND_AP_closed_eyes	FEATURE_mean_frequency_ML_closed_eyes
FEATURE_fractal_dimension_pd_ML_AND_AP_closed_eyes	FEATURE_mean_frequency_AP_closed_eyes
FEATURE_fractal_dimension_pd_ML_AND_AP_opened_eyes	FEATURE_mean_frequency_ML_AND_AP_closed_eyes
FEATURE_length_over_area_ML_AND_AP_opened_eyes	FEATURE_mean_frequency_ML_opened_eyes
FEATURE_maximum_value_ML_closed_eyes	FEATURE_mean_frequency_AP_opened_eyes
FEATURE_confidence_ellipse_area_ML_AND_AP_closed	FEATURE_mean_frequency_ML_AND_AP_opened_eyes
FEATURE_length_over_area_ML_AND_AP_opened_eyes	FEATURE_peak_velocity_neg_SPD_ML_opened_eyes
	FEATURE_peak_velocity_neg_SPD_AP_opened_eyes
	FEATURE_fractal_dimension_cc_ML_AND_AP_opened_eyes
	FEATURE_fractal_dimension_ce_ML_AND_AP_opened_eyes
	FEATURE_mean_peak_Sway_Density_opened_eyes
	FEATURE_phase_plane_parameters_AP_opened_eyes
	FEATURE_phase_plane_parameters_ML_opened_eyes
	FEATURE_sway_area_per_second_ML_AND_AP_opened
	FEATURE_zero_crossing_SPD_AP_closed_eyes
	FEATURE_zero_crossing_SPD_ML_closed_eyes
	FEATURE_principal_sway_direction_ML_AND_AP_closed

Figure 1.2: Variables of the positional and dynamic cluster



Frequency	Stochastic
FEATURE_frequency_mode_Power_Spectrum_Density_ML	FEATURE_short_time_diffusion_Diffusion_ML_closed
FEATURE_frequency_mode_Power_Spectrum_Density_AP	FEATURE_long_time_diffusion_Diffusion_ML_closed
FEATURE_total_power_Power_Spectrum_Density_ML_opened_eyes	FEATURE_critical_time_Diffusion_ML_closed
FEATURE_total_power_Power_Spectrum_Density_AP_opened_eyes	FEATURE_long_time_scaling_Diffusion_ML_closed
FEATURE_power_frequency_50_Power_Spectrum_Density_ML_open	FEATURE_short_time_diffusion_Diffusion_AP_closed
FEATURE_power_frequency_50_Power_Spectrum_Density_AP_open	FEATURE_long_time_diffusion_Diffusion_AP_closed
FEATURE_power_frequency_95_Power_Spectrum_Density_ML_open	FEATURE_critical_time_Diffusion_AP_closed_
FEATURE_power_frequency_95_Power_Spectrum_Density_AP_open	FEATURE_long_time_scaling_Diffusion_AP_closed
FEATURE_centroid_frequency_Power_Spectrum_Density_ML_open	FEATURE_critical_displacement_Diffusion_ML_closed
FEATURE_centroid_frequency_Power_Spectrum_Density_AP_open	FEATURE_critical_displacement_Diffusion_AP_closed
FEATURE_frequency_dispersion_Power_Spectrum_Density_ML_open	FEATURE_short_time_diffusion_Diffusion_ML_opened
FEATURE_frequency_dispersion_Power_Spectrum_Density_AP_open	FEATURE_long_time_diffusion_Diffusion_ML_opened
FEATURE_energy_content_0_05_Power_Spectrum_Density_ML_ope	FEATURE_critical_time_Diffusion_ML_opened
FEATURE_energy_content_0_05_Power_Spectrum_Density_AP_open	FEATURE_critical_displacement_Diffusion_ML_opened
FEATURE_energy_content_05_2_Power_Spectrum_Density_ML_ope	FEATURE_long_time_scaling_Diffusion_ML_opened
FEATURE_energy_content_05_2_Power_Spectrum_Density_AP_open	FEATURE_short_time_diffusion_Diffusion_AP_opened
FEATURE_energy_content_2_inf_Power_Spectrum_Density_ML_ope	FEATURE_long_time_diffusion_Diffusion_AP_opened
FEATURE_energy_content_2_inf_Power_Spectrum_Density_AP_open	FEATURE_critical_time_Diffusion_AP_opened
FEATURE_frequency_quotient_Power_Spectrum_Density_ML_open	FEATURE_critical_displacement_Diffusion_AP_opened
FEATURE_frequency_quotient_Power_Spectrum_Density_AP_open	FEATURE_long_time_scaling_Diffusion_AP_opened
FEATURE_frequency_mode_Power_Spectrum_Density_ML_close	
FEATURE_frequency_mode_Power_Spectrum_Density_AP_close	
FEATURE_power_frequency_50_Power_Spectrum_Density_ML_close	
FEATURE_power_frequency_50_Power_Spectrum_Density_AP_close	
FEATURE_power_frequency_95_Power_Spectrum_Density_ML_close	
FEATURE_power_frequency_95_Power_Spectrum_Density_AP_close	
FEATURE_centroid_frequency_Power_Spectrum_Density_ML_close	
FEATURE_centroid_frequency_Power_Spectrum_Density_AP_close	
FEATURE_frequency_dispersion_Power_Spectrum_Density_ML_close	
FEATURE_frequency_dispersion_Power_Spectrum_Density_AP_close	
FEATURE_energy_content_0_05_Power_Spectrum_Density_ML_clos	
FEATURE_energy_content_0_05_Power_Spectrum_Density_AP_close	
FEATURE_energy_content_05_2_Power_Spectrum_Density_ML_clos	
FEATURE_energy_content_05_2_Power_Spectrum_Density_AP_close	
FEATURE_energy_content_2_inf_Power_Spectrum_Density_ML_clos	
FEATURE_energy_content_2_inf_Power_Spectrum_Density_AP_close	
FEATURE_frequency_quotient_Power_Spectrum_Density_ML_close	
FEATURE_frequency_quotient_Power_Spectrum_Density_AP_close	
FEATURE_total_power_Power_Spectrum_Density_AP_closed_eyes	
FEATURE_total_power_Power_Spectrum_Density_ML_closed_eyes	

Figure 1.3: Variables of the frequency and stochastic cluster

Chapter 2

Exploration of the data set

After having briefly presented our data set we will in this part explore these parameters in more detail and that by not only proposing visualizations of certain features but also determining the distribution which allows us to describe the most exactly the variable and the different parameters of this distribution.

2.1 Patient's data

2.1.1 Continuous variables

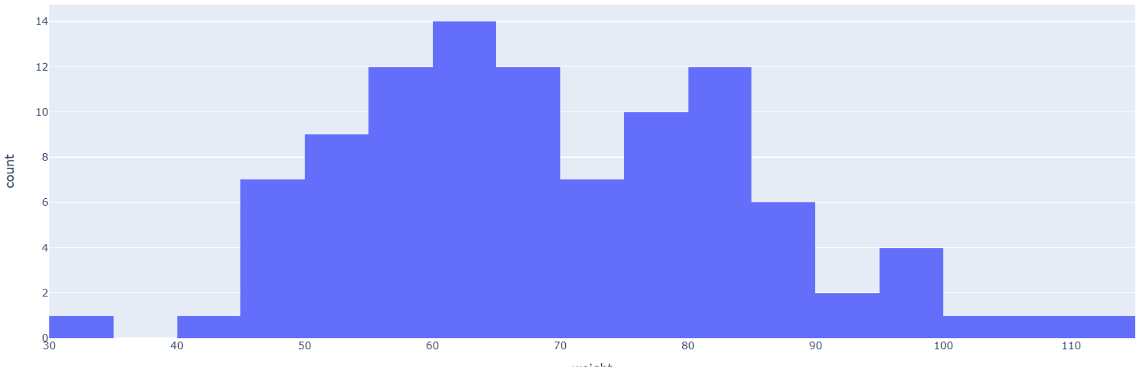


Figure 2.1: Weight of the population

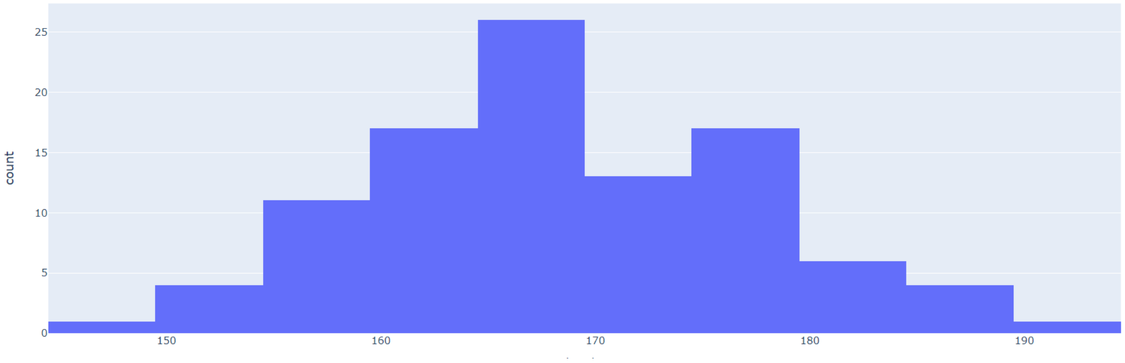


Figure 2.2: Height of the population

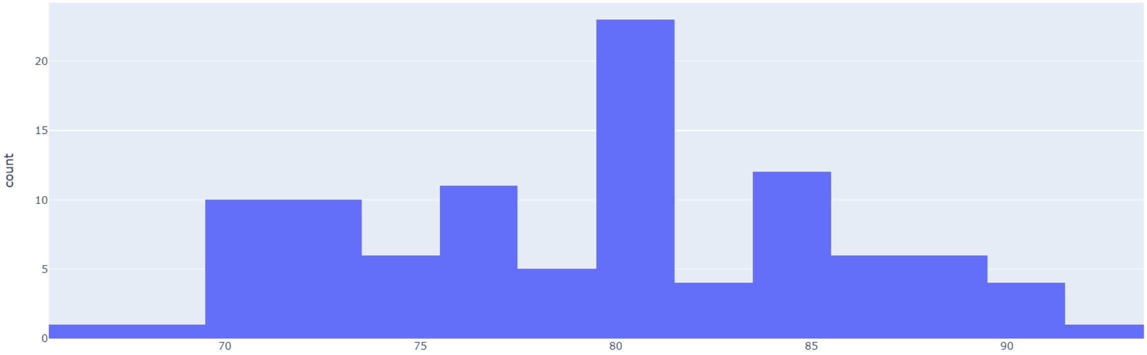
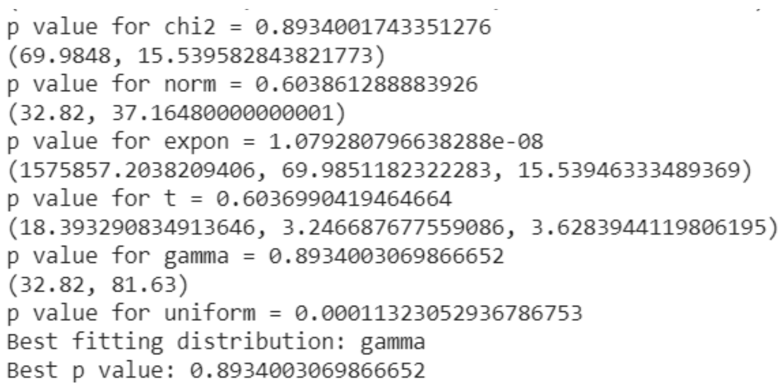


Figure 2.3: Age of the population

The graphs below allow us to have a rather clear idea of the studied population. Indeed this study was carried out on a fairly old population ranging from 66 years to 93 years composed equitably of men and women and having on average a weight of 70 kg and a height of 165 cm. To further analyze these variables we are going to approach them with the most appropriate distribution while determining its parameters and the distribution of one of those parameters. To do so we are using the fit method given by the library scipy that uses **the Maximum likelihood** to determine the parameters of the most common continuous distribution (uniform, normal, exponential, student, gamma, chi2) that fits the best, the variable that we are studying. Once the parameters of the theoretical distribution are determined we apply now **the Kolmogorov test** . The null hypothesis of this test suggests that this variable follow the distribution we are testing which means that while applying this test we will be seeking the distribution that gives us the biggest p-value because in this case, a bigger p-value means that we cannot reject the null hypothesis which implies the more the p-value is big the more we are sure that the variable effectively follows the tested distribution. Testing that on the weight variable gives us the following results:



```
p value for chi2 = 0.8934001743351276
(69.9848, 15.539582843821773)
p value for norm = 0.603861288883926
(32.82, 37.1648000000001)
p value for expon = 1.079280796638288e-08
(1575857.2038209406, 69.9851182322283, 15.53946333489369)
p value for t = 0.6036990419464664
(18.393290834913646, 3.246687677559086, 3.6283944119806195)
p value for gamma = 0.893400306986652
(32.82, 81.63)
p value for uniform = 0.00011323052936786753
Best fitting distribution: gamma
Best p value: 0.893400306986652
```

Figure 2.4: Determining the best fit to the weight variable

In the case of the weight parameter because the gamma distribution has the highest p-value we can conclude that it is the most appropriate distribution for this variable. We had also been able to determine the different parameters of this distribution which are in this case an **Alpha** = 18,39 a **Loc**=3,24 and **Lamda**=3,62 However, these parameters are just estimations and cannot be considered deterministic values. This is why in what follows we are going to apply Bayesian inference on one of those parameters (the alpha parameter) to obtain the range of the possible values of this parameter. To perform a **Bayesian analysis** we would first need the likelihood (the distribution of the variable we are studying in this case the weight variable). The previous study already gave us the fact that it follows a gamma with an **Alpha** = 8,39 and a **Landa** = 3,62. We will then need prior knowledge on the parameter we want to analyze in-depth in our case it will be the alpha parameter of this gamma distribution. Because we don't have any particular information about this alpha parameter we are going to use a non informative prior which is an approach followed when we don t have any idea of the distribution followed by our parameter. In our case, we will be choosing the Uniform prior.

