COMPUTER SCIENCE

FORMAL METHODS – CASE STUDY

PROCESS MINING:

‘INTO ELDERLY LIFE’

AURORA TOMA – 796719

INDEX

INTRODUCTION

This document aims to describe the concept of "*Process Mining*" and its powerful usage, illustrating a university case study of Formal Method course.

Given a certain *Process* regarding a particular domain, the objective of this project is to analyse the former and after that, discover some important information considering different *cases*.

At the end, the acquired knowledge could be used to improve, control, and monitor specific activities or characteristics of that domain.

1. PROCESS MINING

The notion of *Process Mining* was born in 2008 taking the place of the previous most used technique called "*Process Modelling*". The latter was characterized by a procedure that consisted of developing a workflow about a phenomenon; implement it with tools; then a real process was given to it and finally analyse how the obtained model identifies problems or anomalies.

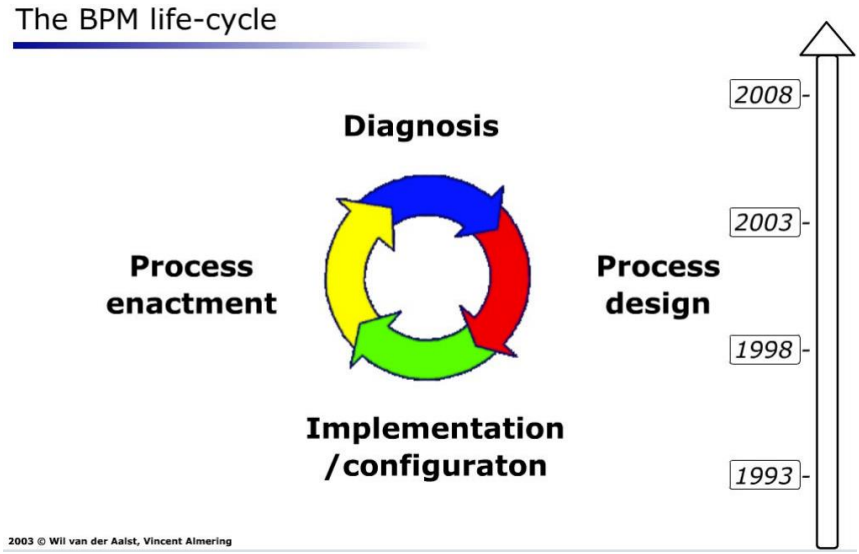


Image1 – The BPM life cycle

Nevertheless, with passing time bigger and bigger amounts of data were available and moreover, there were new technologies that could be used in this field.

Therefore, *Process Mining* was designed making possible to work in the opposite way. In fact, implementing its first step called *Process Discovery*, it's possible to discover which is the process starting from data. After that, there's the second step called *Conformance Checking* that consists of a delta analysis in which, given a schema specified by experts, it's possible to compare the latter with the new model and define how much the two are different. And finally, *Performance Analysis* is computed to identify how it is possible to improve the final model.

To sum up, *Process Mining* aims to analyse event data in order to extract useful information from ICT systems used in organizations or institutions, focusing on bottlenecks, privacy issues, resources optimization and so on.

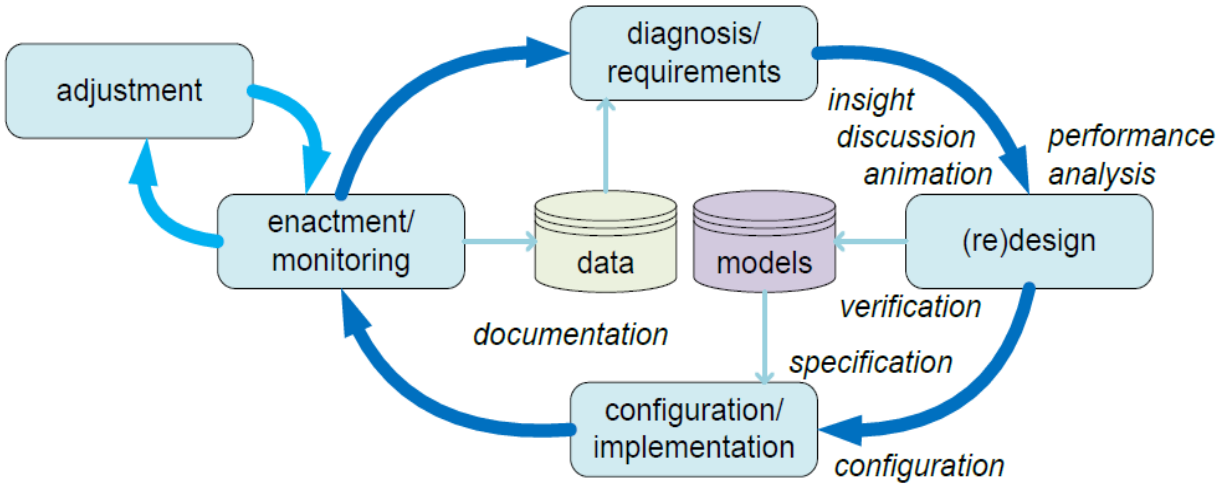


Image2 – Process model life cycle

1.1 DATA NEEDED

As mentioned before, *Process Mining* is a field that involves extracting insights and knowledge from *event logs* to discover, monitor, and improve processes within an organization.

Events and *cases* play essential roles in the context of *Process Mining*.

Introducing the discussion about *Process*, it is a structured set of activities that are performed in a specific order to produce a desired outcome. Each activity represents a task or operation that contributes to the overall process.

Moreover, it is said to be a set of *cases*, in which the latter are composed of *events* (or activities).

A *case* represents a specific instance or execution of a process. It is often identified by a unique *case ID*. Analysing *cases* allows organizations to understand variations in process execution, identify bottlenecks, and discover patterns that may impact organizations' efficiency. As said before, each *case* consists of a series of *events* that occurred during its execution.

Events are recorded in *event logs* and refer to a specific occurrence or activity that takes place at a particular time during the execution of a *process*. *Events* provide the raw data for process mining algorithms to analyse and visualize the sequence of activities within a process.

Event logs capture details such as the timestamp of the *event*, the type of activity and additional attributes that could concern individuals or even machines.

2. TOOLS USED

During the development of this project, different tools were implemented to facilitate and enhance various aspects of the work.

2.1 CSViewer

CSViewer represents a free tool used to show in a fast way large CSV files. It's characterized by an advanced filtering that makes possible to query data, or to explore relationships in data and much more. [1]

It was used to enhance the process of reading dataset and to understand in a clearer way its stored data.

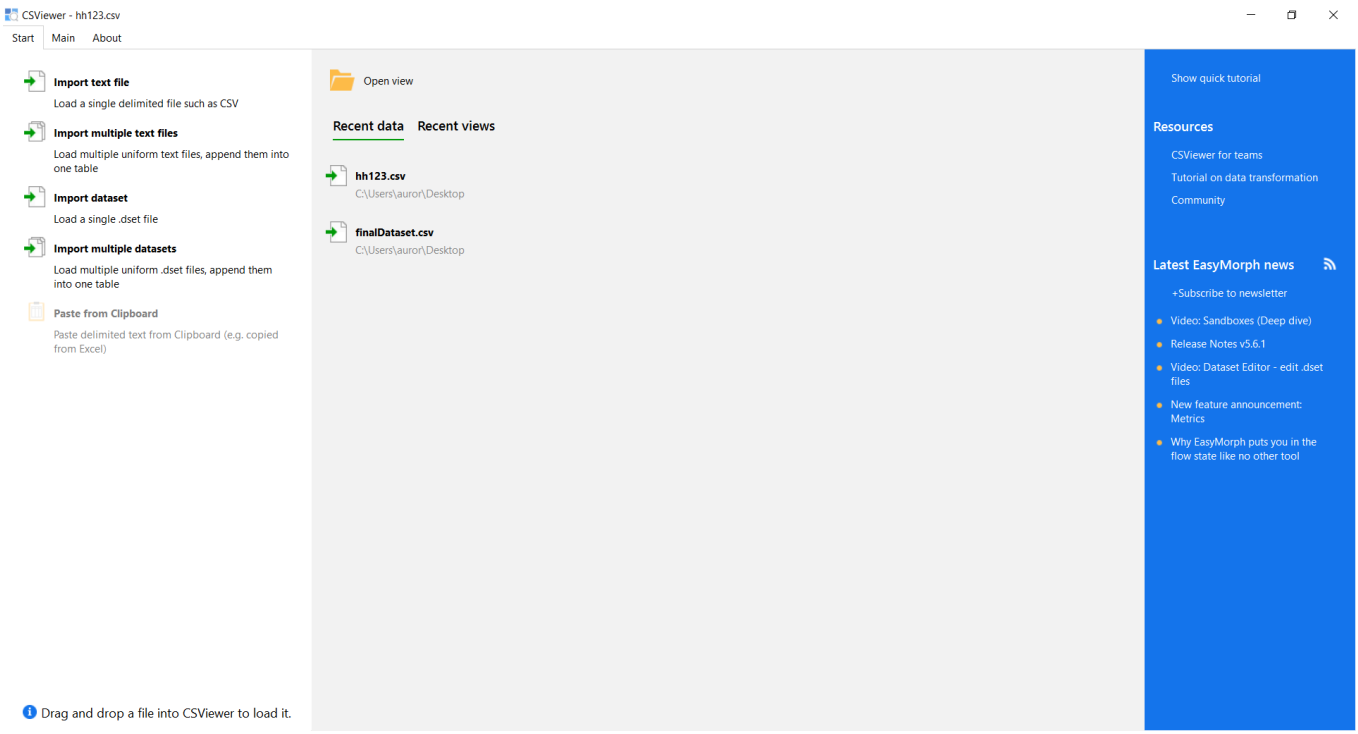


Image3 – CSViewer software launching

2.2 ProM

ProM is the tool predominantly used to analyse data and to extract useful information from a dataset.

ProM is an extensible framework that supports a wide variety of process mining techniques in the form of plug-ins. It is platform independent as it is implemented in Java and can be downloaded free of charge. [2]

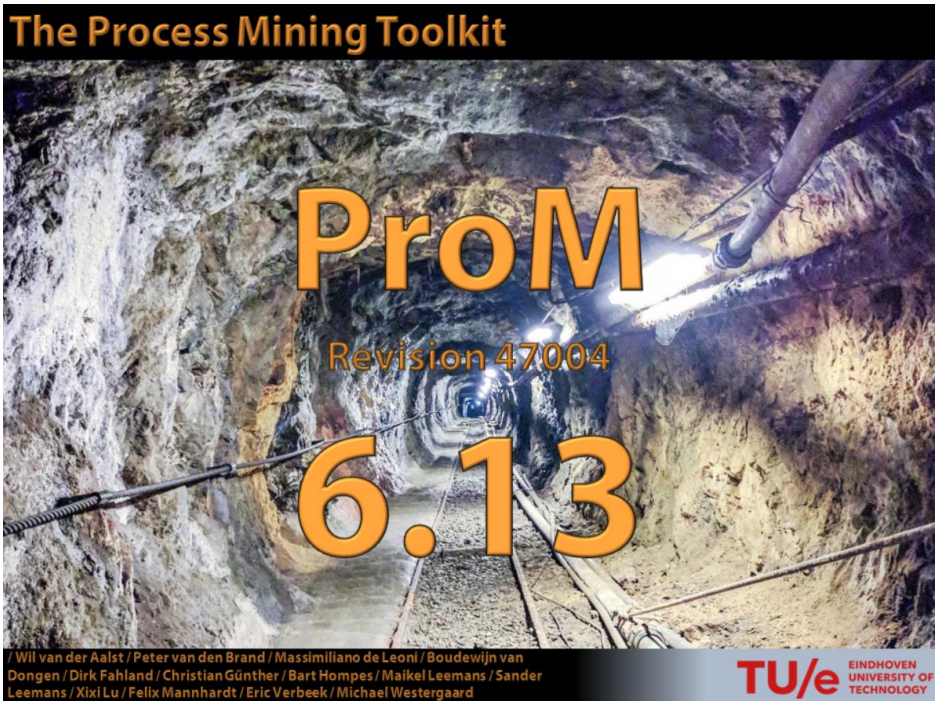


Image4 – ProM software launching

3 DATASET USED

The context of this project is defined by the usage of a dataset that stores information collected by sensors regarding an elderly individual living in a smart home. Indeed, while the increase in life expectancy represents a significant milestone for humanity, it concurrently gives rise to ever-growing challenges related to elderly care, high costs for medical treatments, and more. Therefore, the application of artificial intelligence, robotics, and cutting-edge technologies can aid in addressing these difficulties.