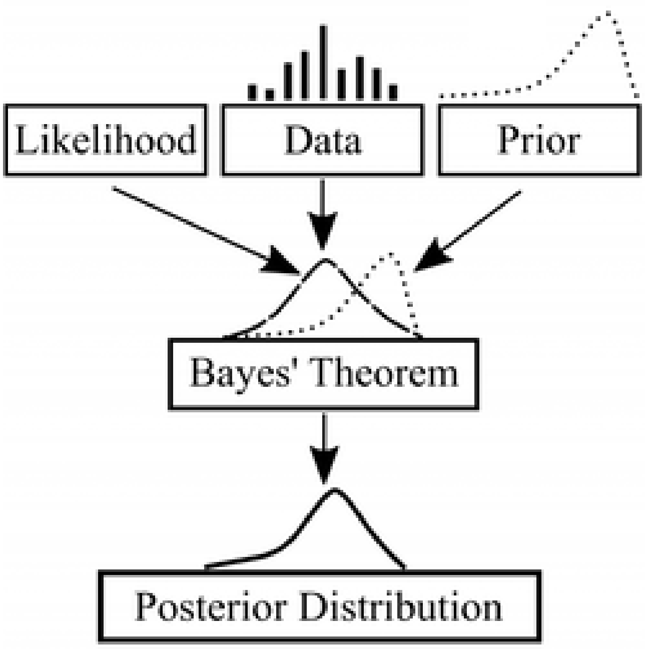


Figure 2.5: Process of Bayesian inference

Knowing the likelihood hood and the prior we can now plot the distribution of the alpha parameter of our weight variable:

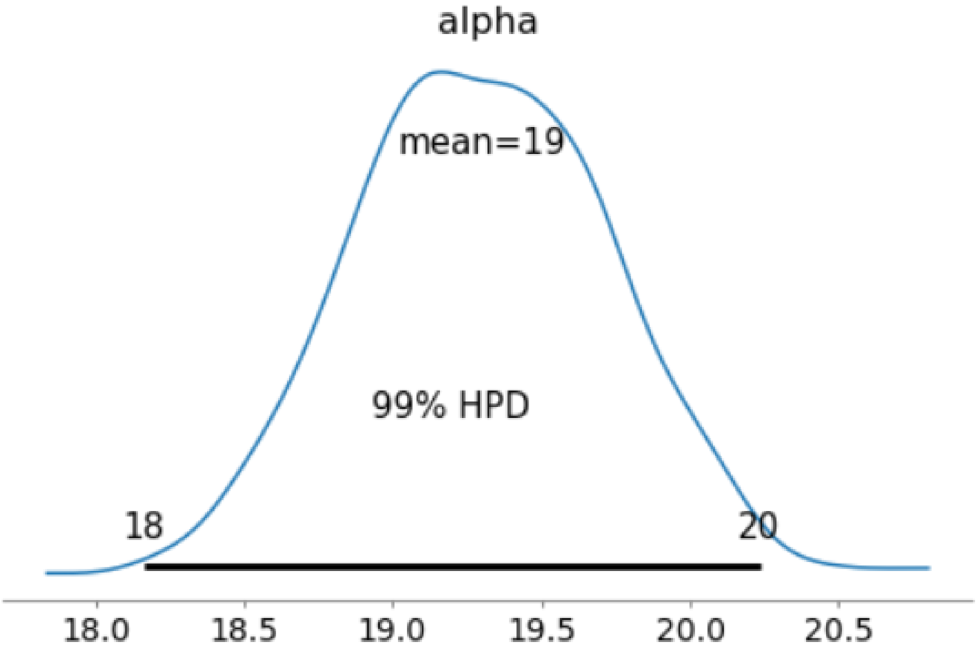
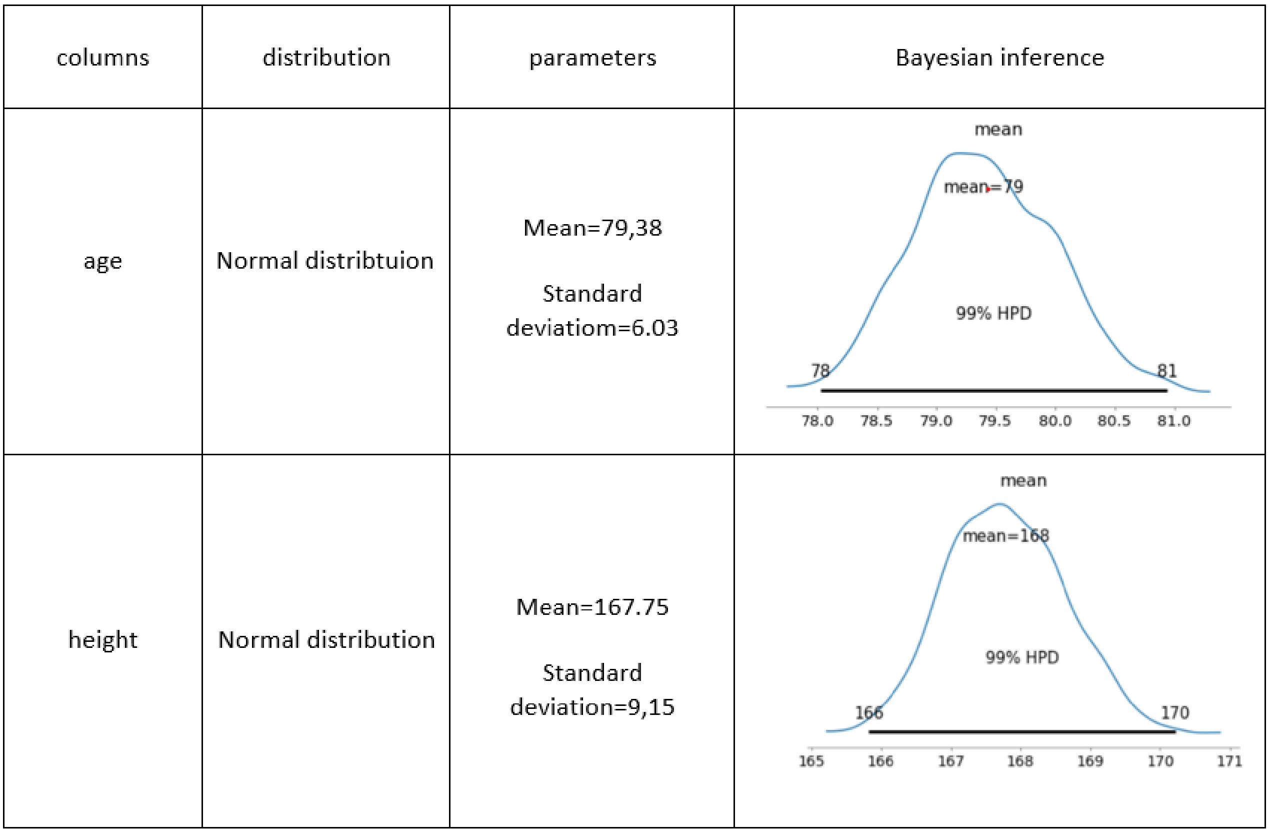


Figure 2.6: Distribution of the alpha parameter

We obtain that the alpha parameter follows a normal distribution with a mean of 19 and that the value given by our fit function (18,39) is included in this distribution. In what follows we will present a summary table where we apply the analysis approach presented above to all the continuous variables concerning the patient.



columns	distribution	parameters	Bayesian inference
age	Normal distribtuion	Mean=79,38 Standard deviatiom=6.03	
height	Normal distribution	Mean=167.75 Standard deviation=9,15	

mean
mean=79
99% HPD
78
81
78.0
78.5
79.0
79.5
80.0
80.5
81.0
mean
mean=168
99% HPD
166
170
165
166
167
168
169
170
171

Figure 2.7: Bayesian inference on the other continuous variables

2.1.2 Categorical variables

To visualize clearly the different categories of the variables drugs and diagnostic we need first to convert those columns into dummies variables. We are then going to plot the count of the patients taking the different drugs and diagnostics in our data set.

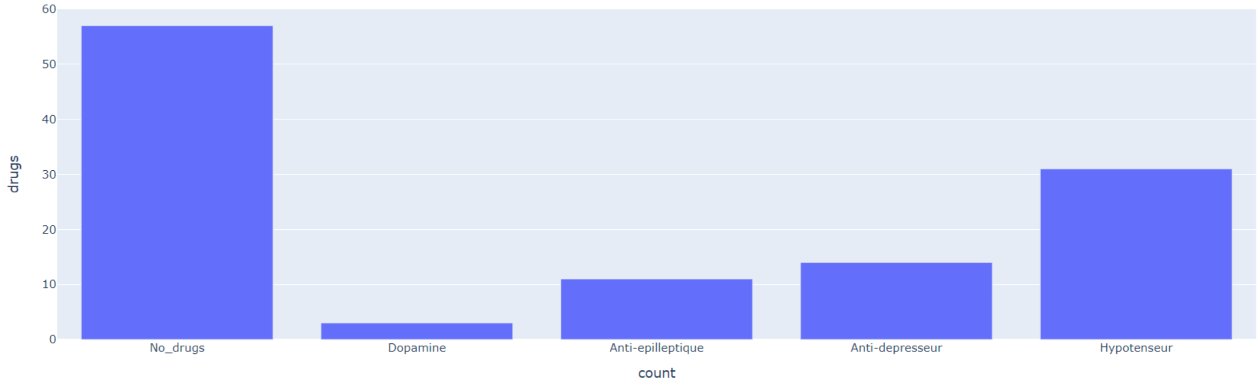


Figure 2.8: Drugs taken by patients

We observe that the majority of patients (57 out of 100) do not take any particular drug. for the remaining 42, they take mostly hypotensive drugs followed by anti-depressants, and a small part of the patients are affected by anti-epileptics and dopamine. Knowing that given our transformation of the drugs column into dummies variables, the same patient may take several drugs simultaneously.

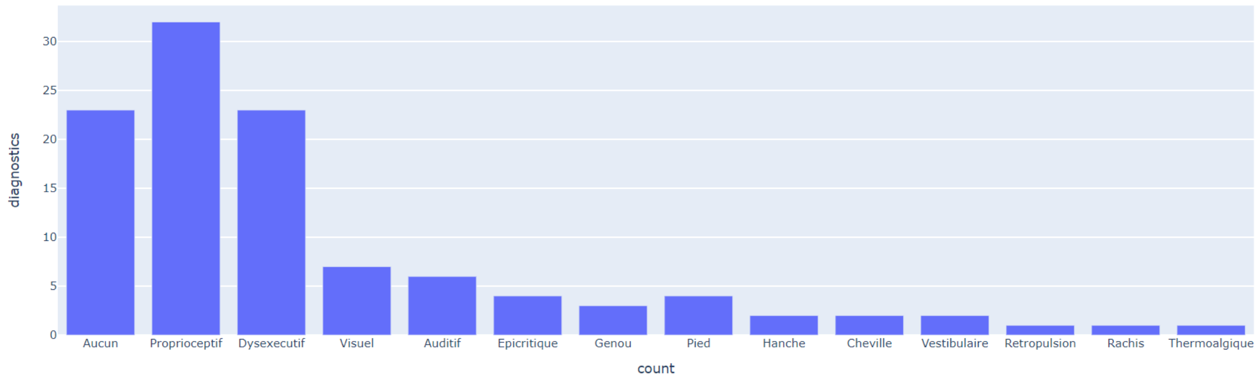


Figure 2.9: Diagnostics of patients

As for the diagnosis the great majority is divided between proprioceptive, dysexecutive, and no diagnostic. As for the rest of the diagnostics, it concerns only a small part of the study population. Once again given our transformation of the diagnostic column into dummies variables, the same patient may have several diagnostics.

2.1.3 Target variable: total fall count

Total fall count is the target variable of our data set and represents the number of falls of the patient.

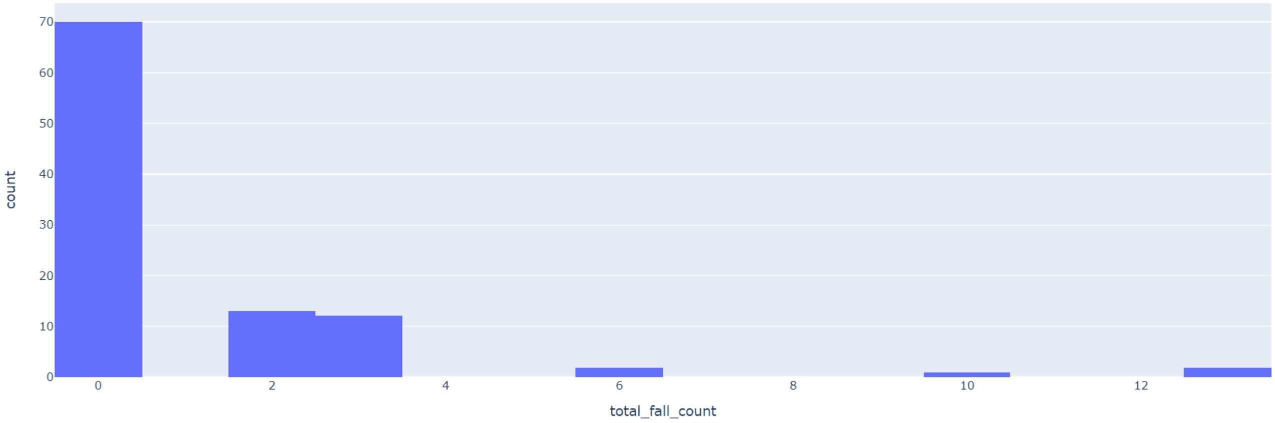


Figure 2.10: total falls of patients

We can see that among the 100 patients studied, 70 did not experience any fall, and the remaining 30 patients experienced from 2 to 13 falls. By reading the literature we found that given a threshold of 2 falls we can consider that an old person falls frequently. Therefore to better target our study and to set up a model that makes more sense we will modify our target variable into a binary variable that we will call is falling and which will take the value of 0 if the elderly person has not experienced any fall and 1 if she had more than 2 falls

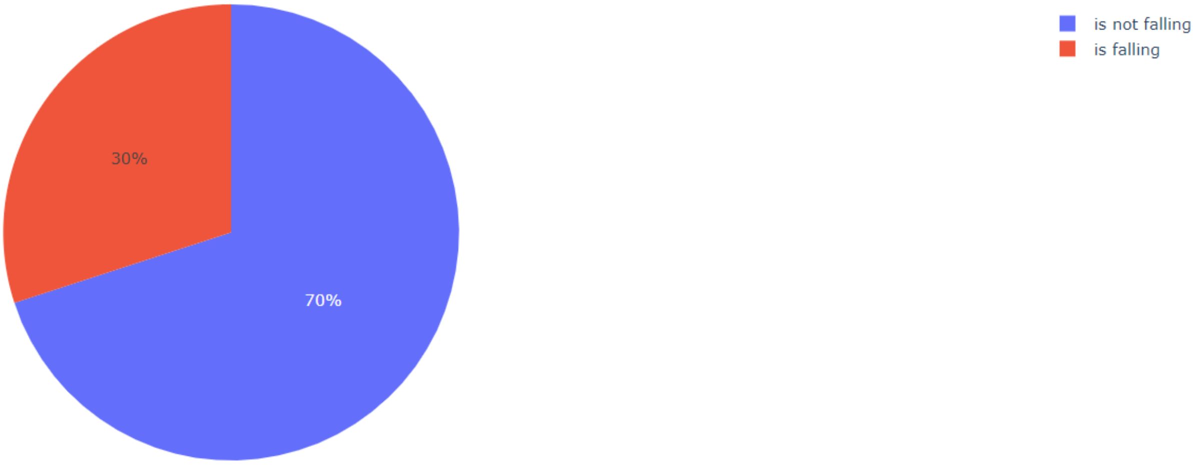


Figure 2.11: Falling rate in elderly

This graph allows us to assume that our target variable is a Bernoulli law with p = 0.3.

However, as in the previous studies, this value is not really deterministic and therefore we decide in what follows to carry out a Bayesian inference on the parameter p of our Bernoulli to determine the range of possible value that this parameter can take. Like the previous case, we take our prior uniform distribution.

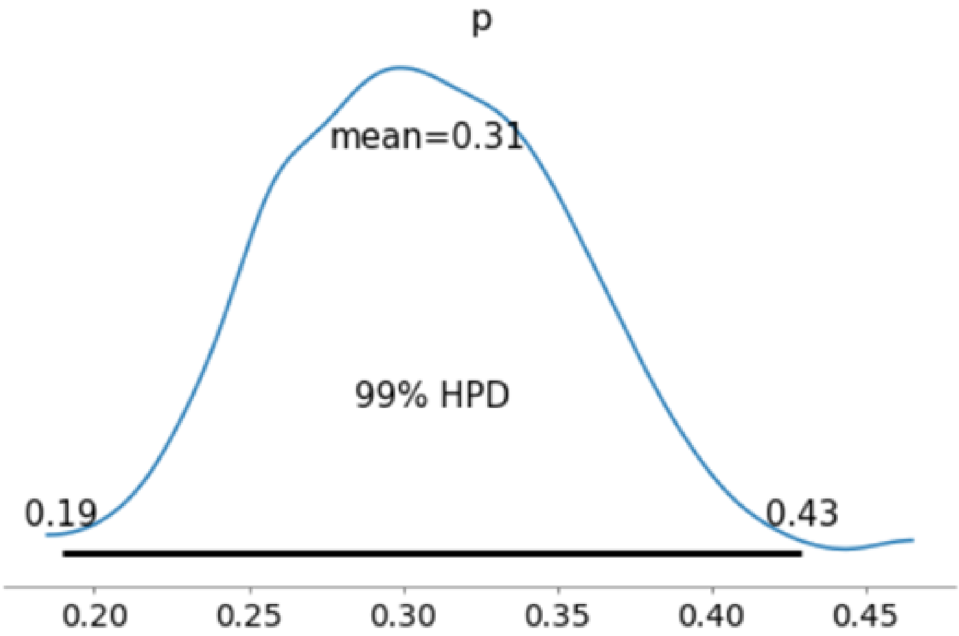


Figure 2.12: Distribution of the p parameter

This analysis allows us to say that the value of the p parameter of our Bernoulli follows a Gaussian of mean 0.31.

2.2 Stabilogram data

Considering the multitude of the variable received by the stabilogram and that these last are all numerical values coming from sensors. We have judged that it is not interesting to study these variables in-depth or to approach them by particular distributions. So in what follows of our study we are going to focus more on the impact of these variables on the prediction of our target variable and not on the signification of each of them.

Chapter 3

Individual effect of predictors on target variable

3.1 Correlation of variables

Because our data set contains several variables, it would not be practical to display their correlation in one go. Therefore, we are going to display the correlation of each cluster defined previously separately then a fifth cluster containing all the information concerning the patient: weight, height, gender, age, drugs, diagnostics. Regarding the last two variables mentioned since they are categorical variables, we have transformed them into binary variables:

$$\textbf{Takes drugs}\begin{cases} 0, & \text{if the patient doesn't take any drug.} \\ 1, & \text{if the patient takes one or more drugs.} \end{cases} \tag{3.1}$$

$$\textbf{Has symptoms}\begin{cases} 0, & \text{if the patient has no diagnostic.} \\ 1, & \text{if the patient has or more diagnostics.} \end{cases} \tag{3.2}$$

To judge that a variable is strongly correlated with our target variable we will also plot the average of the correlations of all the variables of our data set and we will consider in what follows that if the correlation of a variable exceeds in absolute value this average that would mean that the variable is strongly correlated with the target variable.

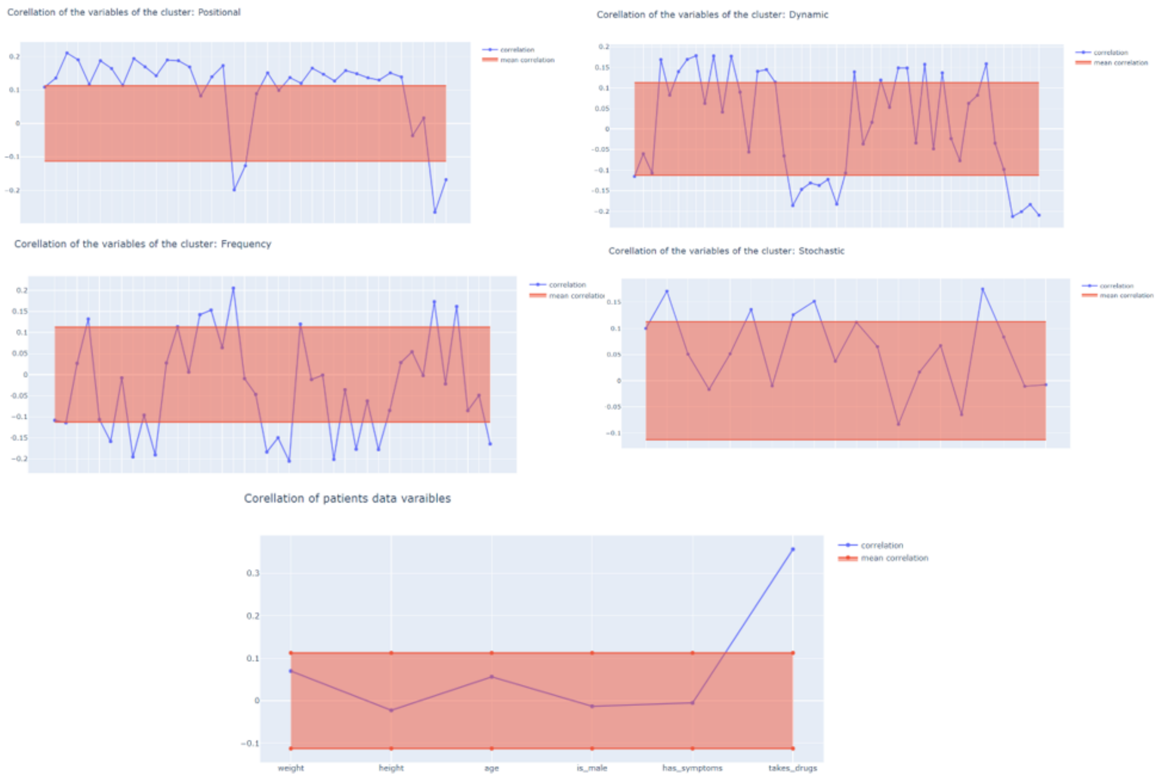


Figure 3.1: Correlation of the variables by clusters

Regarding the variables coming from the stabilogram, we see that the variables belonging to the positional cluster are distinguished from the other since most of these variables are closely related to the target variable. Unlike the stochastic cluster where the majority of the variables are almost not correlated with the target variable. Finally, regarding the variables concerning the patient, it can be seen that the variables hardly correlate except the take drugs variable which is very closely correlated (the strongest correlation, all variables combined). So, in what follows we will focus in more detail on the relation of categorical variables to our target variable.

3.2 Relation between categorical variables and the target variable

In this part, we will study the relationship between the categorical variable (gender, drugs, and diagnostic) and the target variable (is falling). To do this we will transform the categorical variable to have variable dummies to be able to study the impact of the different classes of these variables. Then we will perform a chi2 test which has the null hypothesis:

H0: there is no significant relationship between the two variables.

So, if we get a very small p-value (less than 0.5) we can reject the null hypothesis and therefore say that there is a direct relation between the two variables. But before starting the calculation of the p-value, it is necessary to verify that the different hypotheses of the test are verified. These are:

Strong constraint: each class must be represented (number of observations greater than 0)

Low constraint: each class must be sufficiently represented (number of observations greater than 5)

3.2.1 Gender

Table 3.1: classification of gender by target variable

is falling	is female	is male	ALL
0	34	36	70
1	15	15	30
ALL	49	51	100

We obtain a p-value =0,93 so we can clearly say that there is no relationship between gender and the falling of a patient.

3.2.2 Drugs

Table 3.2: classification of drugs by target variable

is falling	No drugs	Dopamine	Anti-epilleptique	Anti-depresseur	Hypotenseur	ALL
0	48	2	4	5	17	76
1	9	1	7	9	14	40
ALL	57	3	11	14	31	116

We obtain a p-value =0,0005 so we can say with confidence that there is a direct relationship between the medication taken by the patient and the fact that he is falling.

3.2.3 Diagnostic

We notice that when transforming this variable into dummies variable the hypothesis of the test aren't verified anymore (Some class have no representative). To remedy this, we are going to merge those classes so that the resulting ones always have a meaning and that the hypotheses are checked. In this situation we are going to merge the classes: "Epicritique","Genou","Pied","Hanche","Cheville","Vestibulaire","Retropulsion","Rachis","Thermoalgique" Into one class that we are calling: "Jambe".

Table 3.3: classification of diagnostics by target variable

is falling	No Diagnostic	Proprioceptif	Dysexecutif	Visuel	Auditif	Jambe	ALL
0	16	22	13	6	2	9	68
1	7	10	10	1	4	2	34
All	23	32	23	7	6	11	102

We obtain a p-value =0,26 so we can say that there is no relation between the diagnostic and the falling of a patient.

After studying the relationship between the categorical variables and our target variable, we are int what follow going to study the relationship with our continuous variable using **Anova test**.

3.3 Variance analysis

3.3.1 Overview

Our dataset contains also continuous variable that may have an effect on our target variable thus we need to perform a individual analysis on them to see if there is such an effect . In fact controlling by our target variable labels (faller and non faller) may divide our studied variables to two distribution with different characteristic or trends. To keep the analysis simple we assume that a difference between two distributions may introduce a difference in the means.

3.3.2 Examples

We use for example age feature to express what we have discussed in the overview .

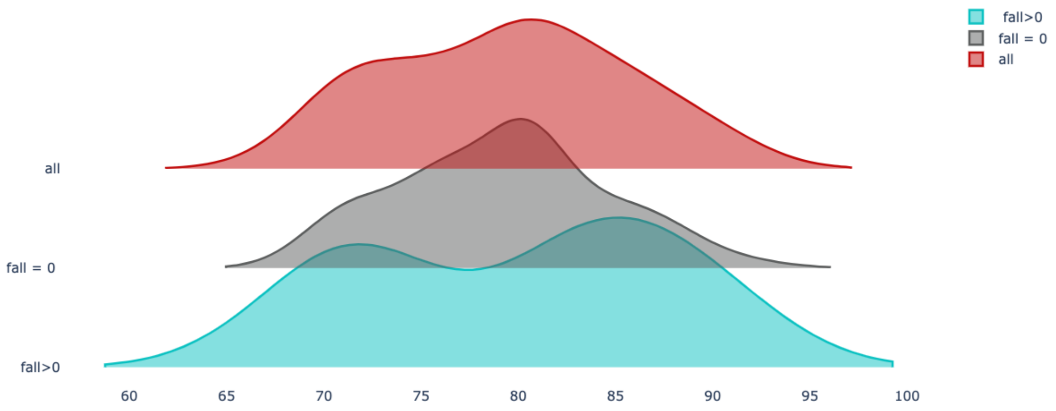


Figure 3.2: Age distribution controlled by target variable's two labels (fallers , non fallers)

As we can see controlling by our target variable, output two Age populations with slightly different mean . We move now to one of the stabilogram features , as we can see the mean values are remarkably different which may imply that the this sampled readings comes from two distributions

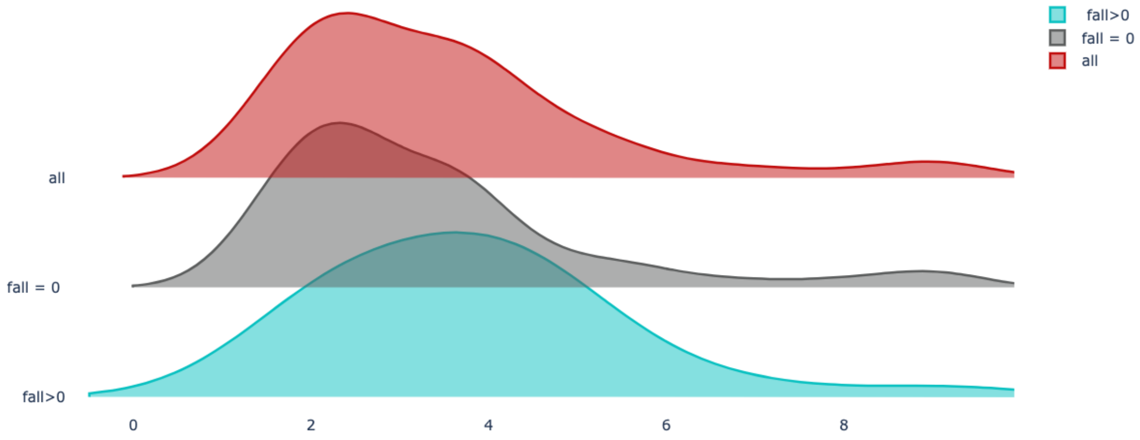


Figure 3.3: **Maximum value radius opened eyes** distribution controlled by target variable's two labels (fallers , non fallers)

3.3.3 ANOVA test application

Thus, we may need a test or metric to quantify this difference in mean when controlling with our target variable .