Given that, dataset employed in this project was found using those provided by Center for Advanced Studies in Adaptive Systems (CASAS).

CASAS serves to meet research needs around testing of the technologies using real data through the use of a smart homes environment located on the Washington State University Pullman campus. CASAS also works in partnership with the Smart Environments Research Center to make the environments in which we live and work safer, healthier and more productive through advanced data analytics and adaptive systems. [3]

The chosen dataset is called ‘hh123’ and stores information about activities carried out by a volunteer over a period of one month: from 2/03/2013 to 1/04/2013. Therefore, the main goal of this project could be to identify habits of the considered individual and define if something could be improved. This task can be carried out taking into account different periods of time: one month, one week, one day.

The original dataset is characterized by the following features:

- ‘Date’. Its format is YYYY-MM-DD, in local time.
- ‘Time’. Its format is HH:MM:SS, 12-hour and in local time.
- ‘Sensor’. It represents the name of the sensor activated.
- ‘Translate01’. It’s the room-level sensor location.
- ‘Translate02’. More detailed sensor.

- ‘Message’. It’s the message generated by the sensor.
- ‘SensorType’. It’s the type of sensor generating the event.
- ‘Activity’. It represents the activity identified by the activated sensor. The set of possible activities is composed of Other_Activity, Watch_TV, Cook_Dinner, Dress, Wash_Dinner_Dishes, Wash_Dishes, Sleep, Cook_Breakfast, Drink, Personal_Hygiene, Toilet, Eat_Breakfast, Cook_Lunch, Groom, Bathe, Eat_Dinner, Enter_Home, Phone, Leave_Home, Wash_Breakfast_Dishes, Relax, Entertain_Guests, Evening_Meds, Morning_Meds, Wash_Lunch_Dishes, Bed_Toilet_Transition, Read, Cook, Eat_Lunch, Step_Out, Work_At_Table, Sleep_Out_Of_Bed, Eat.

	Date	Time	Sensor	Translate01	Translate02	Message	SensorType	Activity
1	3/2/2013	02:33:10 AM	MA014	Bedroom	Bedroom	ON	Control4-Mo...	Sleep
2	3/2/2013	02:33:10 AM	MA007	Bedroom	Bedroom	ON	Control4-Mo...	Sleep
3	3/2/2013	02:33:12 AM	MA007	Bedroom	Bedroom	OFF	Control4-Mo...	Sleep
4	3/2/2013	02:33:12 AM	MA014	Bedroom	Bedroom	OFF	Control4-Mo...	Sleep
5	3/2/2013	02:33:14 AM	MA014	Bedroom	Bedroom	ON	Control4-Mo...	Sleep
6	3/2/2013	02:33:15 AM	MA014	Bedroom	Bedroom	OFF	Control4-Mo...	Sleep
7	3/2/2013	02:33:31 AM	MA015	Bedroom	Bedroom	ON	Control4-Mo...	Sleep
8	3/2/2013	02:33:32 AM	MA015	Bedroom	Bedroom	OFF	Control4-Mo...	Sleep
9	3/2/2013	02:35:00 AM	MA015	Bedroom	Bedroom	ON	Control4-Mo...	Sleep
10	3/2/2013	02:35:01 AM	MA015	Bedroom	Bedroom	OFF	Control4-Mo...	Sleep
11	3/2/2013	02:35:01 AM	MA015	Bedroom	Bedroom	ON	Control4-Mo...	Sleep
12	3/2/2013	02:35:02 AM	MA015	Bedroom	Bedroom	OFF	Control4-Mo...	Sleep
13	3/2/2013	02:38:19 AM	MA007	Bedroom	Bedroom	ON	Control4-Mo...	Other_Activity
14	3/2/2013	02:38:20 AM	MA007	Bedroom	Bedroom	OFF	Control4-Mo...	Other_Activity
15	3/2/2013	02:38:20 AM	MA014	Bedroom	Bedroom	ON	Control4-Mo...	Other_Activity
16	3/2/2013	02:38:21 AM	MA007	Bedroom	Bedroom	ON	Control4-Mo...	Other_Activity
17	3/2/2013	02:38:21 AM	MA015	Bedroom	Bedroom	ON	Control4-Mo...	Other_Activity
18	3/2/2013	02:38:21 AM	MA014	Bedroom	Bedroom	OFF	Control4-Mo...	Other_Activity
19	3/2/2013	02:38:22 AM	MA007	Bedroom	Bedroom	OFF	Control4-Mo...	Other_Activity
20	3/2/2013	02:38:22 AM	MA015	Bedroom	Bedroom	OFF	Control4-Mo...	Other_Activity
21	3/2/2013	02:38:26 AM	MA007	Bedroom	Bedroom	ON	Control4-Mo...	Other_Activity
22	3/2/2013	02:38:27 AM	MA007	Bedroom	Bedroom	OFF	Control4-Mo...	Other_Activity

File: C:\Users\auror\Desktop\hh123.csv | 8×152.365 | 8×152.365

Image5 – Reading chosen dataset using CSViewer

3.1 DATA PREPROCESSING

As mentioned in section 1.1, dataset must satisfy some requirements in order to make possible to implement *Process Mining* techniques.

Repeating, *case* represents the main object needed, and it can be of different nature depending on the main objective of the project. In this case, since the main goal is to define whether individual's habits could be healthier or not, considering different intervals of time, the column 'Date' is chosen.

Then, feature 'Activity' is defined as the one representing possible *events* inside a case. And finally, events are characterized by their starting and ending time.

Looking at image5, it's evident that this dataset lacks the columns that represent in an explicit way the starting time of that activity and the ending time.

It can be noted that there are multiple rows identifying the same activity but in a different time, depending on how often sensors detect activities. Therefore, it's intuitive to think that the first row having a certain activity represents the starting time of that activity, and the last one before changing activity represents the ending time.

In light of the above, first of all, conversion of time from 12h format to 24h format was needed in order to avoid the written 'AM' or 'PM' not accepted during the following adjustments. This was done also creating a new single column called 'DateTime', and then it was put in first position.

After that, in order to delete all the rows between the first row representing the starting time and the last one representing the ending time, for each activity carried out, another column was added to the dataset identifying the 'number_activity' used to enumerate events. Then, all rows were grouped by that value and, finally, only the minimum and the maximum values of 'DateTime' were first selected, and then ordered. At the end, the added column 'number_activity' was dropped due to its current uselessness, and then one saving of this work was done.

Doing all that, ‘Date’ column was lost due to its union with ‘Time’ and so it was needed to obtain it again, putting it as first column.

At this point, there are only two rows left for each activity carried out: the first one with timestamp that corresponds to starting time, and the second one with timestamp that corresponds to ending time.

Now it’s needed to move the important value inside the second row inside the first one, but in another new column named ‘EndDateTime’ and, after that, to drop all those rows became useless.

After one last operation of renaming the left column called ‘DateTime’ into ‘StartDateTime’, dataset can be considered as ready to be used in *ProM*.

Details regarding this work are in following google colab link:

https://colab.research.google.com/drive/1XQDs9VpBkuich90e9_ui6eLoASQqNNqC?usp=sharing

After providing the correct format to dataset, depending on which case is going to be considered, division of dataset is needed.

Indeed, if *case* represents activities carried out during a week, it’s required a dataset containing rows regarding all the available four weeks. The created dataset is called ‘WeekDataset-31days’.

Whereas, if *case* corresponds to activities carried out during each single day of the week, there will be needed seven different datasets containing all rows of each single day considered. In this case the created datasets are called as weekday’s names (Monday, Tuesday and so on).

Unfortunately, with available data is impossible to have *case* representing information about one month. This is due to the lack of data regarding other months. Therefore, storing or generating new data could be a solution to this problem, allowing ProM to learn a more meaningful model.

Details regarding this division are in following google colab link:

https://colab.research.google.com/drive/1XswYzo13ORCauur1l_JrAXNhhhi9Mfzj?usp=sharing

4 USAGE OF ProM

As mentioned before, ProM is the main tool used for this Process Mining work. Thanks to its plug-ins it was possible to convert .csv files in .xes files, necessary for ProM to fulfil its duties.

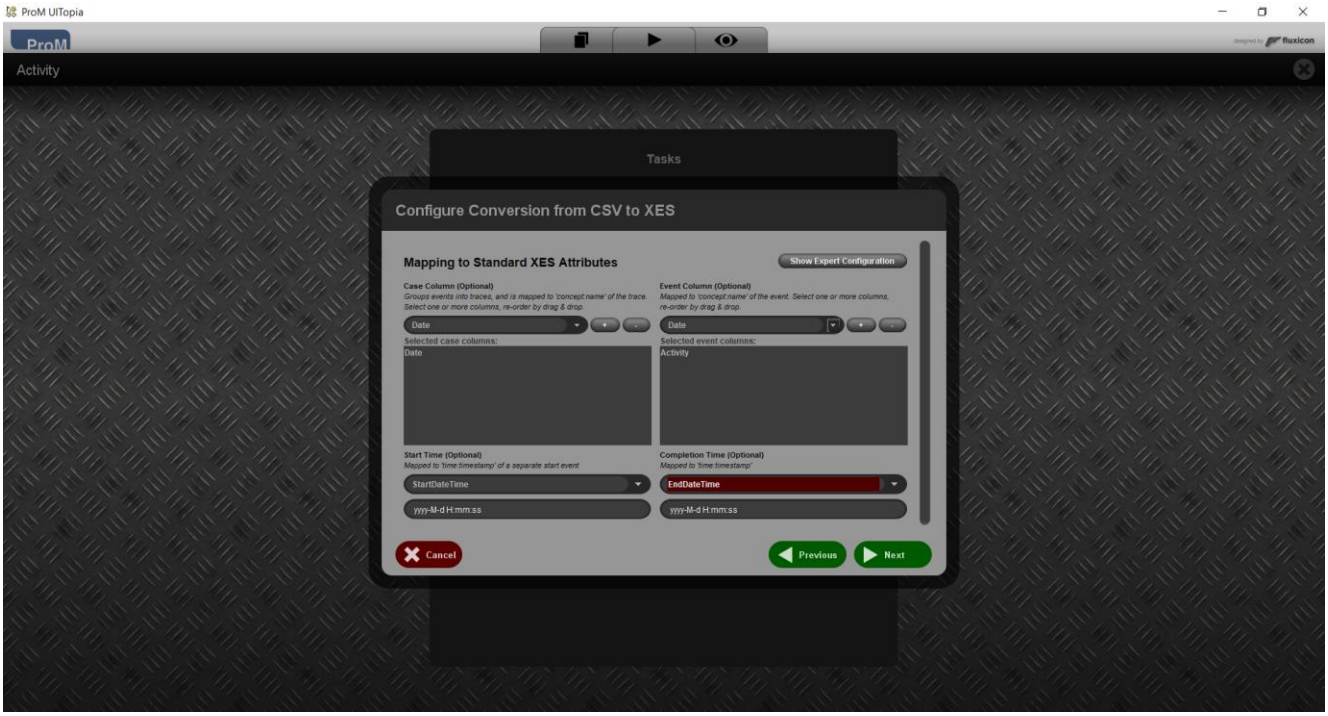


Image6 – Conversion from CSV to XES

Other implemented plug-ins allowed to illustrate details about *cases* or *events*, analyse both and mine information from them. Then, a *Model* was obtained thanks to the plug-in that created a *Petri Net* model and finally, its performance were checked using another plug-in that computed *Precision* and *Fitness* in a 0 - 1 scale.

In theory it's known that *Precision* is concerned with how many of the instances predicted as positive by the model are actually positive, and the closer it is to 1, the less generic the model is.

Whereas *Fitness* (or *Accuracy*) is a broader measure of overall correctness in predictions made by a model, and the closer it is to 1, the more fitting the model is.

4.1 WEEKLY CASE

As concerned *case* corresponding to week, the previous created dataset called ‘WeekDataset-31days’ was given in input to *ProM*.

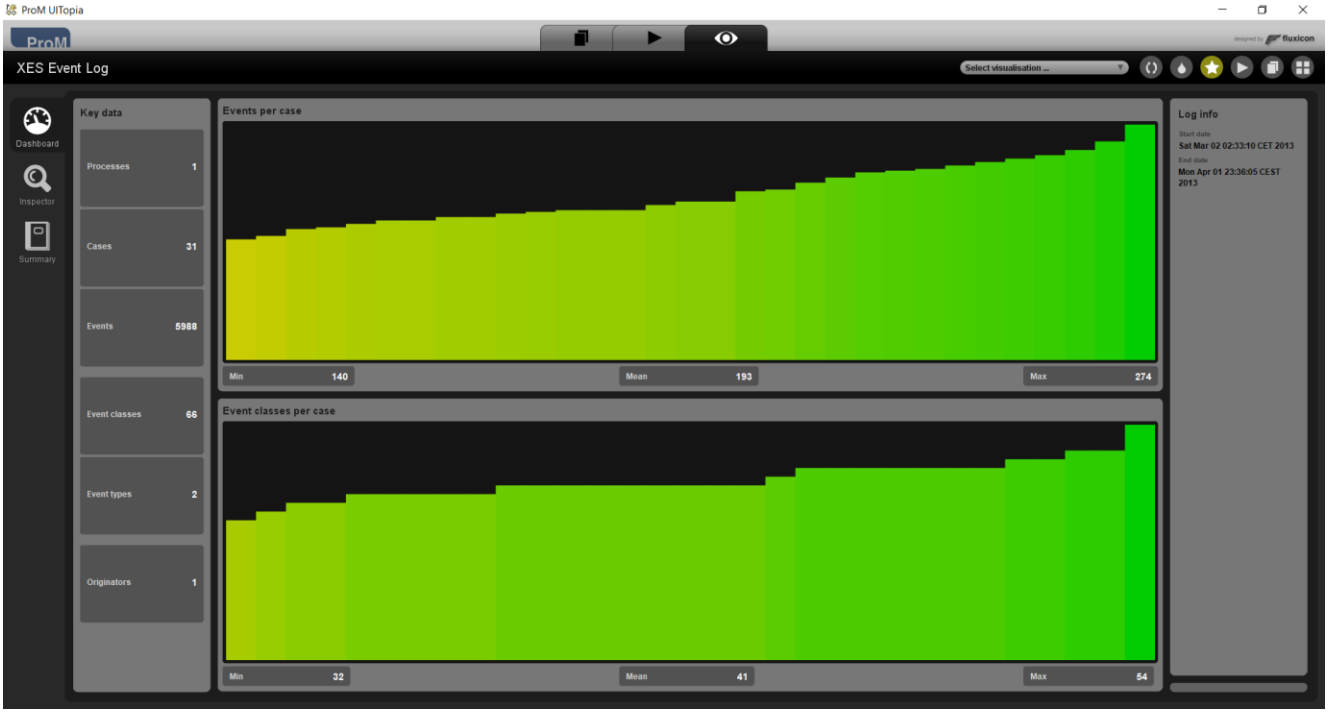


Image7 – Weekly XES event log

As seen in the above image, there are 31cases considered (4weeks in total) and 5988 different events.

Moreover, in the ‘Summary’ section more details about them are shown.

Class	Occurrences (absolute)	Occurrences (relative)
Other_Activity	2924	48,831%
Watch_TV	370	6,179%
Dress	234	3,908%
Drink	232	3,874%
Toilet	216	3,607%
Enter_Home	208	3,474%
Leave_Home	206	3,44%
Cook_Dinner	190	3,173%
Sleep	176	2,939%
Wash_Dishes	130	2,171%
Wash_Dinner_Dishes	128	2,138%
Personal_Hygiene	128	2,138%
Phone	96	1,603%
Cook_Breakfast	96	1,603%
Evening_Meds	74	1,236%
Relax	70	1,169%
Eat_Dinner	70	1,169%
Groom	68	1,136%
Morning_Meds	62	1,035%
Eat_Breakfast	54	0,902%
Eat_Lunch	44	0,735%
Bed_Toilet_Transition	42	0,701%
Cook_Lunch	40	0,668%
Wash_Breakfast_Dishes	28	0,468%
Step_Out	22	0,367%
Wash_Lunch_Dishes	20	0,334%
Bathe	18	0,301%
Work_At_Table	16	0,267%
Entertain_Guests	12	0,2%
Read	6	0,1%
Cook	4	0,067%
Sleep_Out_Of_Bed	2	0,033%
Eat	2	0,033%

Image 8 – All weekly events

From the above image it can be noted that the great part of time of the considered elderly is spent watching TV (370 Occurrences in four weeks), and this could represent a feeling of loneliness or sadness.

On the other hand, activities like ‘Work_At_Table’, ‘Entertain_Guests’ or ‘Read’ are rarer, and therefore it could be necessary to take action and to make elderly more involved in them.

Indeed, the television becomes a companion, a source of distraction from the solitude that can accompany aging. Instead, creating a more interactive and engaging environment can significantly improve the mental well-being of seniors. Additional initiatives such as social gatherings, community events, or

even one-on-one conversations can provide meaningful connections and alleviate the sense of isolation, leading to a healthier lifestyle.

Then, a Model described by a complex Petri Net was obtained as follows.

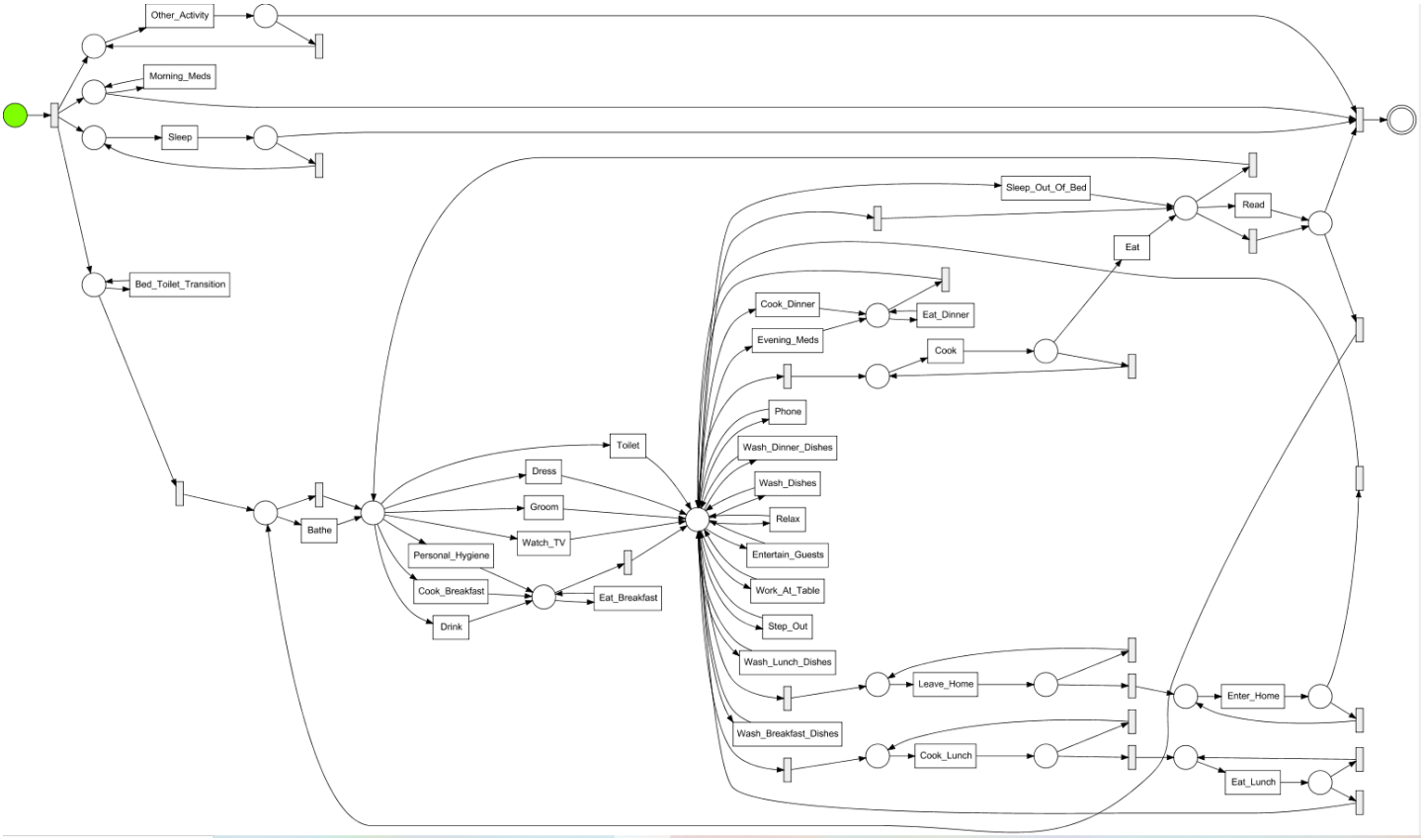
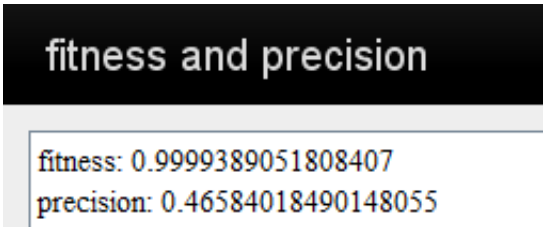


Image9 – Petri Net describing weekly case

And finally, *Precision* and *Fitness* of the model were computed having as result:



4.2 DAILY CASE

Regarding *case* represented by days, the previous created datasets called ‘Monday’, ‘Thursday’, ‘Wednesday’, ‘Tuesday’, ‘Friday’, ‘Saturday’ and ‘Sunday’ were given in input to *ProM* and each of them were analyzed.

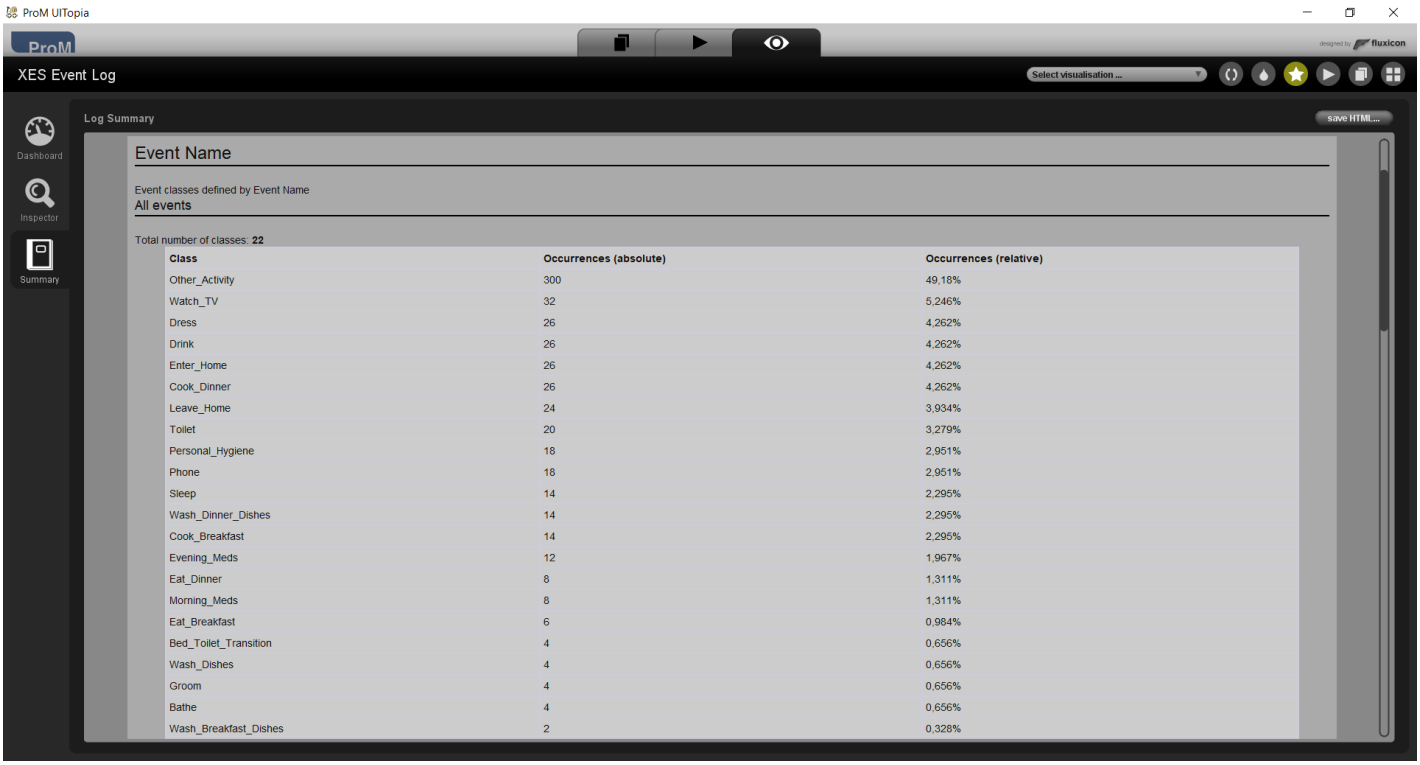
4.2.1 MONDAY

Monday’s dataset is characterized by rows related to the four available Mondays during the month considered.