For this we use the ANOVA test that test if the means are different for our two population and gives us a p-value to see if we can reject our null hypothesis which is in this case the two population have the same mean.

Table 3.4: Anova statistic value and p-value for variables with p-value less than.05

Column name	**Test value**	**P-value**
FEATURE_length_over_area_ML_AND_AP_closed_eyes	7.444100	0.007544
FEATURE_mean_frequency_AP_closed_eyes	4.651233	0.033477
FEATURE_maximum_value_AP_opened_eyes	4.550980	0.035397
FEATURE_fractal_dimension_ce_ML_AND_AP_closed_...	4.508288	0.036250
FEATURE_frequency_mode_Power_Spectrum_Density_...	4.316302	0.040364
FEATURE_energy_content_2_inf_Power_Spectrum_De...	4.302826	0.040671
FEATURE_power_frequency_50_Power_Spectrum_Dens...	4.135991	0.044683
FEATURE_mean_frequency_ML_AND_AP_closed_eyes	4.135083	0.044706
FEATURE_length_over_area_ML_AND_AP_opened_eyes	4.032209	0.047392

As we can see not all variables have important difference in mean regarding the target variable. But for the ones(presented in the table) with low p-values they may help us infer the target of our observations or unseen ones.

After studying the relationship of all of the variables of our data set with our target variable we can start selecting the features that have the most impact. To do we are starting by measuring the predictive power of each of those variables.

3.4 Selecting features through their individual effect

3.4.1 Method

To find the most relevant features that explain balance disorder in elderlies, we will measure for each feature how much it can detect an elderly's fall.

To do so, firstly we will consider a feature and then build a logistic regression model that predicts the elderly's fall based on it. We will measure the accuracy of the model by considering a threshold of 0.3 as illustrated below, and we will preserve this convention throughout all the later sections:

$$\text{elderly is considered} \begin{cases} \text{faller,} & \text{if probability} \geq 0.3 \\ \text{not faller,} & \text{otherwise} \end{cases} \tag{3.3}$$

The choice of the 0.3 as a threshold is justified by the fact that we are dealing with unbalanced data where 0.3 of the population are faller and the remaining one are not. Secondly, we will attribute to the considered feature a score which is the accuracy of the corresponding model. Finally, we will sort all features by their score to obtain a table of features ranked by their importance in explaining elderly falls.

3.4.2 Result

After applying the previous method, we obtain the graph below. For readability, we restricted ourselves to showing the 3^{rd} quartile of features.

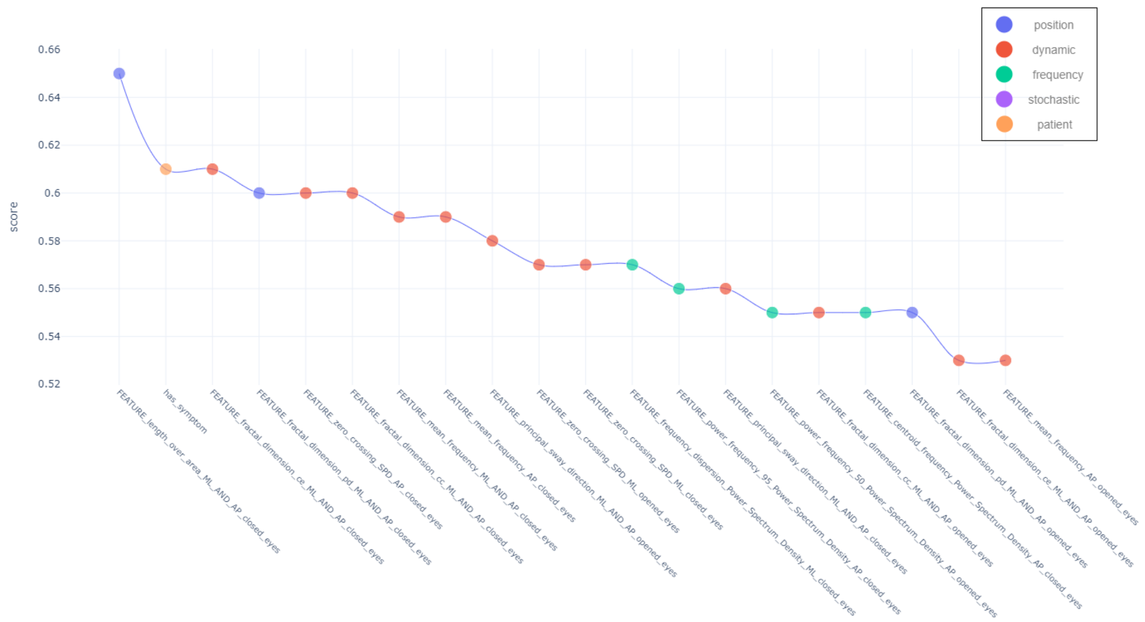


Figure 3.4: Plot that shows the 3^{rd} quartile of features ranked by their corresponding logistic regression accuracy score

We notice that the first feature "length over area ML AND AP closed eyes" – which belongs to the position cluster – stand out significantly from the other features. We also observe that all clusters are present in the first 10 first features except for the stochastic cluster. On the other hand, we remark that dynamic cluster dominates the other one in the sense that appears successively and several time ($\frac{12}{22}$ time).

3.4.3 Discussion

If we are interested in selecting only one variable to detect/explain elderlies' fall, we would definitely choose "length over area ML AND AP closed eyes" which as the results show has a high score compared to the other features.

Regarding the followed approach, we think that it has a major drawback. If within a cluster we have several features that are highly correlated or multicollinear with each other, they would automatically have relatively a similar score and thereby appear in a similar rank. This suggestion might explain both the absence of the stochastic cluster in the 10 first features and dominance of the dynamic cluster since the first places would be taken by a feature and its multicollinear peers.

Another major drawback of this approach is not considering the combined effect of features. In fact, we may have two features that when taken independently do not have a great impact in explaining elderlies fall. Yet, their combined effect is significant in that.

3.5 Conclusion

By regrouping all the previous results we see that a certain number of variables were able to pass the Anova test and are part of the third quartile of the features allowing the most accuracy and also have an important correlation with the target variables. These variables are:

Table 3.5: Best features given by the univariate analysis

Features	Cluster
Length over area ML AND AP closed eyes	Position
Mean frequency AP closed eyes	Dynamic
Fractal dimension ce ML AND AP closed eyes	Dynamic
Frequency mode Power Spectrum Density AP closed eyes	Frequency
Power frequency 50 Power Spectrum Density ML opened eyes	Frequency
Mean peak Sway Density closed eyes	Dynamic
Length over area ML AND AP opened eyes	Position

However, this does not allow us to conclude with regard to the other variables and as explained previously it does not give us any idea on the impact of the combination of these variables on the prediction of our target variable hence the importance of the following chapter which we will help determine the linear relationship between the different variables of our data set.

Chapter 4

Effect of predictors Linear combination on target

4.1 Overview

Within this approach, we will tackle the previously mentioned issues. More precisely, we will not manipulate features independently, rather we will try to find the best linear combination of them that better explain elderlies' fall. To do so, we primally need to filter features in each cluster so that to remove all multicollinear ones. We will use the varaince inflation factor (VIF) to measure multicollinearity and decide on a threshold to keep the relevant features in each cluster. Later, we will use forward stepwise procedure to obtain the best linear combination of features.

4.2 Preselection of features using VIF

4.2.1 Method

Given a set of features $X_1, X_2, ..., X_n$, we will use the Variance Inflation Factor (VIF) to measure the multicollinearity of a given features X_i and the other remaining ones. The key idea behind VIF is to build a linear regression model that predicts X_i using the other features and compute its R^2. The more the R^2 approach 1 the more it means that X_i is a linear combination of the other features, in other words $X_i \approx \sum_{j \neq i} \alpha_i X_j$ where the $(\alpha_j)_j$ are the linear coefficients.

On the other hand, we obtain the inverse effect when R^2 approaches 0. Based on that, the VIF score is defined as

$$VIF\left(X_i\right) = \frac{1}{1 - R^2}$$

A high VIF score means high multicollinearity between the X_i and other features, and vice versa.

Now we need to decide on a VIF threshold above which we will consider X_i a linear combination of the other variables and consequently remove it. We will assume that X_i is literally a linear combination of the other variables if the R^2 of the obtained linear regression model exceeds 0.99. following this assumption, we will remove features whose VIF score exceeds 100.

$$R^2 \geq 0.99 \Leftrightarrow \frac{1}{1 - R^2} \geq \frac{1}{0.01} \Leftrightarrow VIF \geq 100$$

Note that we will not preselect features from patient data (age, height, weight, take drugs, ...) since it contains binary features that are not suited for linear regression model.

4.2.2 Result

For each cluster namely position, dynamic, frequency, and stochastic, we compute the VIF score of its features and keep the features whose VIF are below 100. The final set of features has a total of 51 features. In the following graphs, we represent the percentage of features that verify the criteria $VIF \ \leq 100$ within each cluster and their partition in the final set:

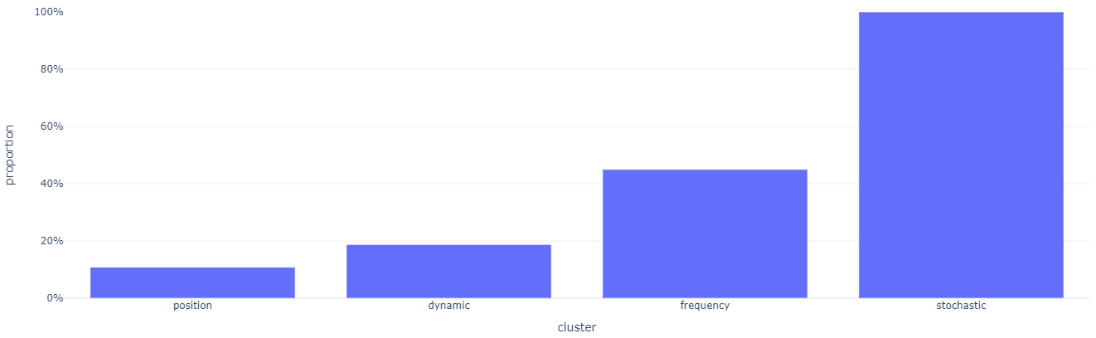


Figure 4.1: Bar chart that represents the retained proportion from each cluster after preselection using VIF

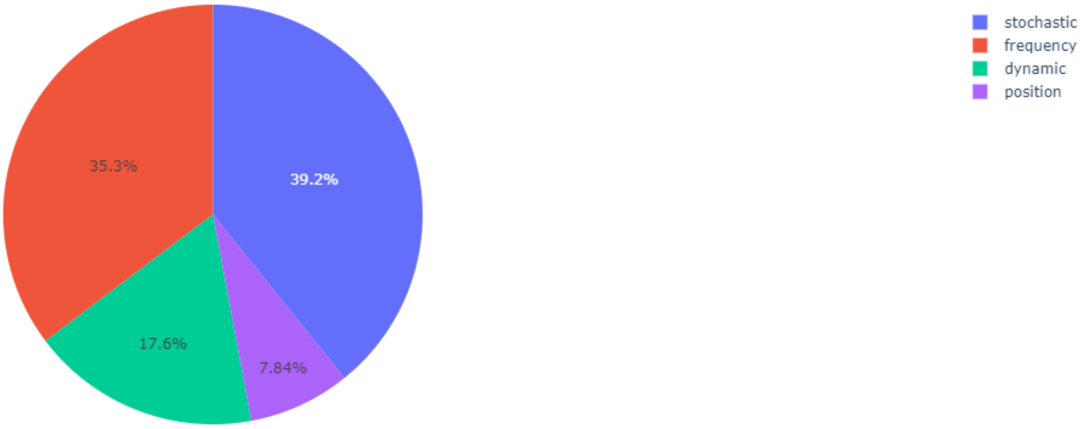


Figure 4.2: Pie chart that represents the proportion of each cluster in the final set of preselected features

We observe that only 11% and 20% of the features in the position and dynamic cluster respectively were kept whereas all and almost half of the features in the stochastic and frequency cluster respectively were preserved. This means that there is high multicollinearity within the position and dynamic cluster compared to a low one within the stochastic and frequency. On the other hand, we notice that the final set consists meanly of stochastic and frequency features (almost 75%) while the remaining 25% are dynamic and position features.

4.3 Best linear combination using forward stepwise

4.3.1 Method

In our case, we will start with an empty set S that will be holding the best linear combination of features. In each stage and within the set of preselected features, we will look for the features V_{i_k} that maximize the accuracy of the logistic regression model built using features $S \cup \{V_{i_k}\{$ and then append it to S. We will repeat this procedure until there is no remaining feature in the set of preselected features.