Therefore, there have been found four different cases and 610 different events.

Image10 – Monday XES event log

In the ‘Summary’ section it’s possible to analyse details about events.



Class	Occurrences (absolute)	Occurrences (relative)
Other_Activity	300	49,18%
Watch_TV	32	5,246%
Dress	26	4,262%
Drink	26	4,262%
Enter_Home	26	4,262%
Cook_Dinner	26	4,262%
Leave_Home	24	3,934%
Toilet	20	3,279%
Personal_Hygiene	18	2,951%
Phone	18	2,951%
Sleep	14	2,295%
Wash_Dinner_Dishes	14	2,295%
Cook_Breakfast	14	2,295%
Evening_Meds	12	1,967%
Eat_Dinner	8	1,311%
Morning_Meds	8	1,311%
Eat_Breakfast	6	0,984%
Bed_Toilet_Transition	4	0,656%
Wash_Dishes	4	0,656%
Groom	4	0,656%
Bathe	4	0,656%
Wash_Breakfast_Dishes	2	0,328%

Image11 – All Monday events

It’s evident that during Monday days the considered elderly usually carries out only 22 out of 33 of the recognized possible activities.

Indeed, he/she isn’t used to read, or work at table or entertain with other guests.

Another important fact to consider could be that during Mondays nights he/she goes to the toilet twice.

Then, a Model described by Petri Net was obtained as follows:

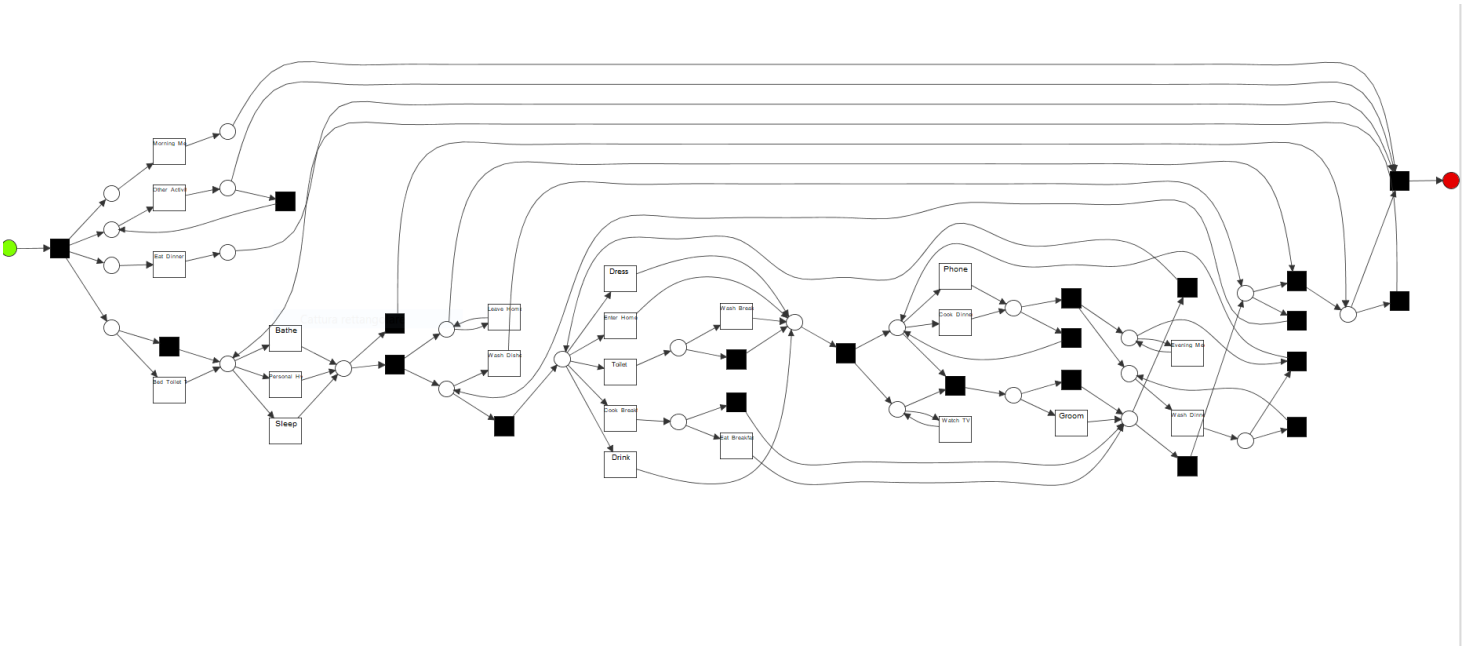
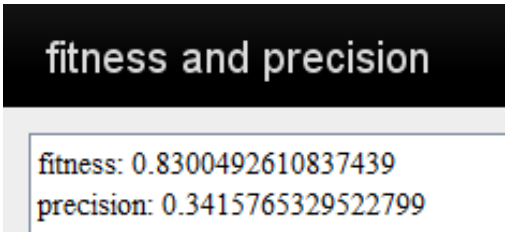


Image12 – Monday Petri Net

And finally, Precision and Fitness were computed having as result:



4.2.2 TUESDAY

Tuesday's dataset is characterized by rows related to the four available Tuesdays during the month considered.

Therefore, there have been found four different cases and 978 different events.

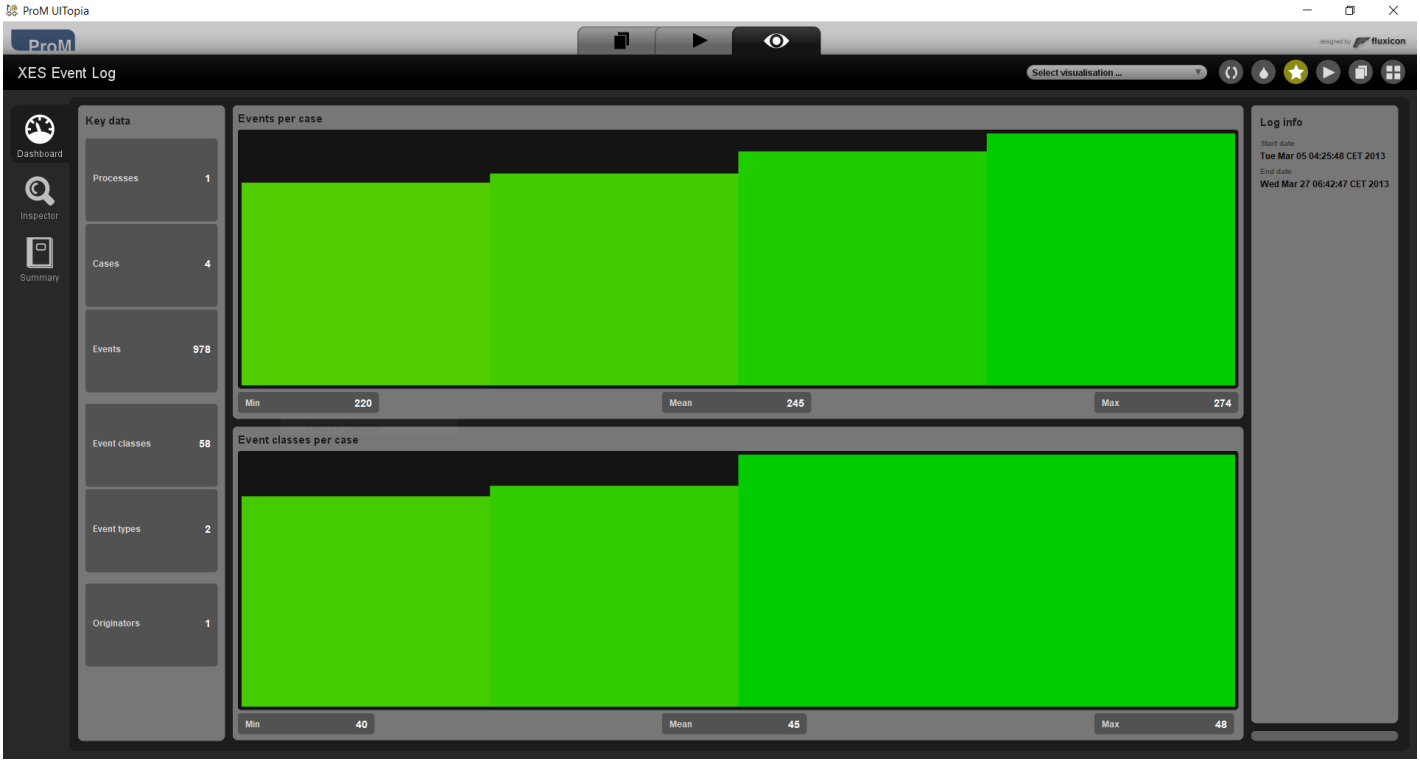


Image13 - Tuesday XES event log

In the ‘Summary’ section it’s possible to analyse details about events.

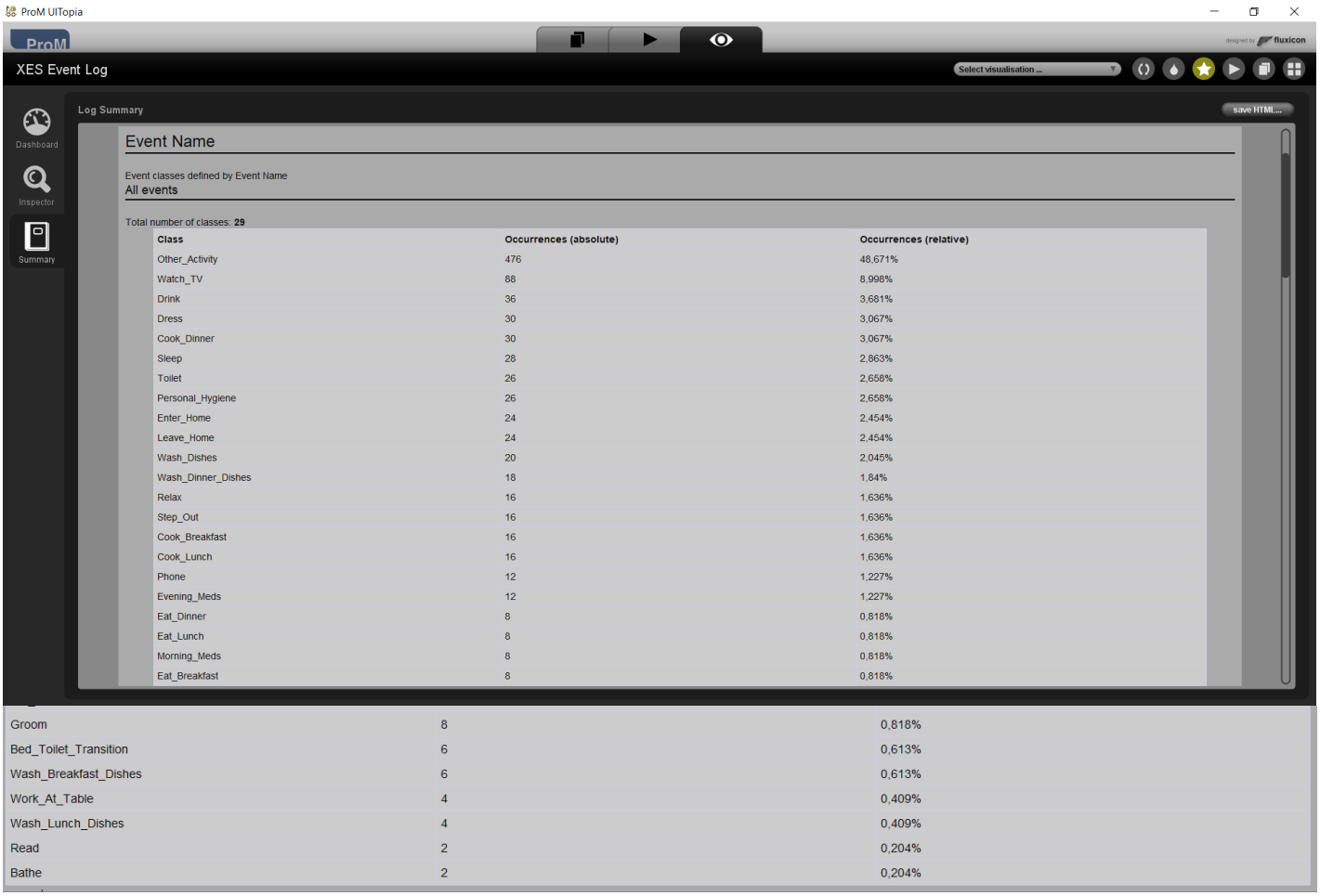


Image14 – All Tuesday events

It’s evident also here that during Tuesday days the considered elderly usually carries out 29 out of 33 of the recognized possible activities.

Indeed, during this day he/she is used to read once or work at table twice, but still, he/she doesn’t entertain with other guests.

Also here, another important fact to consider could be that during Tuesdays nights he/she goes to the toilet three times.

Then, a Model described a quite complex by Petri Net was obtained as follows:

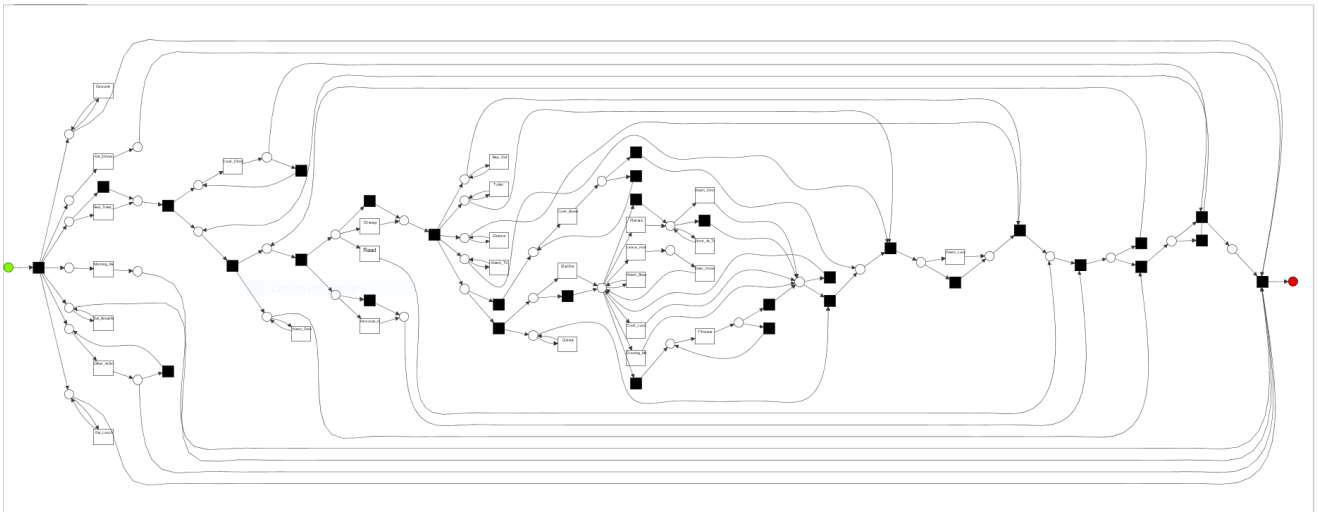
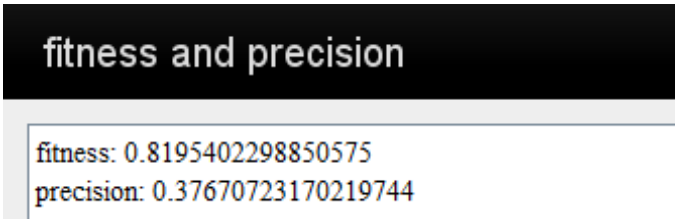


Image15 – Tuesday Petri Net

And finally, Precision and Fitness were computed having as result:



4.2.3 WEDNESDAY

Wednesday's dataset is characterized by rows related to the four available Wednesdays during the month considered.

Therefore, there have been found four different cases and 738 different events.

Image16 – Wednesday XES event log

In the ‘Summary’ section it’s possible to analyse details about events.

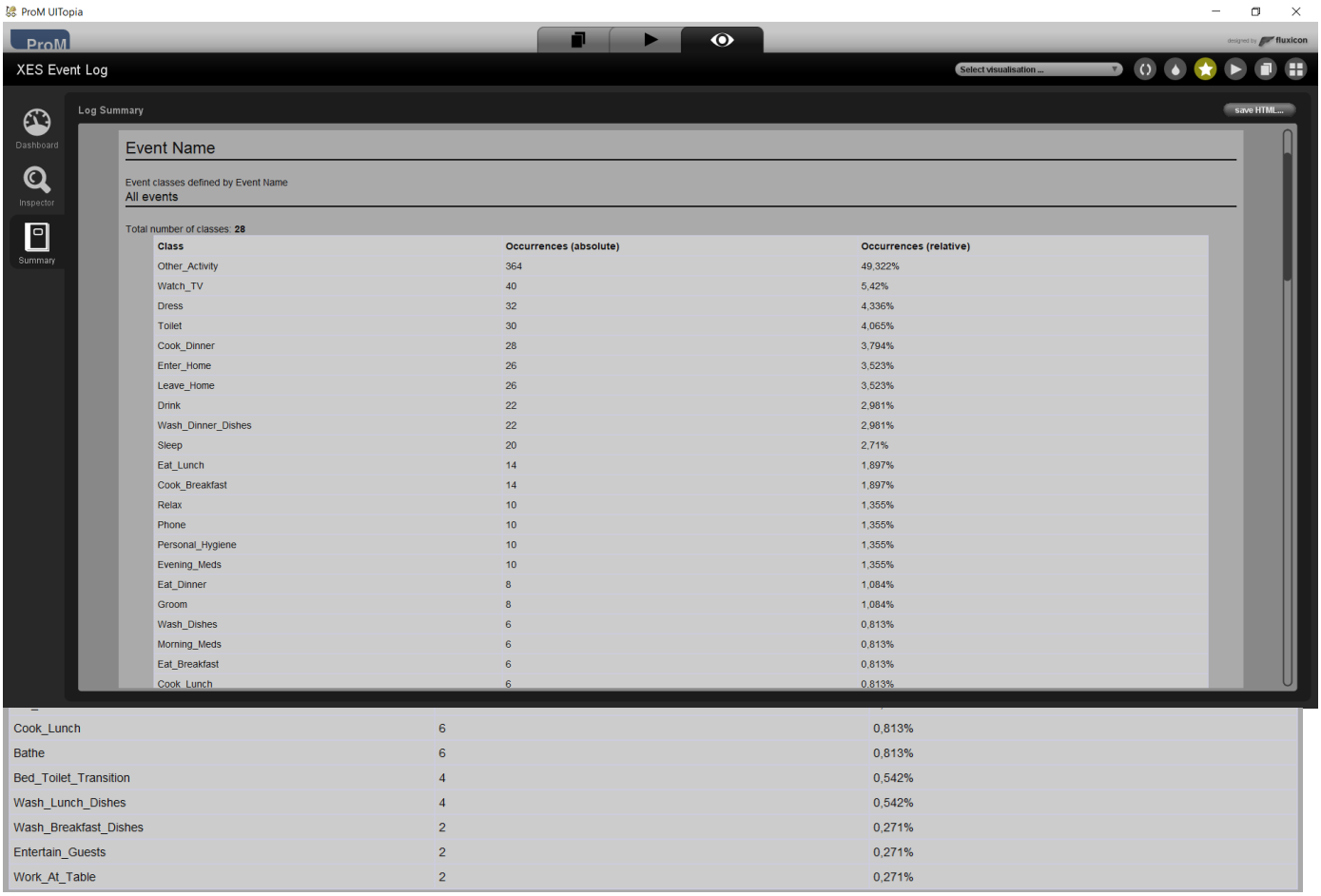


Image17 – All Wednesday events

During Wednesday days the considered elderly usually carries out 28 out of 33 of the recognized possible activities.

Indeed, during this day finally he/she is used to entertain with guests, even if just once; he/she works also at table twice, but he/she doesn’t read.

Another important fact to consider could be that during Wednesdays nights he/she goes to the toilet twice.

Then, a Model described by a quite complex Petri Net was obtained as follows:

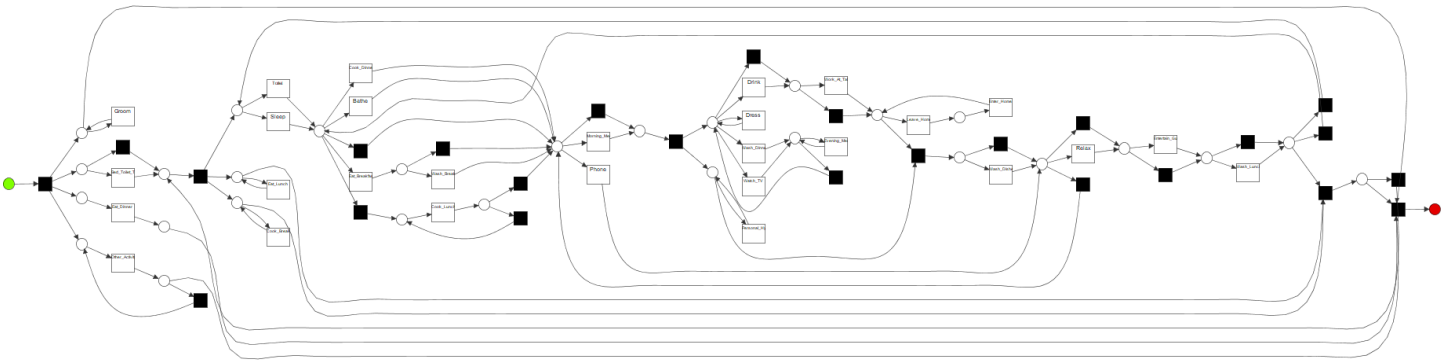
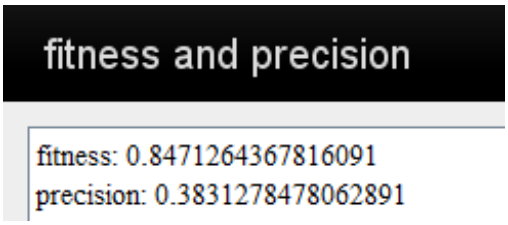


Image17 – Wednesday Petri Net

And finally, Precision and Fitness were computed having as result:



4.2.4 THURSDAY

Thursday's dataset is characterized by rows related to the four available Thursdays during the month considered.

Therefore, there have been found four different cases and 774 different events.

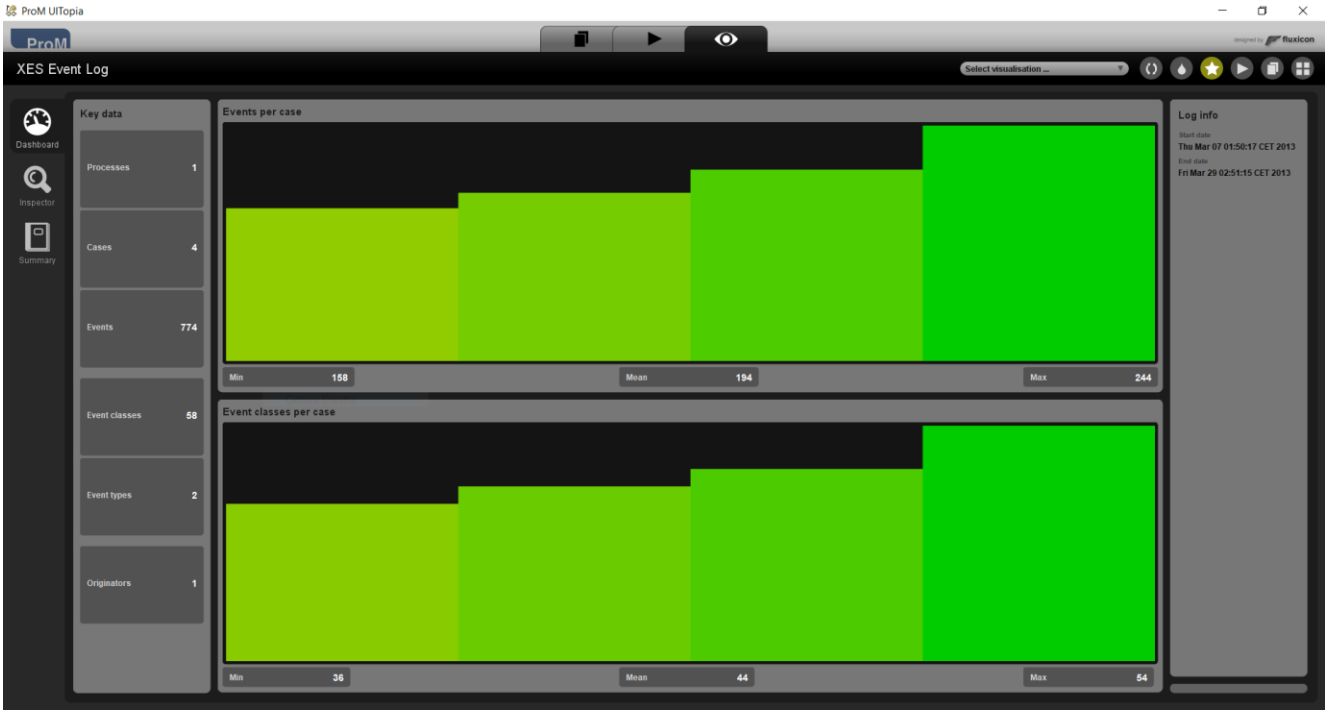


Image18 – Thursday XES event log

In the ‘Summary’ section it’s possible to analyse details about events.