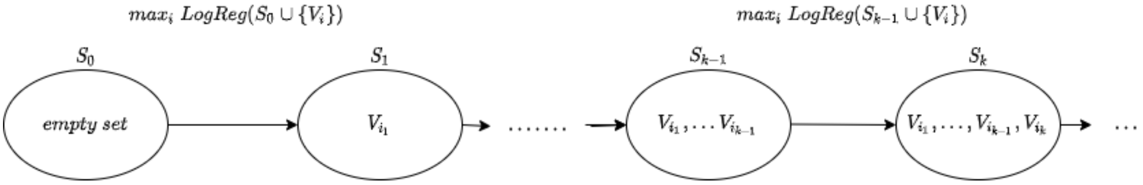

$max_i\ LogReg(S_0 \cup \{V_i\})$
$max_i\ LogReg(S_{k-1} \cup \{V_i\})$
S_0
S_1
S_{k-1}
S_k
empty set
V_{i_1}
....
$V_{i_1}, \dots V_{i_{k-1}}$
$V_{i_1}, \dots, V_{i_{k-1}}, V_{i_k}$
...

Figure 4.3: A schema that illustrates the forward stepwise procedure

As a result, we will label features by a number that represents the iteration in which the feature was added, and a score that represents the accuracy of logistics regression built using the feature and previously added ones.

4.3.2 Result

After applying the previously mentioned procedure, we make a graph where in the x-axis, we rank features according to the order in which they were added to the set of best linear combinations S. On the y-axis, we show the accuracy of logistic regression build using the current variables and all previously added ones, in other terms, the score of logistic regression built using S in the given stage. In another graph, we also represent the proportion of each cluster in the final set of best linear combination of features.

For the sake of readability, we show only the 12 first features in the first graph. You can refer to the Appendix section to view the complete result.

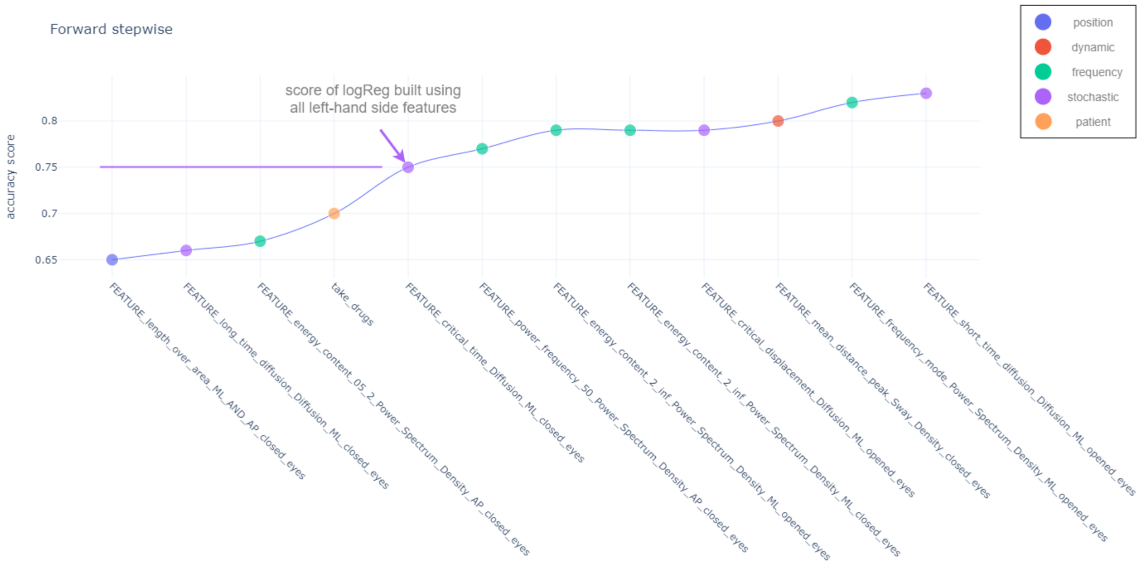


Figure 4.4: Plot that shows the evolution of logistic regression accuracy when appending features

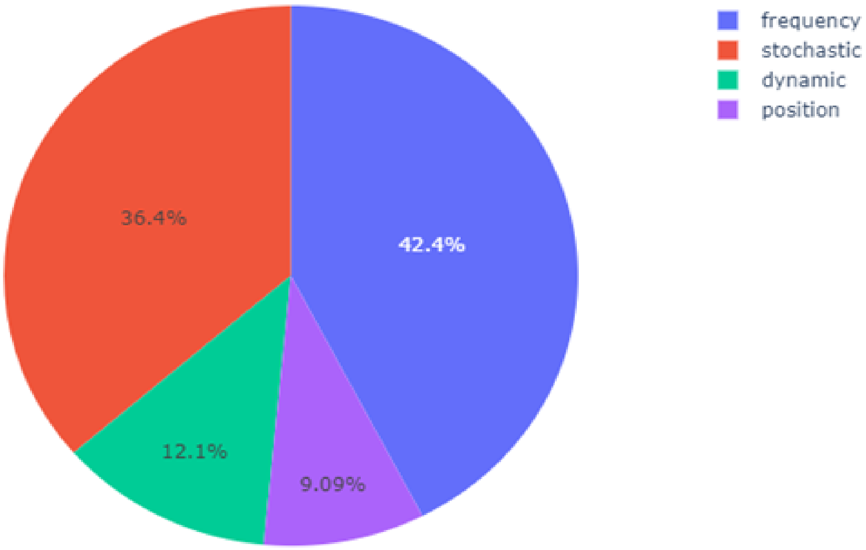


Figure 4.5: Pie plot that shows the proportion of each cluster in the final set of best linear combination of features

We observe that the first 4 features belong to 4 different clusters: position, stochastic, frequency, and patient. Besides, using only 12 features we were able to reach a 0.83 accuracy of the model. If we extend this to 37 features – which is in our case the max number of features that can be included in logistic regression – we can reach an accuracy of 0.89.

On the other hand, we see a clear dominance of the frequency and stochastic clusters over the dynamic and position in the sense that the former represents 88.8% (42.2%, 36.4%) of the whole final set whereas the latter represents only 22.2%.

4.4 Conclusion

As a recap of this 2^{nd} approach, we have determined the best linear combination of variables that better explain elderlies' fall. Particularly, we preselected our features that were initially 148 using VIF so that to remove all multicollinear features, which hold the same information, and therefore we retained 51 features. Thereafter, we applied a stepwise forward procedure to select the best linear combination.

Regarding the results, we noticed that the first 4 features belong to 4 different clusters which suggest a relation of complementarity between clusters, in other terms, each cluster has its own role in explaining elderly balance impairment

Besides, we perceive that after applying stepwise, the proportions of cluster in the final set remained relatively similar to the proportions obtained after VIF selection. This imply that our selection method is biased.

Finally, this approach enabled us to reconfirm that “length over area ML AND AP closed eyes” is the feature that one could rely on to detect elderlies’ fall. Indeed, this feature not only survived VIF preselection but also was the first feature to be selected and included in the set of the best linear combination when applying stepwise.

Nevertheless, one major objection that might be addressed to this approach is the fact that it is restricted to linear combination and does not take into consideration interactions (nonlinear relation) between features. In the next section, we will suggest a new approach that will tackle this issue.

Chapter 5

Effect of variable interaction on target variable

5.1 Overview

In the last section we identified the best linear combination of our dataset features that have the optimal predictive power . Although the method results are acceptably relative to the baseline . It does not exploit the interaction effect of variables in other word it only rely on the individual effect of each variable . In this section we use an entropy based method to detect pairs of variables of with high interaction effect in terms of predictive power, and eventually add them to our pair wise pool so they can compete with individual variables, given us the best combination between individual features and other synthetic variables that model interaction between a pair of features .

5.2 Variable interaction effect

5.2.1 Entropy framework

To talk about variables effect on our target. It's mandatory to introduce the framework in which we will quantify this effect. We use the entropy framework to quantify this effect. In fact, our target variable have an entropy value that characterize the uncertainty in it . Our goal using predictors is to reduce this entropy by conditioning on them . That's what we call **Information gain** .

Mathematically we can express the entropy as :

$$H(Y) = \sum_{i=1}^{n} -P(y_i)logP(y_i)$$

$$Y : target\ variable\ ,\ y_i : ith\ outcome\ of\ target\ Y$$

5.2.2 Individual information gain

As for the information gain we condition by a predictor variable X by splitting by one of its possible values a and we note it as $IG(X, a)$. This split gives us two populations where the entropy is less than the original population .

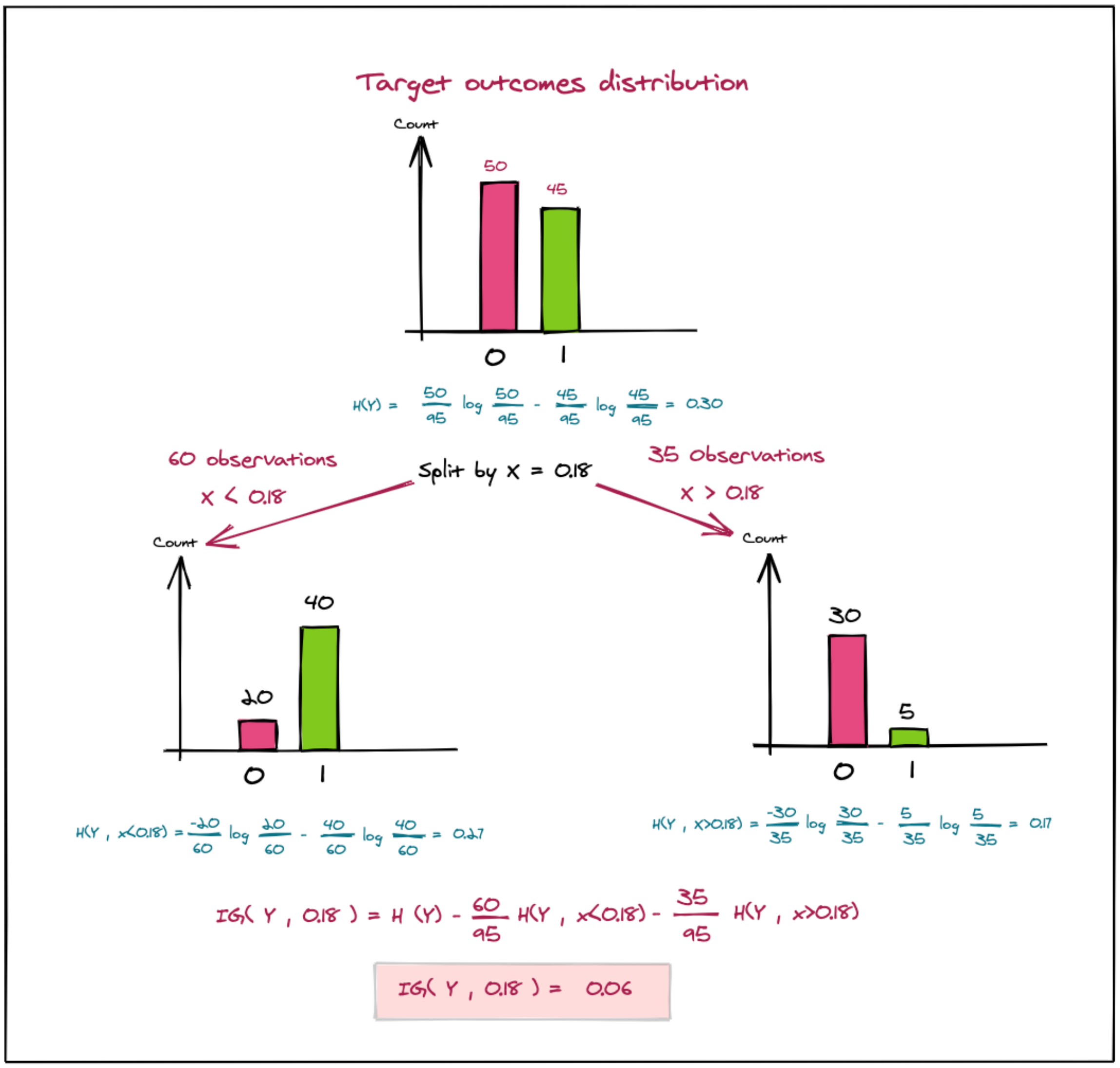

Target outcomes distribution
Count
50
45
0
1
H(Y) = 50/95 log 50/95 - 45/95 log 45/95 = 0.30
60 observations
x < 0.18
Split by x = 0.18
35 observations
x > 0.18
Count
40
20
0
1
Count
30
5
0
1
H(Y , x<0.18) = -20/60 log 20/60 - 40/60 log 40/60 = 0.27
H(Y , x>0.18) = -30/35 log 30/35 - 5/35 log 5/35 = 0.17
IG(Y , 0.18) = H (Y) - 60/95 H(Y , x<0.18) - 35/95 H(Y , x>0.18)
IG(Y , 0.18) = 0.06

Figure 5.1: this figure visualize how controlling by a variable can reduce the entropy of our target , this can be captured via information gain quantity

As you can see in the figure below, controlling by a predictor reduced our initial entropy. That is why it contains a information gain of 0.06. This operation can be generalized to compute the expected information gain of a predictor by averaging on its information gains caused by its possible split which gives us an estimation of information gain when we observe this predictor variable .

$$IG(Y, X) = E_{a \sim p(X)}(IG(Y, a))$$

5.2.3 Joint information gain

The definition of information gain can be expanded to a pair of variables by splitting by two values one for the first predictor and one for the second predictor which gives us **4 sub-population** and we note it as $IG(X, Y, a, b)$ which can be generalized as earlier to define the information gain after introducing two predictors which we note :

$$IG(Y, X_1, X_2) = E_{(a,b)\sim p(X_1,X_2)}(IG(Y, a, b))$$

5.2.4 Interaction information gain

All the quantities that we have defined are just tools to extract the interaction information gain. Consequently we define the interaction information gain as :

$$IG(Y, X_1 \cap X_2) = IG(Y, X_1, X_2) - IG(Y, X_1) - IG(Y, X_2)$$

This information gain could help us extract variables that have an interaction effect that could enhance our predictive power .

5.3 Results

After defining the tools that we need to quantify the interaction power we're ready to apply it to our use case using a simple workflow that compute the interaction information gain for each pair of variables in our data set .

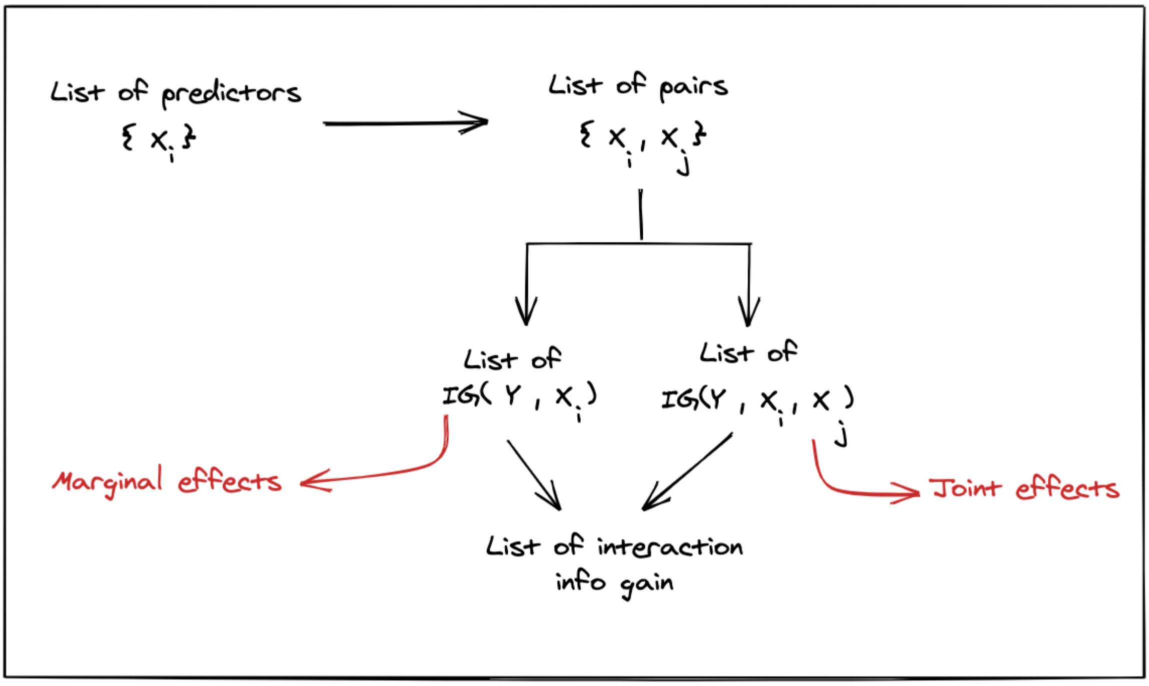


Figure 5.2: Main workflow to compute different information gain for each pair of variables

Running our script (see the code base) for each pair of predictors gives us this results (see fig. 5.3), where we plot distributions of marginal , joint and interaction information gain . So we can compare them and see if it is worthy to continue our study of interaction . In fact the interaction effect distribution contain two classes a set of pairs with 0 interaction and other with non-null effect . Plus comparing the interaction and marginal effect distribution in terms of magnitude(information gain values) help us state that the effect of interaction can not be discarded . In other words some pair of variables have an interaction gain that can be equivalent to the marginal effect of some variables which endorse the effect of variable interaction on predicting our target .

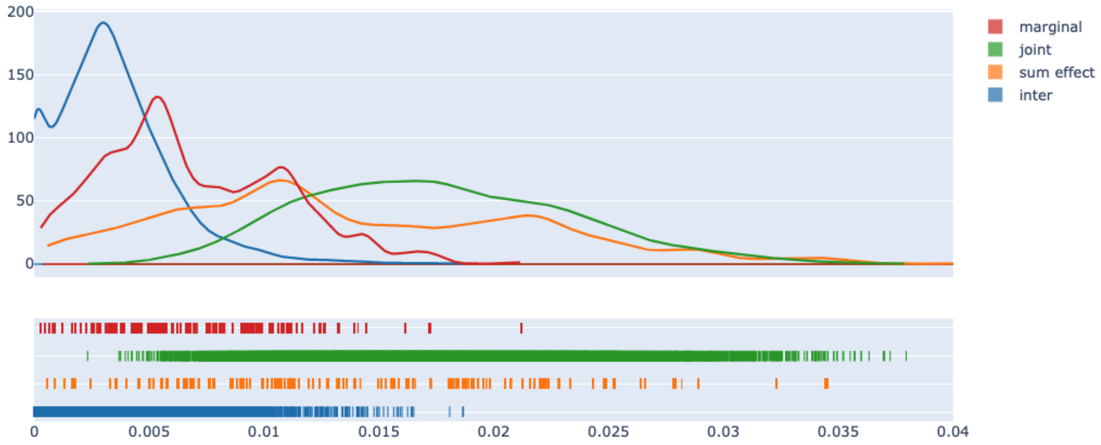


Figure 5.3: distribution of different type of information gain : individual , joint and interaction

5.3.1 Interaction effect test

Hypothesis formulation :

To quantify these observations we've computed the sum effect of each pair to compare it to the joint effect . In fact if the two distribution are the same this can be an argument that the interaction effect is minimal and all the reduction in entropy comes from the marginal effects . This can be our null hypothesis . This latter can be formulated as a **Kolmogorov** test that compare the two distributions .

Test result :

the test gives us a p-value of 0.002 thus we can reject the null hypothesis and state that the interaction effect can not be discarded .

5.3.2 Interaction graph

Now that we have our interaction terms for each pair of variable we can compute a graph of interaction that help us extract and visualize variables with high interaction and hence integrate them in our pair wise logistic regression pool .

Interaction graph definition

Before showing our interaction graph we first define it as indirected graph where nodes are our data set variables and with weighted edges that quantify the information gain between pairs of variables . This graph choice visualization will help us not only see the variable that interact but also how the predefined clusters interact between them to reduce entropy in our target variable . We also can filter level of information gain so we can limit our visualization to variable that interact the most .

Result

the graph below represent the top 5 (95 quantile) of our population which summarize the best interacting variables. At first glance we can see that the interaction is unbiased by variable class which means that the interaction between variables is independent on the cluster . In other words the cluster type does not define a community structure in our graph .

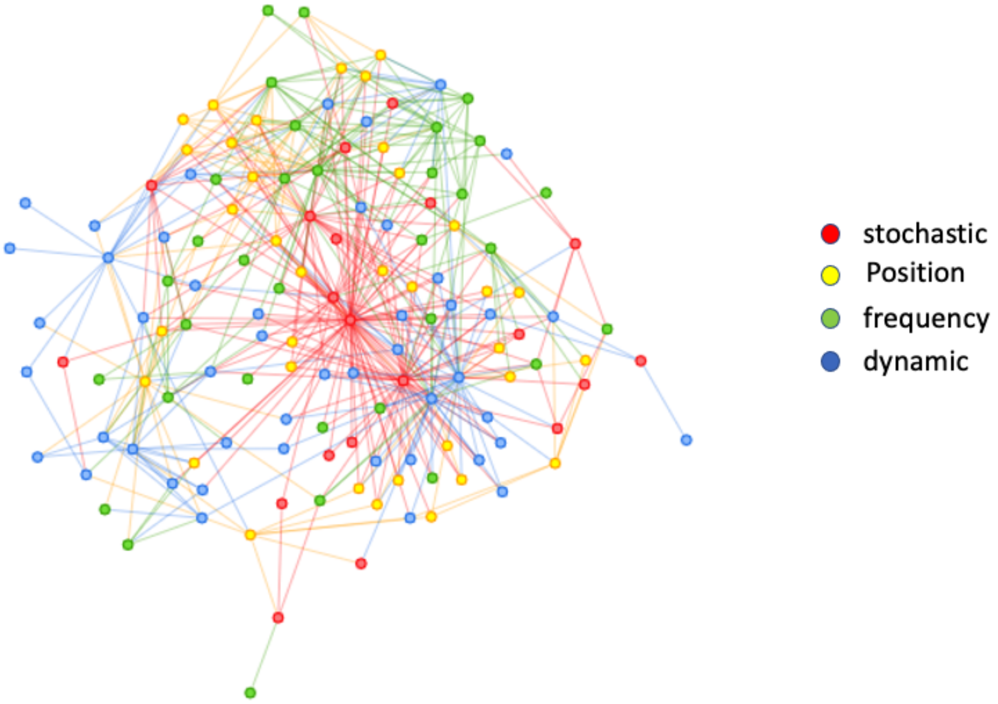
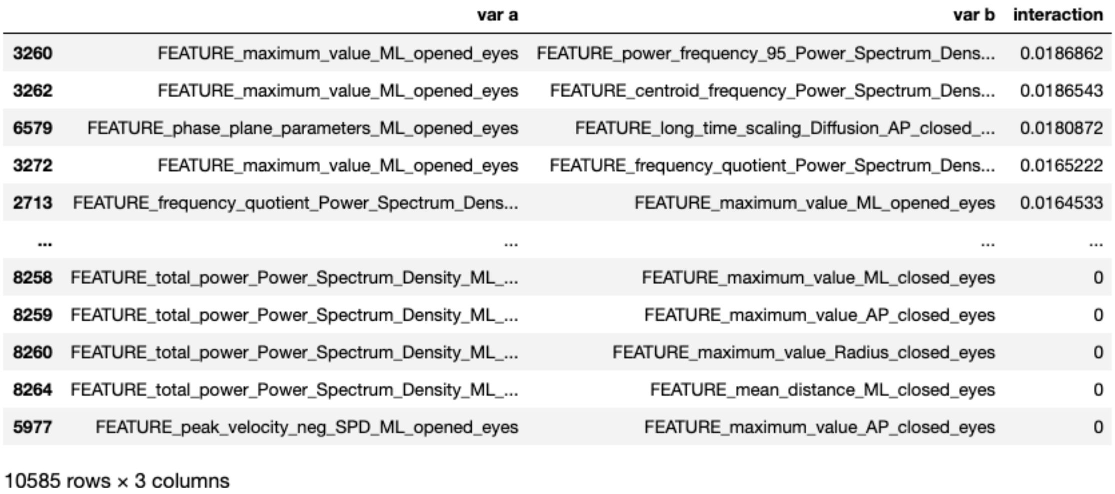


Figure 5.4: Interaction graph of top 5 % of variable pairs (last 5 % percentile)

We return now to our main goal which is to extract variables with high interaction information gain the table in fig. 5.5 gives us interacting pairs sorted by their interaction power



	var a	var b	interaction
3260	FEATURE_maximum_value_ML_opened_eyes	FEATURE_power_frequency_95_Power_Spectrum_Dens...	0.0186862
3262	FEATURE_maximum_value_ML_opened_eyes	FEATURE_centroid_frequency_Power_Spectrum_Dens...	0.0186543
6579	FEATURE_phase_plane_parameters_ML_opened_eyes	FEATURE_long_time_scaling_Diffusion_AP_closed_...	0.0180872
3272	FEATURE_maximum_value_ML_opened_eyes	FEATURE_frequency_quotient_Power_Spectrum_Dens...	0.0165222
2713	FEATURE_frequency_quotient_Power_Spectrum_Dens...	FEATURE_maximum_value_ML_opened_eyes	0.0164533
...	...	...	...
8258	FEATURE_total_power_Power_Spectrum_Density_ML_...	FEATURE_maximum_value_ML_closed_eyes	0
8259	FEATURE_total_power_Power_Spectrum_Density_ML_...	FEATURE_maximum_value_AP_closed_eyes	0
8260	FEATURE_total_power_Power_Spectrum_Density_ML_...	FEATURE_maximum_value_Radius_closed_eyes	0
8264	FEATURE_total_power_Power_Spectrum_Density_ML_...	FEATURE_mean_distance_ML_closed_eyes	0
5977	FEATURE_peak_velocity_neg_SPD_ML_opened_eyes	FEATURE_maximum_value_AP_closed_eyes	0

10585 rows × 3 columns

Figure 5.5: this table present sorted interaction term for each pair by interaction power

In order to add them to our logistic pool we need to create new variables to quantify the interaction and keep the linearity assumption. To do so we choose to create artificial variables that take the product of a pair and eventually stand as a interaction term .

$$\textbf{pair \ \ of \ \ variables} \ : (X_i, X_j) \rightarrow \textbf{interaction \ \ variable} : X_{ij} = X_i * X_j$$

5.3.3 Step wise feature selection – interaction included

Using the same workflow presented in section 2 (linear combination) we apply a step wise feature selection to our new pool of variables that include in addition to our data set features a set of interaction variables.

Result

The plot in figure 5.6 present how the accuracy change by adding new best variables we can see that some of our interaction variables were added in early iteration which show their predictive power. Furthermore ,comparing to our previous model that only exploit linear combination of feature we can see an augmentation in our accuracy from 0.89 to 0.93.

Finally, and the most **interesting observation** is that some of the variables that were included in the ending iterations of our linear model or even omitted by the model are now included via their interaction term with other variables as amplitude which was included in the 5th iteration, although it wasn't even chosen in the linear model. This observation explains the power of interaction and give the model more expensiveness thus higher accuracy .