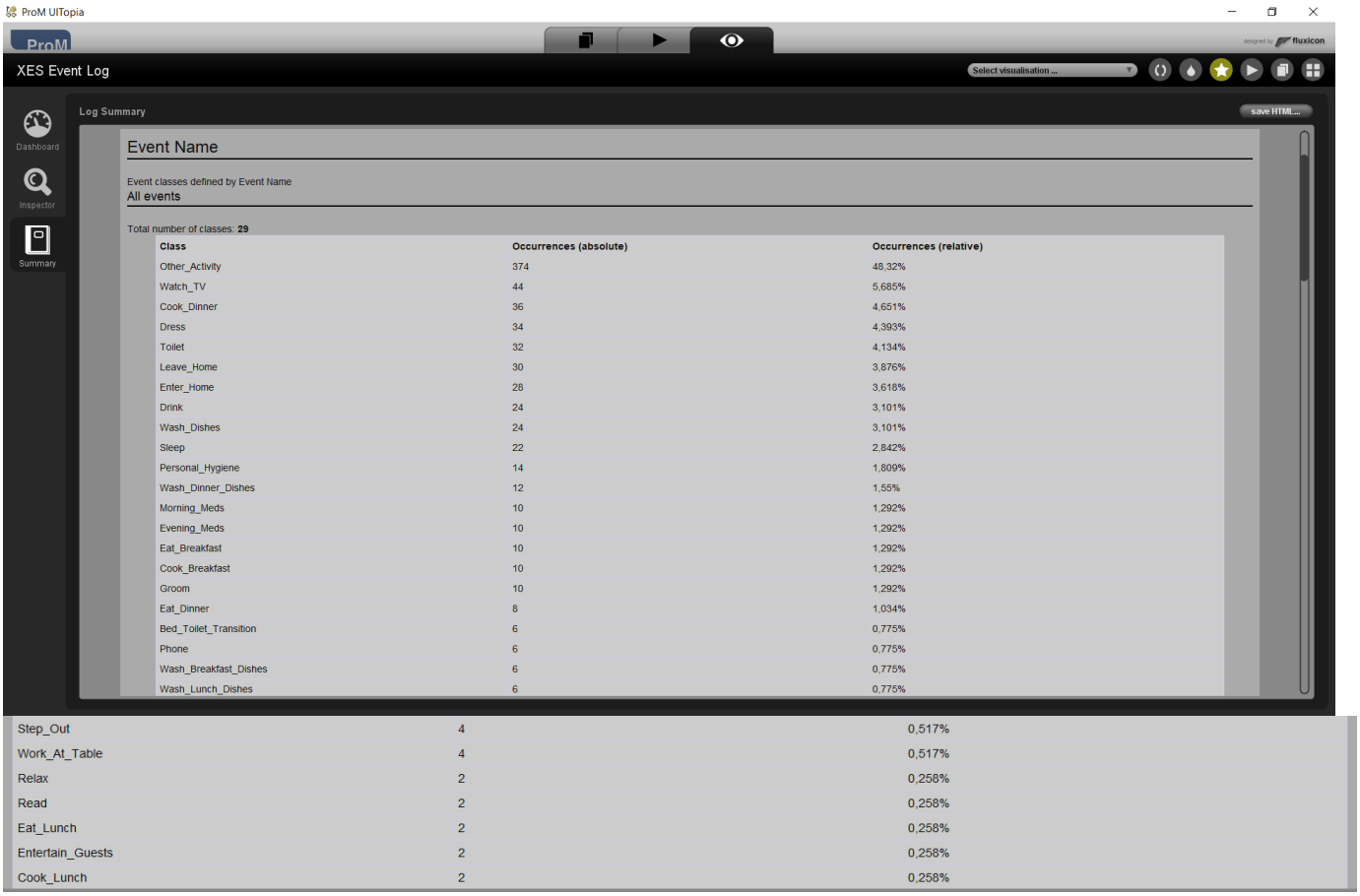


Image19 – All Thursday events

During Thursday days the considered elderly usually carries out 29 out of 33 of the recognized possible activities.

Indeed, during this day he/she is used again to entertain with guests, even if just once; he/she works also at table twice and he/she also read.

Another important fact to consider could be that during Thursday nights he/she goes to the toilet three times.

Then, a Model described by a complex Petri Net was obtained as follows:

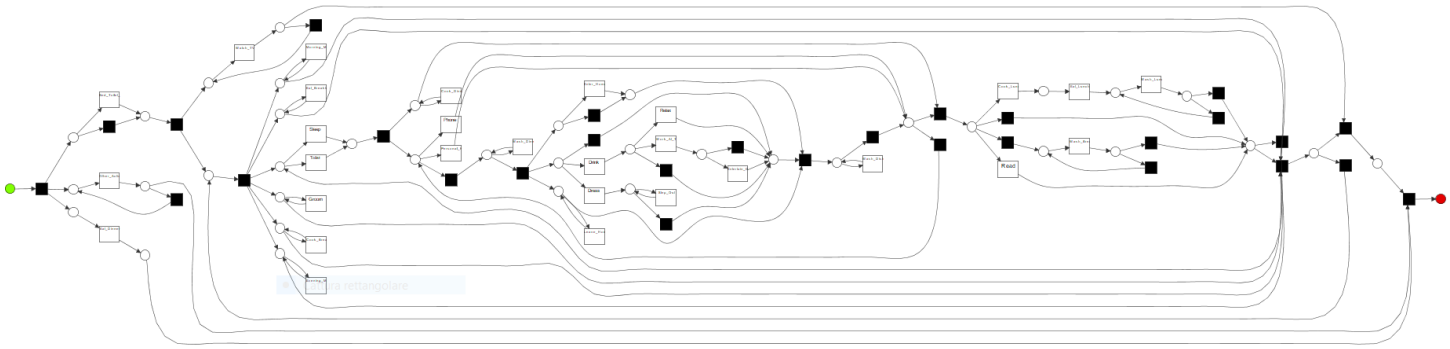
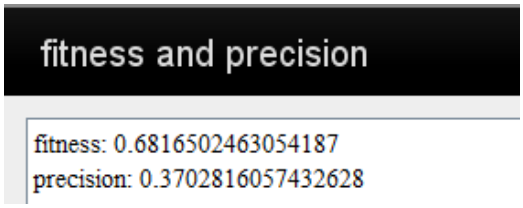


Image20 – Thursday Petri Net

And finally, Precision and Fitness were computed having as result:



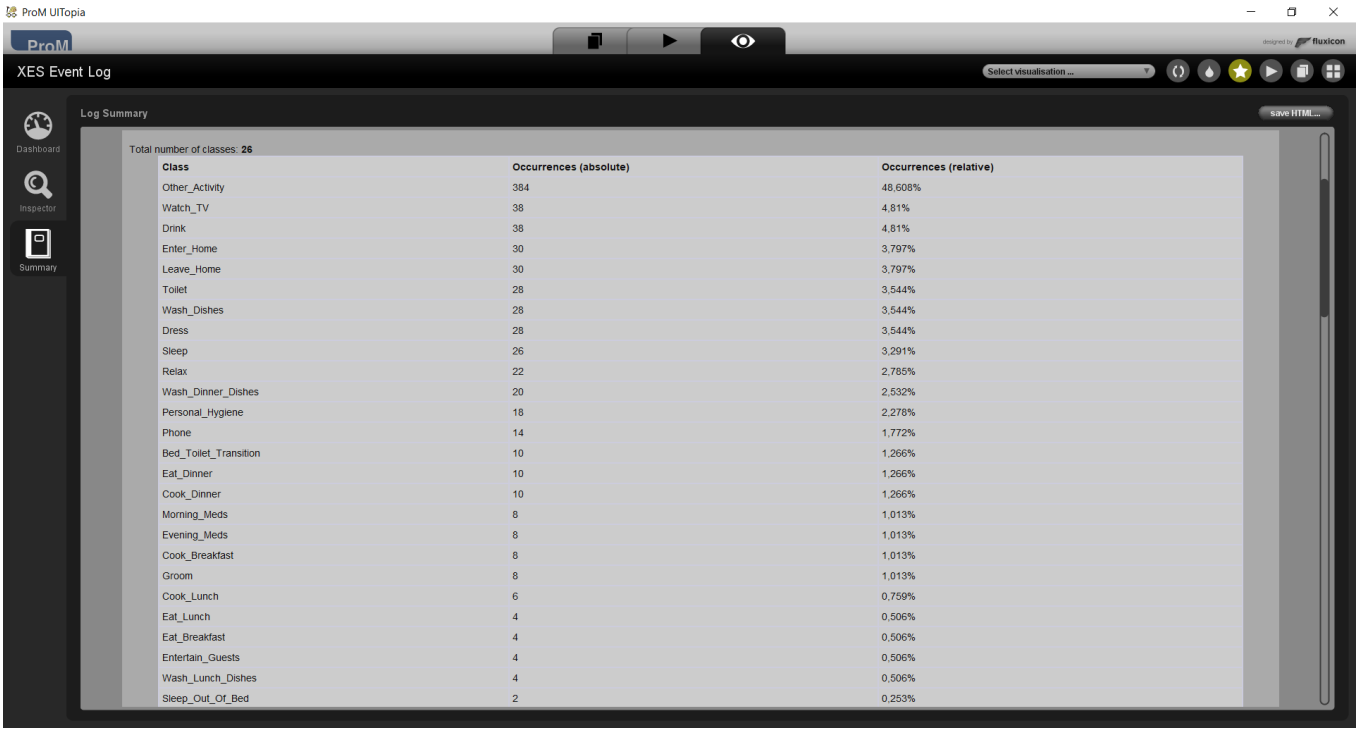
4.2.5 FRIDAY

Friday's dataset is characterized by rows related to the four available Fridays during the month considered.

Therefore, there have been found four different cases and 790 different events.

Image21 – Friday XES event log

In the ‘Summary’ section it’s possible to analyse details about events.



Class	Occurrences (absolute)	Occurrences (relative)
Other_Activity	384	48,608%
Watch_TV	38	4,81%
Drink	38	4,81%
Enter_Home	30	3,797%
Leave_Home	30	3,797%
Toilet	28	3,544%
Wash_Dishes	28	3,544%
Dress	28	3,544%
Sleep	26	3,291%
Relax	22	2,785%
Wash_Dinner_Dishes	20	2,532%
Personal_Hygiene	18	2,278%
Phone	14	1,772%
Bed_Toilet_Transition	10	1,266%
Eat_Dinner	10	1,266%
Cook_Dinner	10	1,266%
Morning_Meds	8	1,013%
Evening_Meds	8	1,013%
Cook_Breakfast	8	1,013%
Groom	8	1,013%
Cook_Lunch	6	0,759%
Eat_Lunch	4	0,506%
Eat_Breakfast	4	0,506%
Entertain_Guests	4	0,506%
Wash_Lunch_Dishes	4	0,506%
Sleep_Out_Of_Bed	2	0,253%

Image22 – All Friday events

During Friday days the considered elderly usually carries out 26 out of 33 of the recognized possible activities.

Indeed, during this day he/she is used to entertain with guests twice, quite unusual, but he/she doesn’t work at table or read.

Another important fact to consider could be that during Friday nights he/she goes to the toilet five times.