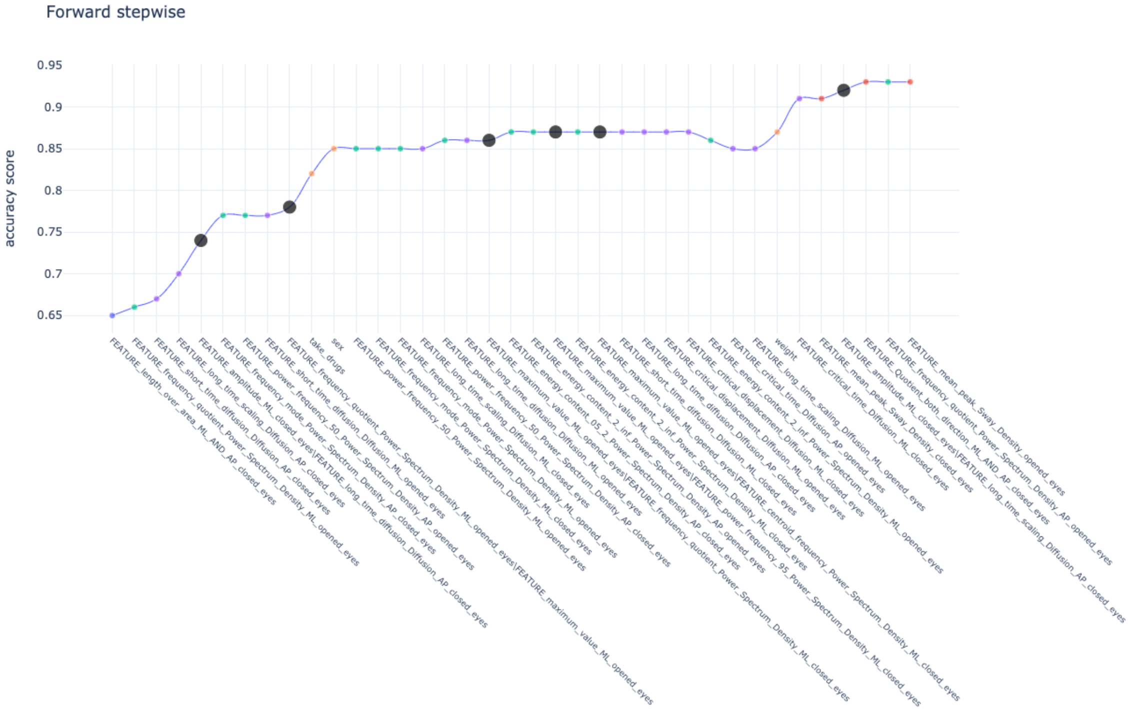


Figure 5.6: Forward step wise feature selection pulling from individual and interaction variables(**Black dots**)

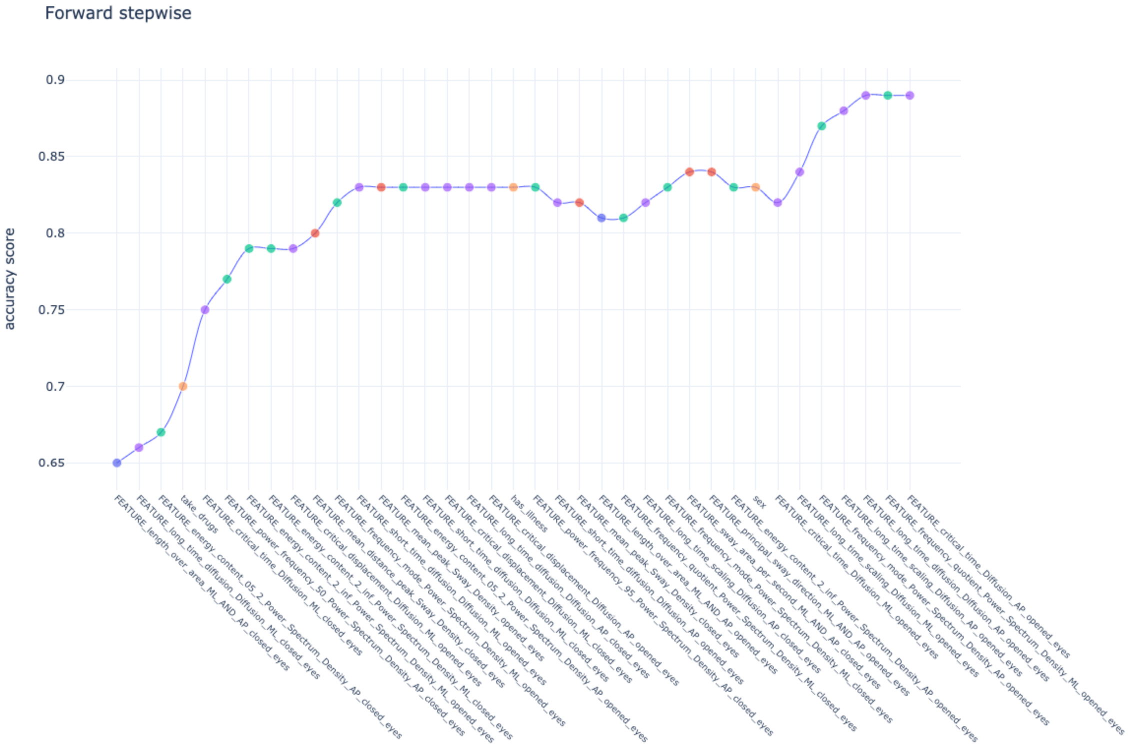


Figure 5.7: Forward step wise feature selection pulling from individual variables

5.4 Summary

Taking in consideration interaction variables improved our model accuracy and boosted its expressivity since he takes more than simple linear relations between variables and include non linearity that may link predictors to our target .

To conclude this section we summarize our key findings in this table ,as for the set of selected features you may find their specif names in the appendix

Table 5.1: Summary table of model accuracy after including interaction effect (P:position ,F:frequency , S:stochastic , D:dynamic , I :interaction)

Accuracy	**Accuracy improvement**	**Cluster ratios**
0.93%	**4.4%**	1(P),17(F),12(S),2(D),7(I)

Chapter 6

Discussion

6.1 Results benchmark

In the last chapters we tried a stack of methods to boost our model accuracy starting from the individual effect of our dataset predictors, then the collective effect via their linear combination. Finally we exploited the interaction effect of our predictors and integrated them as new artificial variables in our model in order to go beyond linearity and thus boost its accuracy .

The table bellow summarize our logistic regression accuracy improvement. As you can see the model contain the interaction variables gives the best result. Although we stay really close to simple linear combination. This may be explained by the lack of observation which make it harder to add more variable hence stay in the logistic regression assumption (p $\ll N$)

Table 6.1: Summary table of model accuracy in different scenarios

Scenario	**Accuracy**	**Accuracy improvement**
Individual effect%	**52% - 66%**	-
linear combination%	**89.6%**	59%
interaction included%	**93.0%**	4.4%

6.2 Cluster bias analysis

The level of accuracy of our model imply that we have successfully extracted features that may help us detect fallers from non fallers . But if we recall to our analysis in the last chapters we can say that our results depends heavily on the the existing of well defined clusters of variables thus we can exploit each one independently . For example calculating the VIF score of our variables was calculated for each cluster to reduce the number of variable thus stay under the assumption of linear regression since we have a small set of observation . This solution could help us continue our work but we may have got sub optimal result since we may discarded some variables that have predictive power but couldn't make it in the preselection via the VIF scoring.

The question thus is what would happen if we change our initial clusters. Can we still get the same result in other words the variables selected under the previous cluster will still stand out or there may be a variability in our results.

6.2.1 Clustering using PCA

Clusters are used to group variables that share same characteristics, we build one the same idea and use PCA algorithm which reduce the dimension of our data and build new dimensions that capture the maximum variance. This reduction will help us present our variables in a small dimension space and thus use of the shelf clustering methods as **K means** to compute new clusters .

We run the PCA algorithm and get the projection of our variables on the first two components that capture 71% of the variance . Then we extract new clusters using k-mena . The choice of 4 clusters and not 3 can be justified by the fact that with 3 clusters we will get a high number of variables in one of the clusters thus dividing it to two new clusters.

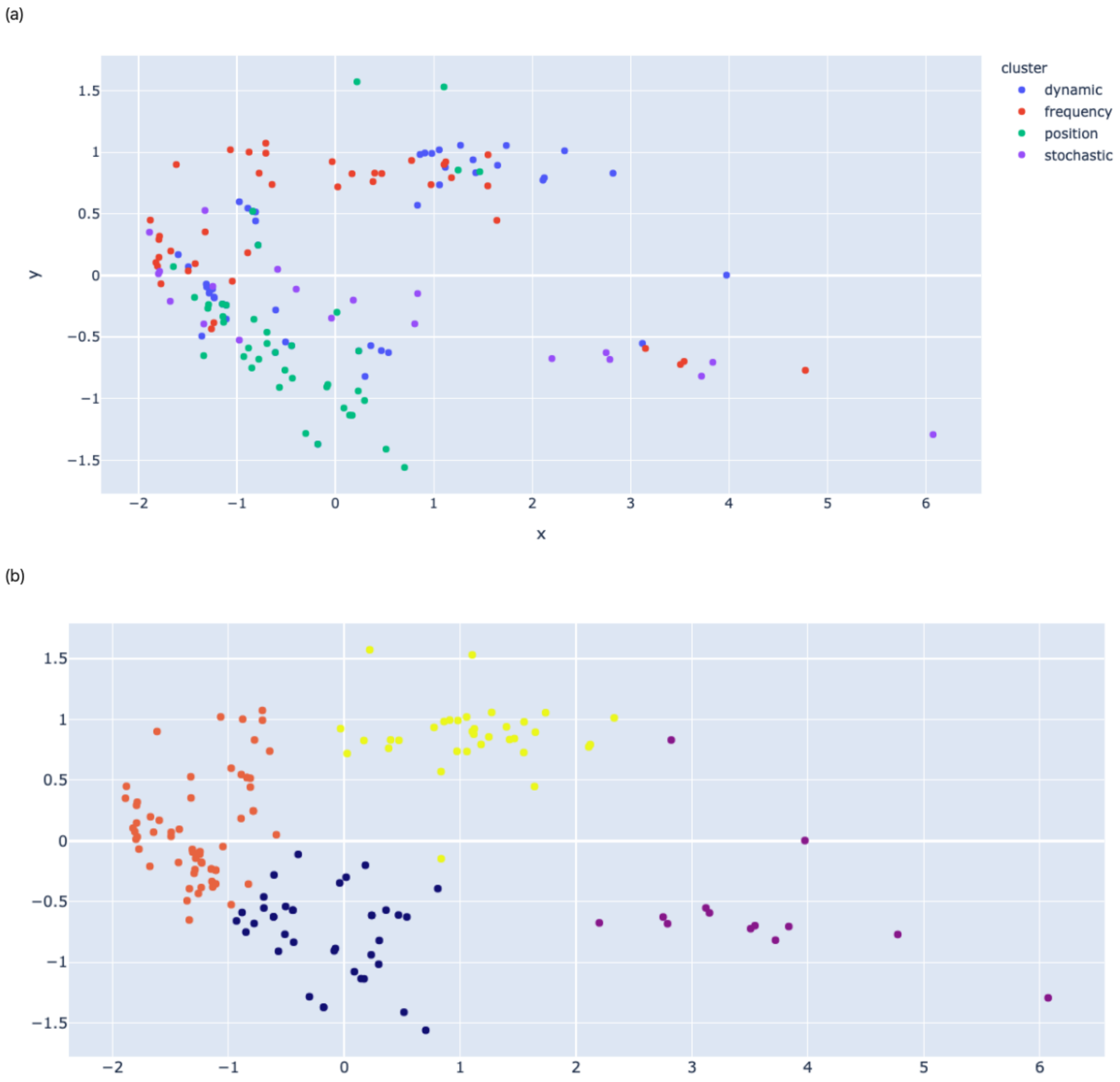


Figure 6.1: (a) Stabilo variables projection on the two first components ,(b) new clusters using K-means(k=4)

6.2.2 Results using new clusters

Presenting the results

After obtaining new clusters, we run the same workflow that we have done using old clusters. In other words, we apply the same methodology introduced previously except that now we rely on the new clusters obtained through PCA to preselect features.

As a recall of the methodology, we start by filtering out features and whereby removing multicollinear ones using VIF. Then following a stepwise forward procedure, we select the best linear combination of features. Finally, we introduce new features that incorporate interaction between features and reapplied stepwise.

We summarize our results in a table that gathers the accuracy of logistic regression and a pie chart that compares cluster proportions for both literature and PCA clusters. For further details, we also include a graph of the evolution of the score.

Table 6.2: Accuracy of logistics regression for the best linear combination in two cases: with-/without considering the interaction between features

Clusters	**Without interaction**	**With interaction**
Literature clusters	**89%**	**93%**
PCA clusters%	**88%**	**88%**

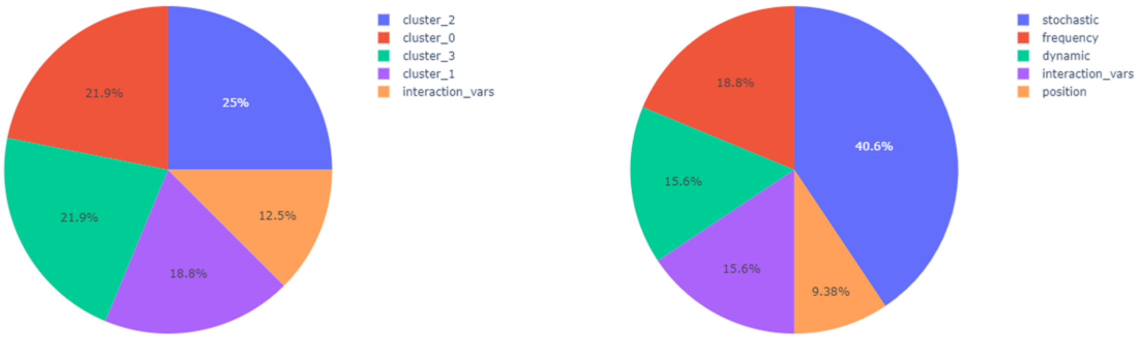


Figure 6.2: Pie chart that compares clusters proportion: literature clusters are in the right whereas PCA clusters are on the left.

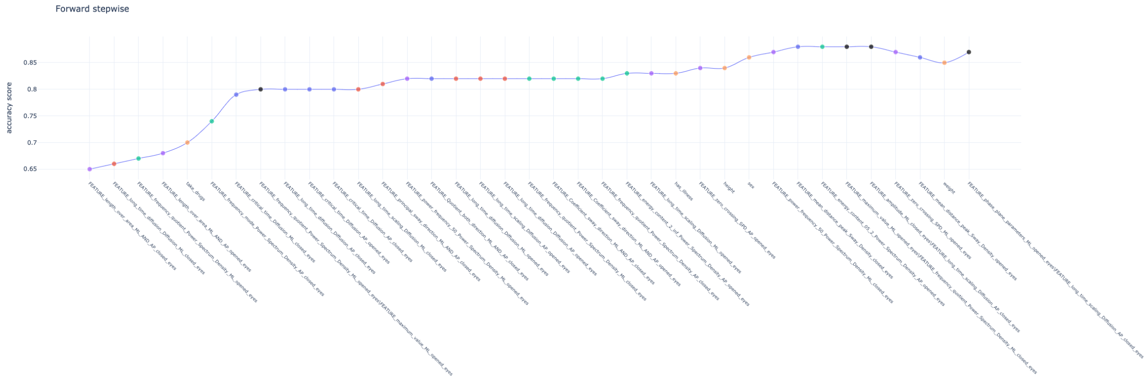


Figure 6.3: Forward step wise feature selection pulling from individual and interaction variables with PCA clusters

Describing and interpreting results

We notice that the accuracy obtained for PCA clusters is slightly less than the accuracy obtained for literature clusters (0.01%, 0.05%).

Regarding clusters proportion in the final set, we see almost an equipartition of PCA clusters (mean: 20% and std: 4%) as opposed to literature clusters where the stochastic cluster dominates over the others.

Moreover, the trend that we have discussed in the interaction effect section stays true. Indeed, the model always includes interaction variables in early iterations.

Seemingly, literature clusters appear to have the best results. Yet, the difference remains small to conclude that they are the ideal ones. When applying PCA, we have obtained clusters whose proprieties in terms of numbers of features, proportions in the final set are clearly different from the literature clusters. However, the final results are relatively the same which suggests that no matter how the clusters were chosen they yield the same conclusion. **Hence this reinforce that feature selection is unbiased by the choice of clusters and therefore we may preserve the original clustering due to their literature significance**

Chapter 7

Conclusion

To conclude we can say that during this study we tried to select the variables having the most impact on the prediction of our target variable and that through several procedures. We also studied the sensitivity of our results on cluster defined by the literature. However, the results found should be taken with great caution because the studies were carried out on a population of 100 patients and this number remains very low compared to the number of variables manipulated during our studies. This limits our ability to generalize the results found.

Chapter 8

Appendix

8.1 Table of features ranked by their logistic regression accuracy (3^{rd} quartile)

Feature	Cluster	Logistic regression score
length_over_area_ML_AND_AP_closed_eyes	position	0.65
has_symptom	patient	0.61
fractal_dimension_ce_ML_AND_AP_closed_eyes	dynamic	0.61
fractal_dimension_pd_ML_AND_AP_closed_eyes	position	0.60
zero_crossing_SPD_AP_closed_eyes	dynamic	0.60
fractal_dimension_cc_ML_AND_AP_closed_eyes	dynamic	0.60
mean_frequency_ML_AND_AP_closed_eyes	dynamic	0.59
mean_frequency_AP_closed_eyes	dynamic	0.59
principal_sway_direction_ML_AND_AP_opened_eyes	dynamic	0.58
zero_crossing_SPD_ML_opened_eyes	dynamic	0.57
zero_crossing_SPD_ML_closed_eyes	dynamic	0.57
frequency_dispersion_Power_Spectrum_Density_ML...	frequency	0.57
power_frequency_95_Power_Spectrum_Density_AP_c...	frequency	0.56
principal_sway_direction_ML_AND_AP_closed_eyes	dynamic	0.56
power_frequency_50_Power_Spectrum_Density_AP_o...	frequency	0.55

fractal_dimension_cc_ML_AND_AP_opened_eyes	dynamic	0.55
centroid_frequency_Power_Spectrum_Density_AP_c...	frequency	0.55
fractal_dimension_pd_ML_AND_AP_opened_eyes	position	0.55
fractal_dimension_ce_ML_AND_AP_opened_eyes	dynamic	0.53
mean_frequency_AP_opened_eyes	dynamic	0.53
power_frequency_50_Power_Spectrum_Density_AP_c...	frequency	0.52
power_frequency_95_Power_Spectrum_Density_AP_o...	frequency	0.52
mean_frequency_ML_closed_eyes	dynamic	0.52
frequency_mode_Power_Spectrum_Density_AP_close...	frequency	0.52
centroid_frequency_Power_Spectrum_Density_AP_o...	frequency	0.52
sex	patient	0.51
power_frequency_50_Power_Spectrum_Density_ML_o...	frequency	0.51
length_over_area_ML_AND_AP_opened_eyes	position	0.50
frequency_dispersion_Power_Spectrum_Density_AP...	frequency	0.49
centroid_frequency_Power_Spectrum_Density_ML_c...	frequency	0.49
frequency_quotient_Power_Spectrum_Density_AP_o...	frequency	0.49
centroid_frequency_Power_Spectrum_Density_ML_o...	frequency	0.49
mean_frequency_ML_AND_AP_opened_eyes	dynamic	0.49
mean_frequency_ML_opened_eyes	dynamic	0.48
frequency_dispersion_Power_Spectrum_Density_ML...	frequency	0.48
frequency_quotient_Power_Spectrum_Density_AP_c...	frequency	0.47
mean_peak_Sway_Density_closed_eyes	dynamic	0.47
zero_crossing_SPD_AP_opened_eyes	dynamic	0.47
frequency_mode_Power_Spectrum_Density_ML_close...	frequency	0.47
height	patient	0.47

8.2 Table of the best linear combination of features

Note: The table below represents the evolution (from top to bottom) of logistic regression accuracy when appending new features to the set of best linear combination of features.

Features	**Score**	**Cluster**
length_over_area_ML_AND_AP_closed_eyes	0.65	position
long_time_scaling_Diffusion_AP_closed_eyes	0.66	stochastic
critical_displacement_Diffusion_ML_closed_eyes	0.67	stochastic
long_time_diffusion_Diffusion_AP_closed_eyes	0.68	stochastic
frequency_mode_Power_Spectrum_Density_AP_close...	0.74	frequency
take_drugs	0.76	patient
frequency_quotient_Power_Spectrum_Density_AP_o...	0.78	frequency
mean_peak_Sway_Density_closed_eyes	0.80	dynamic
energy_content_2_inf_Power_Spectrum_Density_ML...	0.81	frequency
long_time_scaling_Diffusion_AP_opened_eyes	0.81	stochastic
long_time_diffusion_Diffusion_ML_closed_eyes	0.81	stochastic
power_frequency_50_Power_Spectrum_Density_AP_o...	0.81	frequency
long_time_diffusion_Diffusion_AP_opened_eyes	0.81	stochastic
energy_content_2_inf_Power_Spectrum_Density_AP...	0.81	frequency
frequency_mode_Power_Spectrum_Density_ML_opene...	0.81	frequency
power_frequency_50_Power_Spectrum_Density_ML_o...	0.81	frequency
age	0.81	patient
frequency_mode_Power_Spectrum_Density_AP_opene...	0.83	frequency
height	0.83	patient
confidence_ellipse_area_ML_AND_AP_closed_eyes	0.83	position
critical_time_Diffusion_ML_opened_eyes	0.84	stochastic
energy_content_05_2_Power_Spectrum_Density_AP_....	0.84	frequency
critical_displacement_Diffusion_AP_closed_eyes	0.85	stochastic
frequency_quotient_Power_Spectrum_Density_ML_o...	0.86	frequency
frequency_mode_Power_Spectrum_Density_ML_close...	0.87	frequency
Quotient_both_direction_ML_AND_AP_closed_eyes	0.88	dynamic
frequency_quotient_Power_Spectrum_Density_AP_c...	0.90	frequency
critical_time_Diffusion_AP_opened_eyes	0.91	stochastic

critical_time_Diffusion_AP_closed_eyes	0.91	stochastic
mean_peak_Sway_Density_opened_eyes	0.92	dynamic
long_time_scaling_Diffusion_ML_opened_eyes	0.91	stochastic
weight	0.92	patient
power_frequency_50_Power_Spectrum_Density_AP_c...	0.92	frequency
principal_sway_direction_ML_AND_AP_opened_eyes	0.91	dynamic
length_over_area_ML_AND_AP_opened_eyes	0.91	position
short_time_diffusion_Diffusion_ML_opened_eyes	0.91	stochastic
frequency_quotient_Power_Spectrum_Density_ML_c...	0.92	frequency

8.3 Table of the accuracy of the best linear combination of features (With interaction)

Note: The table below represents the evolution (from top to bottom) of logistic regression accuracy when appending new features to the set of best linear combination of features.

Features	**Accuracy**	**Cluster**
length_over_area_ML_AND_AP_closed_eyes	0.65	position
long_time_diffusion_Diffusion_ML_closed_eyes	0.66	stochastic
frequency_quotient_Power_Spectrum_Density_ML_o...	0.67	frequency
frequency_quotient_Power_Spectrum_Density_ML_o...	0.68	interaction_vars
maximum_value_ML_opened_eyes\centroid_frequenc...	0.74	interaction_vars
power_frequency_50_Power_Spectrum_Density_ML_o...	0.77	frequency
frequency_quotient_Power_Spectrum_Density_ML_c...	0.78	frequency
frequency_mode_Power_Spectrum_Density_ML_opene...	0.78	frequency
confidence_ellipse_area_ML_AND_AP_closed_eyes	0.78	position
mean_distance_peak_Sway_Density_closed_eyes	0.80	dynamic
short_time_diffusion_Diffusion_ML_closed_eyes	0.81	stochastic

principal_sway_direction_ML_AND_AP_opened_eyes	0.81	dynamic
short_time_diffusion_Diffusion_AP_opened_eyes	0.81	stochastic
energy_content_2_inf_Power_Spectrum_Density_ML...	0.80	frequency
length_over_area_ML_AND_AP_opened_eyes	0.80	position
amplitude_ML_closed_eyes\long_time_scaling_Dif...	0.79	interaction_vars
critical_displacement_Diffusion_AP_opened_eyes	0.79	stochastic
sex	0.79	patient
sway_area_per_second_ML_AND_AP_closed_eyes	0.79	dynamic
long_time_diffusion_Diffusion_AP_closed_eyes	0.79	stochastic
critical_time_Diffusion_ML_closed_eyes	0.78	stochastic
has_illness	0.78	patient
long_time_scaling_Diffusion_AP_closed_eyes	0.77	stochastic
power_frequency_95_Power_Spectrum_Density_AP_c...	0.80	frequency
energy_content_2_inf_Power_Spectrum_Density_AP...	0.81	frequency
energy_content_05_2_Power_Spectrum_Density_AP_...	0.81	frequency
long_time_diffusion_Diffusion_ML_opened_eyes	0.81	stochastic
power_frequency_50_Power_Spectrum_Density_AP_c...	0.80	frequency
frequency_mode_Power_Spectrum_Density_ML_close...	0.81	frequency
Quotient_both_direction_ML_AND_AP_closed_eyes	0.81	dynamic
power_frequency_50_Power_Spectrum_Density_AP_o...	0.80	frequency
frequency_quotient_Power_Spectrum_Density_AP_c...	0.81	frequency
take_drugs	0.86	patient
short_time_diffusion_Diffusion_ML_opened_eyes	0.88	stochastic
maximum_value_ML_opened_eyes\power_frequency_9...	0.88	interaction_vars
Coefficient_sway_direction_ML_AND_AP_opened_eyes	0.90	dynamic
energy_content_05_2_Power_Spectrum_Density_AP_...	0.93	frequency

8.4 Table of the accuracy of the best linear combination of features (with interaction for PCA clusters)

Note: The table below represents the evolution (from top to bottom) of logistic regression accuracy when appending new features to the set of best linear combination of features.

Features	Accuracy	cluster
length_over_area_ML_AND_AP_closed_eyes	0.65	cluster_3
long_time_diffusion_Diffusion_ML_closed_eyes	0.66	cluster_1
frequency_quotient_Power_Spectrum_Density_ML_o...	0.67	cluster_2
length_over_area_ML_AND_AP_opened_eyes	0.68	cluster_3
take_drugs	0.70	patient
frequency_mode_Power_Spectrum_Density_AP_close...	0.74	cluster_2
critical_time_Diffusion_ML_closed_eyes	0.79	cluster_0
frequency_quotient_Power_Spectrum_Density_ML_o...	0.80	interaction_vars
long_time_diffusion_Diffusion_AP_closed_eyes	0.80	cluster_0
critical_time_Diffusion_AP_opened_eyes	0.80	cluster_0
critical_time_Diffusion_AP_closed_eyes	0.80	cluster_0
long_time_scaling_Diffusion_ML_closed_eyes	0.80	cluster_1
principal_sway_direction_ML_AND_AP_closed_eyes	0.81	cluster_1
power_frequency_50_Power_Spectrum_Density_ML_o...	0.82	cluster_3
Quotient_both_direction_ML_AND_AP_closed_eyes	0.82	cluster_0
long_time_diffusion_Diffusion_ML_opened_eyes	0.82	cluster_1
long_time_scaling_Diffusion_AP_opened_eyes	0.82	cluster_1
long_time_diffusion_Diffusion_AP_opened_eyes	0.82	cluster_1
frequency_quotient_Power_Spectrum_Density_ML_c...	0.82	cluster_2
Coefficient_sway_direction_ML_AND_AP_closed_eyes	0.82	cluster_2
Coefficient_sway_direction_ML_AND_AP_opened_eyes	0.82	cluster_2
frequency_quotient_Power_Spectrum_Density_AP_c...	0.82	cluster_2

energy_content_2_inf_Power_Spectrum_Density_AP...	0.83	cluster_2
long_time_scaling_Diffusion_ML_opened_eyes	0.83	cluster_3
has_illness	0.83	patient
zero_crossing_SPD_AP_opened_eyes	0.84	cluster_3
height	0.84	patient
sex	0.86	patient
power_frequency_50_Power_Spectrum_Density_ML_c...	0.87	cluster_3
mean_distance_peak_Sway_Density_closed_eyes	0.88	cluster_0
energy_content_05_2_Power_Spectrum_Density_AP_...	0.88	cluster_2
maximum_value_ML_opened_eyes\frequency_quotien...	0.88	interaction_vars
amplitude_ML_closed_eyes\long_time_scaling_Dif...	0.88	interaction_vars
zero_crossing_SPD_ML_opened_eyes	0.87	cluster_3
mean_distance_peak_Sway_Density_opened_eyes	0.86	cluster_0
weight	0.85	patient
phase_plane_parameters_ML_opened_eyes\long_tim...	0.87	interaction_vars