Then, a Model described by a quite complex Petri Net was obtained as follows:

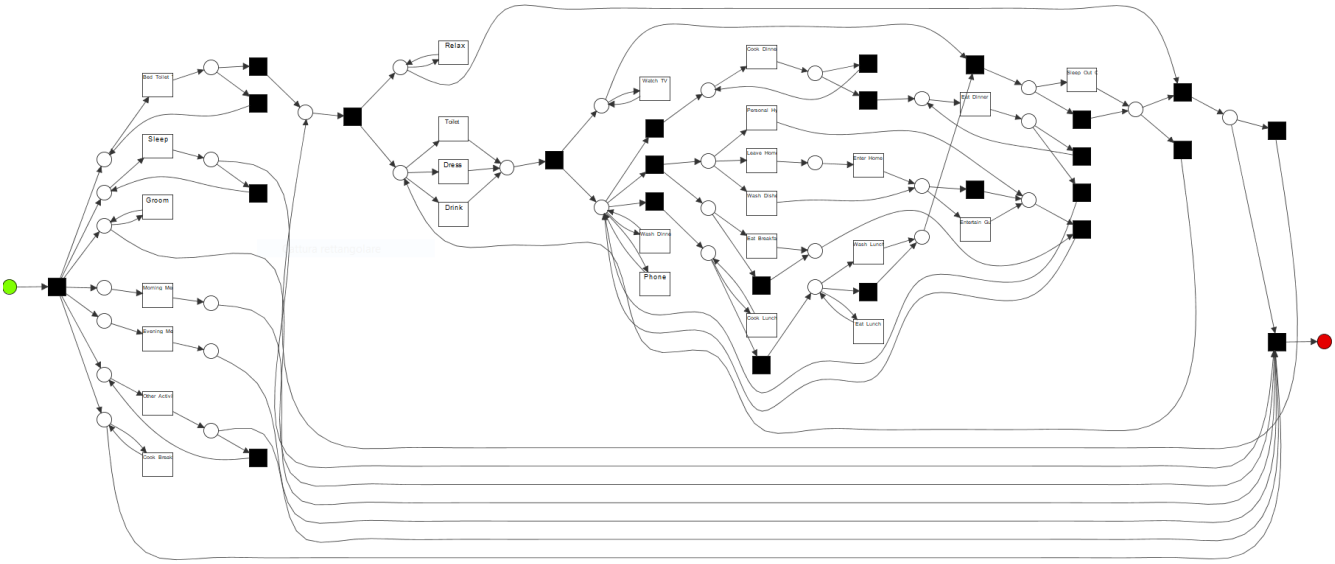
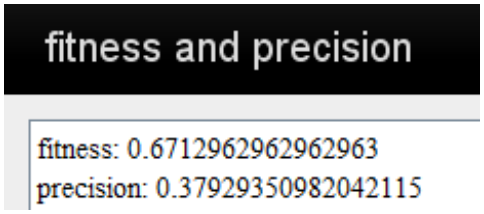


Image23 – Friday Petri Net

And finally, Precision and Fitness were computed having as result:



4.2.6 SATURDAY

Saturday's dataset is characterized by rows related to the four available Fridays during the month considered.

Therefore, there have been found four different cases and 660 different events.

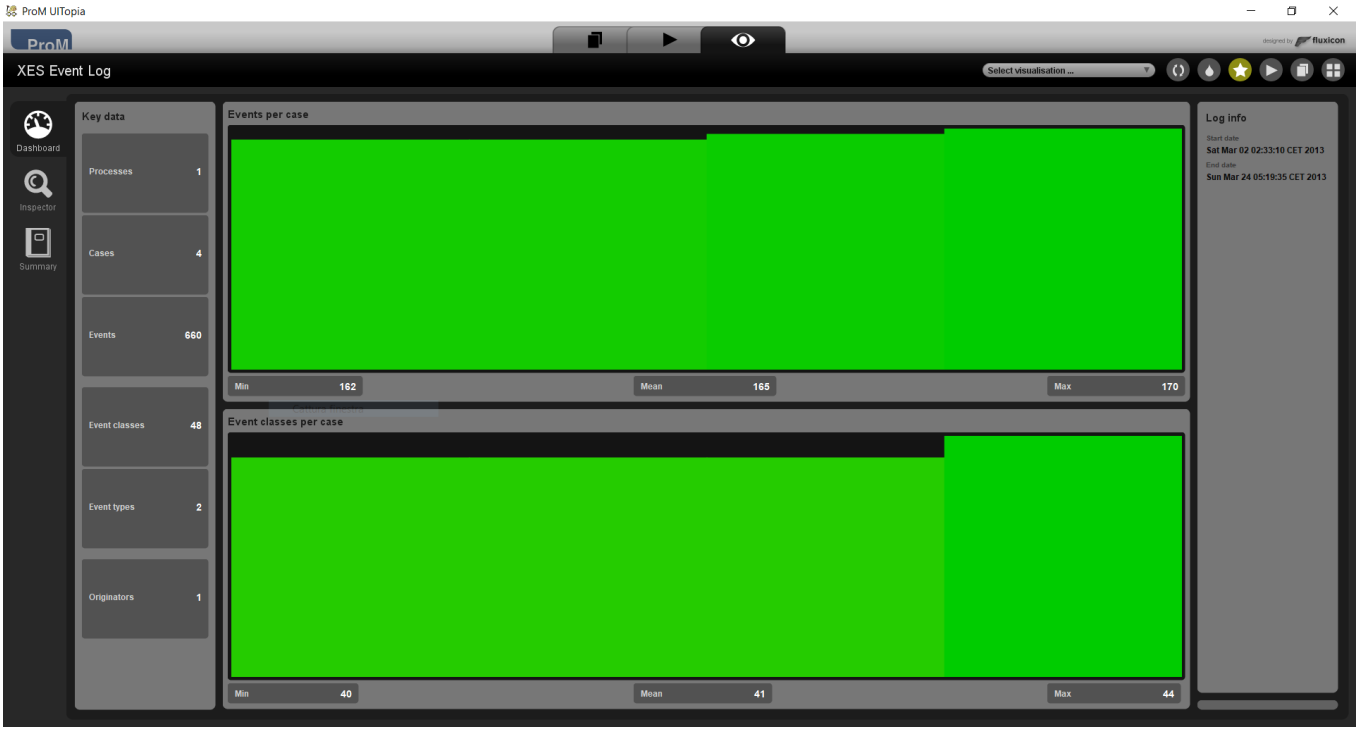


Image24 – Saturday XES event log

In the ‘Summary’ section it’s possible to analyse details about events.

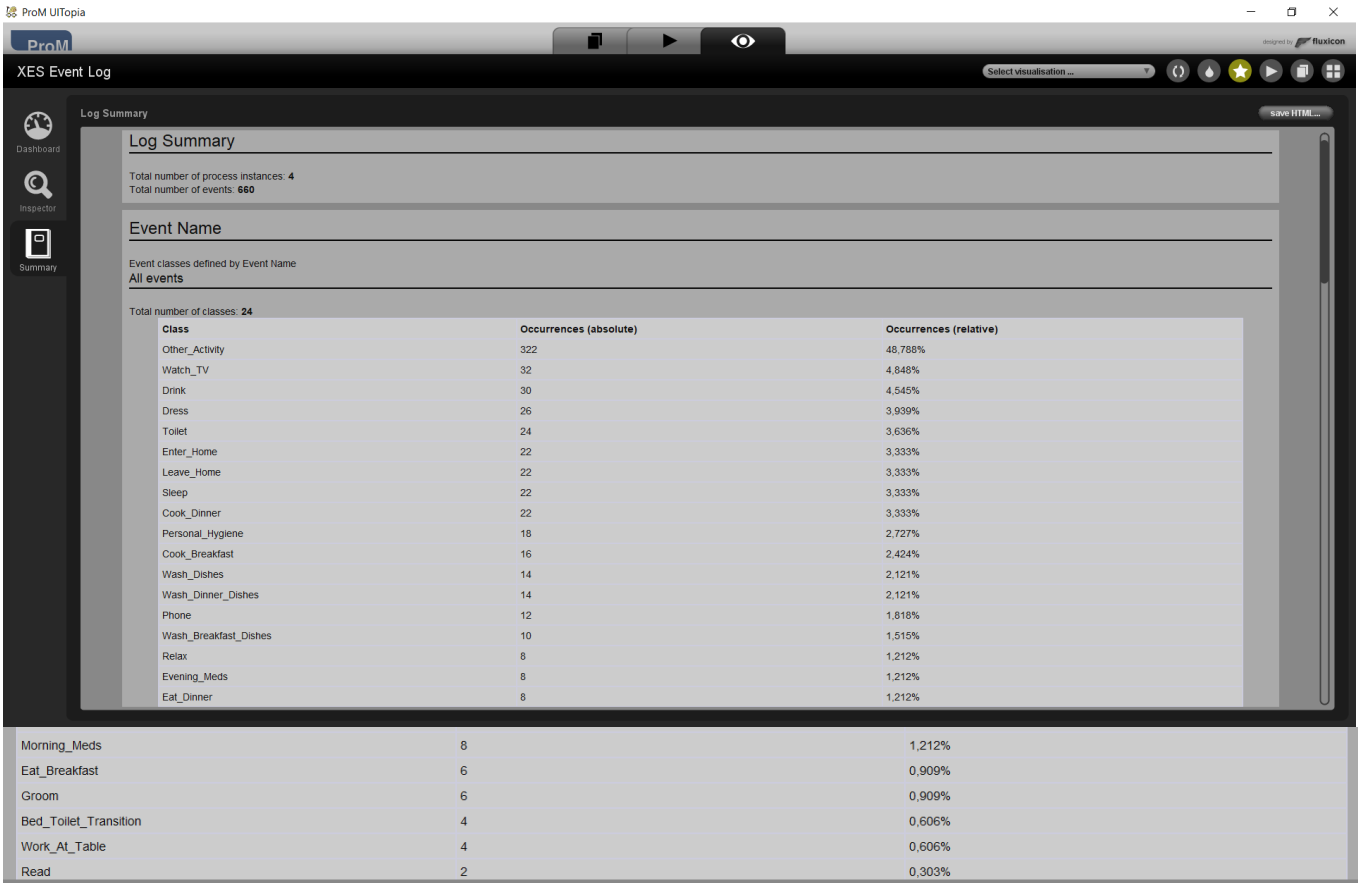


Image25 – All Saturday events

During Saturday days the considered elderly usually carries out 24 out of 33 of the recognized possible activities.

Indeed, during this day he/she is used to work at table twice and to read once, but without entertaining with guests.

Also here, another important fact to consider could be that during Saturday nights he/she goes to the toilet twice.

Then, a Model described by a quite complex Petri Net was obtained as follows:

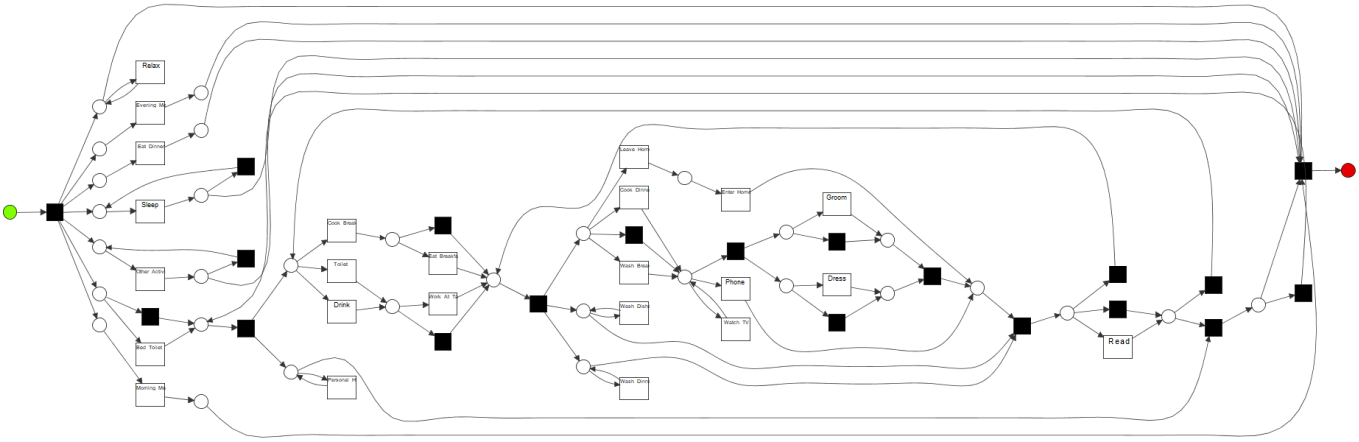
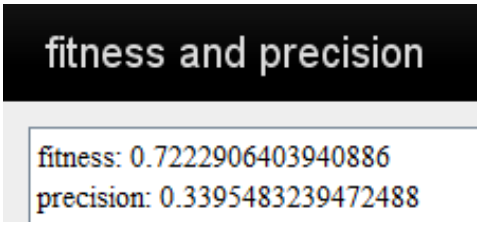


Image26 – Saturday Petri Net

And finally, Precision and Fitness were computed having as result:



4.2.7 SUNDAY

Sunday's dataset is characterized by rows related to the four available Sunday during the month considered.

Therefore, there have been found four different cases and 840 different events.

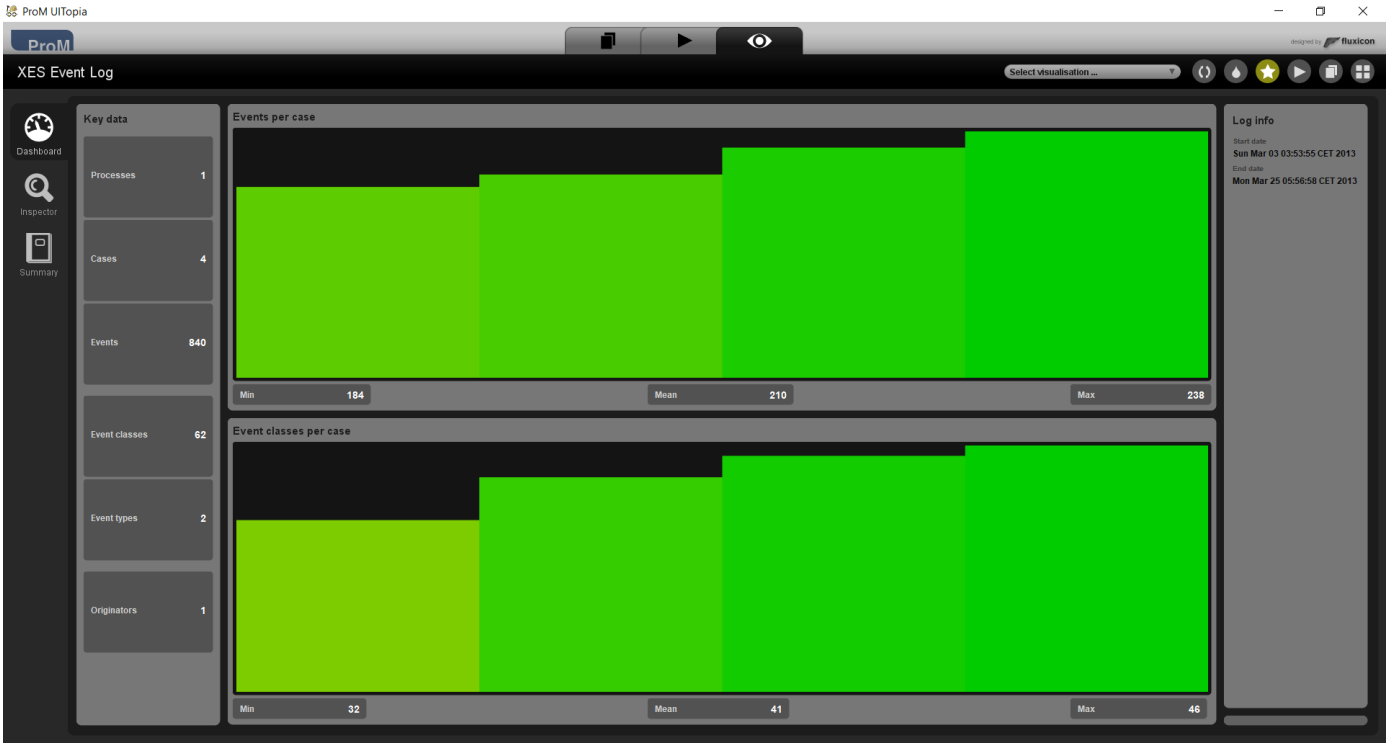


Image27 – Sunday XES event log

In the ‘Summary’ section it’s possible to analyse details about events.

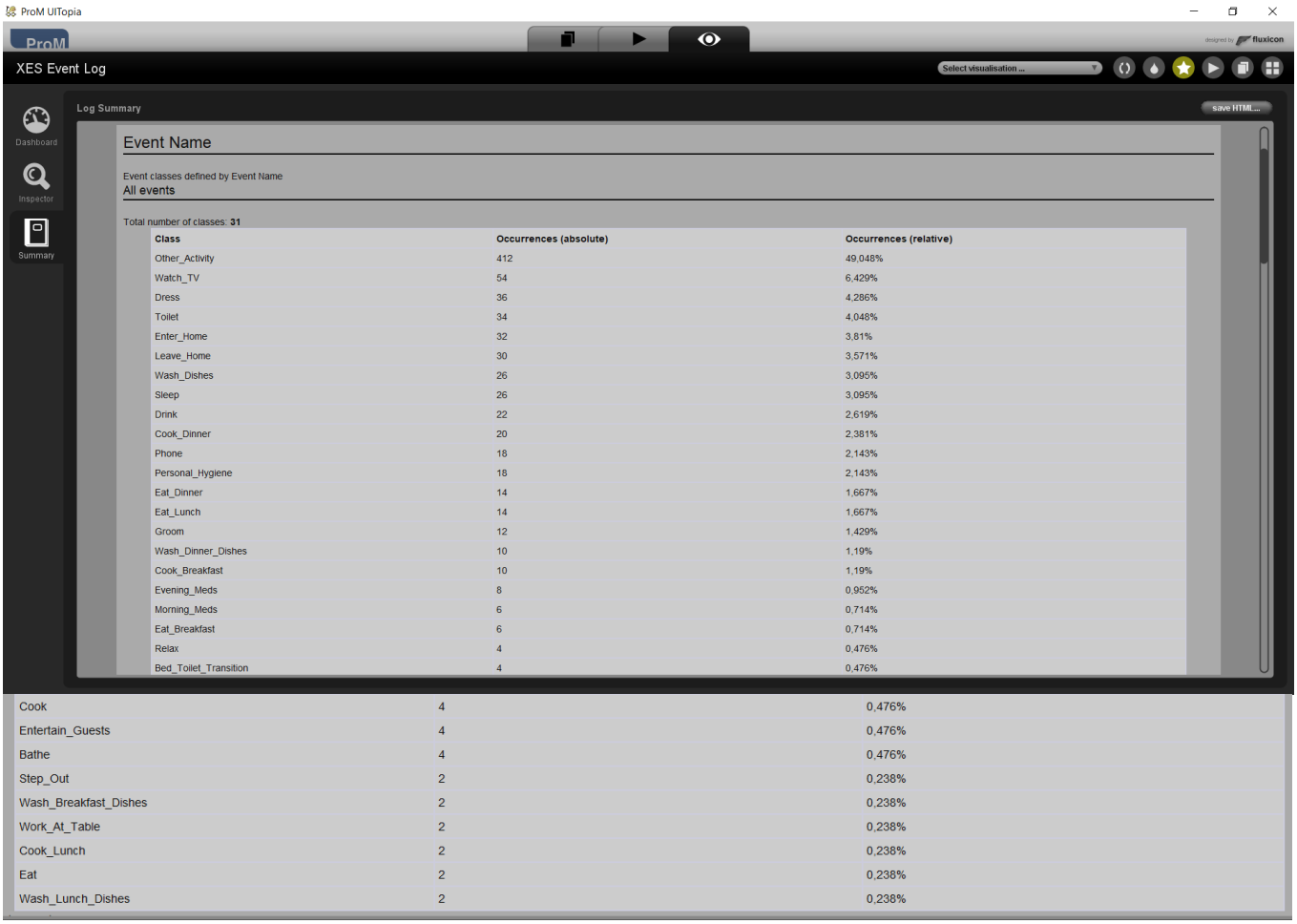


Image28 – All Sunday events

During Sunday days the considered elderly usually carries out 31 out of 33 of the recognized possible activities.

Indeed, during this day he/she is used to work at table once and to entertain with guests twice, but without reading.

Also here, another important fact to consider could be that during Sunday nights he/she goes to the toilet twice.

Then, a Model described by a complex Petri Net was obtained as follows:

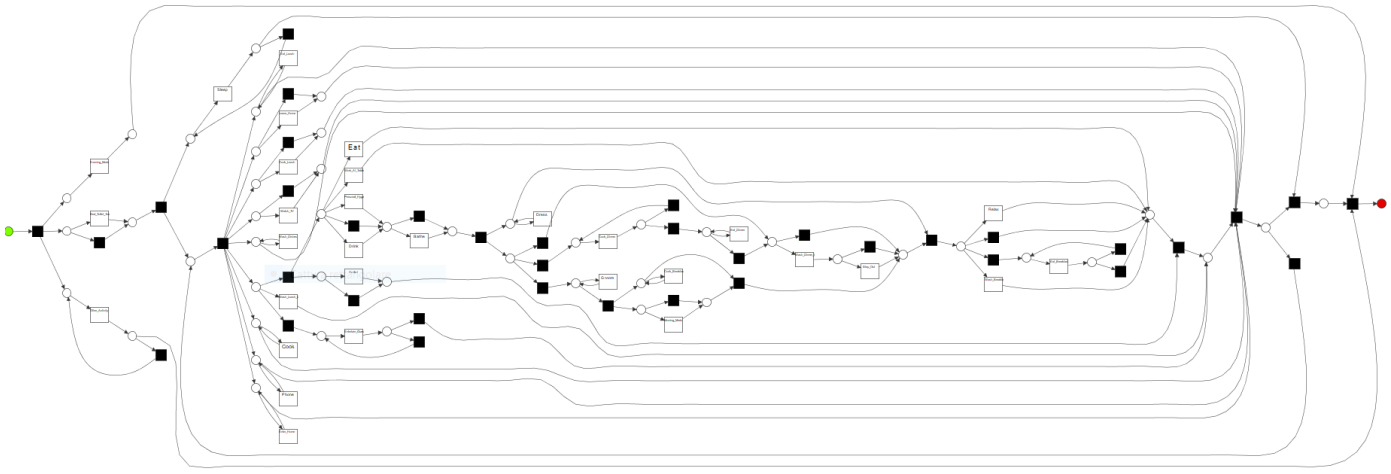
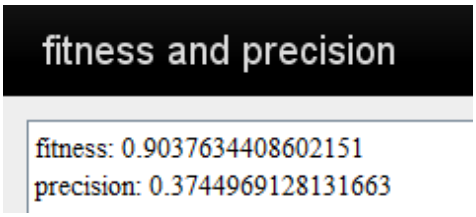


Image26 – Saturday Petri Net

And finally, Precision and Fitness were computed having as result:



CONCLUSION

In conclusion, the utilization of *ProM* has been usefull in extracting valuable insights for the project. Through extensive preprocessing efforts and the adaptation of data into a suitable format in order to be a Prom input, a comprehensive understanding of the elderly individual's various weekly and daily habits has been achieved.

Indeed, the software has facilitated the discovery of crucial information, enhancing the overall effectiveness of the project's objectives.

To sum up, the most significant derived information reveal that the elderly individual spends excessive time watching television, exhibiting signs of self-isolation, potentially indicative of feelings of sadness and loneliness. This observation holds true for both the daily and weekly *cases*.

On the other hand, there is minimal engagement in social activities, which involve interaction with guests or the application of personal skills in activities such as reading or manual work.

These findings underscore the importance of addressing potential emotional and social well-being concerns in the elderly person's routine, highlighting areas where intervention and support may be beneficial for their overall quality of life.

Additionally, the weekly *case* suggest that the elderly individual may require specialized care, as there is a frequent occurrence of nocturnal bathroom visits. This behaviour could be indicative of potential infections or illnesses that, when detected through the use of sensors as in this case, can be addressed and treated, even though the average elderly person tends to deny such needs.

In conclusion, the power of *Process Mining* in analysing data derived from the habits of an elderly individual is evident in its ability to uncover valuable insights that may otherwise remain concealed. Through the utilization of tools such as the Prom software, a comprehensive understanding of daily and weekly routines has been achieved, shedding light on crucial aspects of the

individual's lifestyle. This approach not only highlights patterns in activities but also reveals potential concerns related to emotional well-being, social engagement, and health.

LINKS

https://csviewer.com/ [1]

https://promtools.org/ [2]

https://casas.wsu.edu/ [3]