1. Raw Data Conversion Process

The conversion that took place, that I started working on is from the ‘merged.csv’ file that Joshua created. Everything that he did before that is documented in his GitHub file. All the post-processing took place in the notebooks analyse\_masterfile and analyse\_masterfile2.

The data in the file ‘merged.csv’ is saved, where each row represents a patient. Each cell contains a list with all the values for that patient regarding that specific attribute. Some cells of certain attributes have data that corresponds to each day that the patient was being studied, while others had data that corresponds to a fraction of the days that the patient was in the hospital. The attributes that do not have a recorded value on every day, had another attribute that corresponds to the days that this data is recorded. The goal is to spread the patient values by creating a large dataframe, where each row corresponds to a day of a certain patient. The patient days are stacked on top of each other, and each row has a value that corresponds to the patient number, and the day that the patient visited, along with all the attributes related to the patient on that day.

The first step was to identify the attributes that correspond to each day and the other attributes that depend on another timing approach. The attributes that had a value for each day that the patient is present at the hospital were easy. These values are distributed to correspond to each day that the patient is present at the hospital. As for the attributes that have less data than the number of days, the values of the attributes that correspond to a certain day are inserted. However, if no data is available for certain days, ‘nan’ is inserted into that cell as a starting point.

The attributes that had a value for each day are:

['heathrate\_min', 'heathrate\_max', 'blood\_pressure\_min', 'blood\_pressure\_max', 'mean\_art.bloodpressure\_min', 'mean\_art.bloodpressure\_max', 'temp\_min', 'temp\_max', 'breathing\_rate\_min', 'breathing\_rate\_max', 'bicarbonate\_v\_min', 'bicarbonate\_v\_max', 'bicarbonate\_min', 'bicarbonate\_max', 'ph\_min', 'ph\_max', 'PF\_ratio\_min', 'pO2\_at\_lowest\_PF\_ratio', 'FIO2\_at\_lowest\_PF\_ratio', 'pCO2\_at\_lowest\_PF\_ratio', 'type\_of\_ventilation(0:n-1:noninvasive-2:invasive)', 'urine\_output(mL)', 'urine\_collection(h)', 'renal\_replacement(0:n-1:y)', 'Pneumo(0:n-1:y)', 'Trach\_secration(0:little-1:abundant-2:ab\_with\_prulent)', 'Infiltrate(0:n-1:diffuse-2:localized)', 'ARDS(0:n-1:y)', 'eye\_response(4:spontaneous-3:after\_prompt-2:on\_pain\_sti-1:no)', 'Verbal(5:clear-4:confused-3:single\_words-2:single\_sounds-1:no)', 'Motor(6=>1\_getting\_worse)', 'GCS(1:raised-2:estimated)', 'is\_infected(1:y)', 'degree(1:clinical-2:lab)', 'infection\_origin(1:outpatient-2:nosocomial)', 'Localization(0:n-1:y)', 'Pneumonia(0:n-1:y)', 'other(0:n-1:y)', 'Thoracic(0:n-1:y)', 'Gastro(0:n-1:y)', 'Primary\_bacteria(0:n-1:y)', 'Cathater\_infection(0:n-1:y)', 'Bones\_soft\_tissue(0:n-1:y)', 'Wound(0:n-1:y)', 'intraabdominal(0:n-1:y)', 'ZNS(0:n-1:y)', 'Cardio(0:n-1:y)', 'Urogenital(0:n-1:y)', 'other\_2(0:n-1:y)', 'Brain(0:n-1:y-9:unknown)', 'Thromb(0:n-1:y-9:unknown)', 'Hypox(0:n-1:y-9:unknown)', 'Hypot(0:n-1:y-9:unknown)', 'Dysfu(0:n-1:y-9:unknown)', 'Azid(0:n-1:y-9:unknown)', 'Schock(0:n-1:y-9:unknown)', 'Sepsis\_day', 'Sepsis\_hour', 'INFNEU\_B', 'kreatinin\_min', 'kreatinin\_max', 'crp\_max', 'pct\_max', 'hematocrit\_min', 'hematocrit\_max', 'leukocytes\_min', 'leukocytes\_max', 'thrombocit\_min', 'pip\_give(0:n-1:y)', 'pip\_give\_type(1:empric-2:targetted)', ‘Bolus\_delivery(0:n-1:y)', ‘how\_applied(1:contd-2:prolonged-3:short)', 'contd\_inf\_rate(mL/h)', 'pip\_dosage(mg)', 'blood\_sampling(0:n-1:y)']

The following attributes had multiple entries for the same day. These entries were averaged by finding the mean:

['heathrate\_min', 'heathrate\_max', 'mean\_art.bloodpressure\_min', 'mean\_art.bloodpressure\_max', 'temp\_min', 'temp\_max', 'breathing\_rate\_min', 'breathing\_rate\_max', 'ph\_min', 'ph\_max', ‘PF\_ratio\_min', 'pO2\_at\_lowest\_PF\_ratio', 'FIO2\_at\_lowest\_PF\_ratio', 'urine\_output(mL)', ‘urine\_collection(h)', 'kreatinin\_min', 'crp\_max', 'pct\_max', 'leukocytes\_min', 'leukocytes\_max', 'thrombocit\_min', 'pip\_dosage(mg)', 'kreatinin\_max', 'contd\_inf\_rate(mL/h)', 'bicarbonate\_min', 'bicarbonate\_max']

Attribute new\_dosage was set to 5.0 on the first day.

Any attribute that is date-related, reason-related, or location-related will not be used for the genetic algorithm, but they were relevant for finding the days that some attributes occurred:

['Sepsis\_day', 'Sepsis\_hour', 'reason', 'pip\_start\_hour', 'pip\_stop\_hour', 'Visite\_day\_pip\_clean\_reshape', 'Visite\_day\_sep\_clean\_reshape', 'Start\_day\_pip\_clean\_reshape', 'Start\_hour\_pip\_clean\_reshape', ‘START\_D', 'STARTZ\_Z', 'STARTGRUND\_E', 'STOPP\_D', 'STOPPZ\_Z', 'STOPPGRUND\_E', 'STOPPGRUND\_C', 'Visite\_day\_ccmc\_clean\_reshape', 'site', 'Visite\_day\_chir\_clean\_reshape', 'Date\_micobio\_results', 'Start\_day', 'Start\_hour', 'Start\_reason(1=>6)', 'STARTGRUND\_C', 'Stop\_day', 'Stop\_time', 'Stop\_reason(1=>8)', 'Stop\_reason(why\_8)', 'blood\_sampling\_day', 'blood\_sampling\_hour', 'why\_therapy\_stop(1=>8)', 'Stop\_pip\_day', 'Stop\_pip\_hour', 'new\_dosage\_hour', ‘new\_dosage\_reason(1:TDM-2:side-effect-3:other)', 'KONZENT\_D\_new', 'day\_visited', 'hospital\_admission\_date', 'hospital\_admission\_hour', 'ICU\_date', 'ICU\_hour', 'type\_of\_assigment', 'where\_before\_ICU']

These attributes had very little information and were therefore removed from the dataset:

['bicarbonate\_v\_min', 'bicarbonate\_v\_max', 'APPLI\_E', 'GESDOSIS\_N', 'Antimic(0:n-1:y)', 'CHECK1\_E(day)', 'CHECK2\_E', 'CHECK3\_E', 'CHECK4\_E', 'pip\_mic(mg/l)']

These attributes had less than 0.05 variance threshold from the continuous values and they were removed: 'bicarbonate\_min', 'bicarbonate\_max'.

All the missing values for attributes were dealt with the same way. If the first day is missing, the first value that appears for that patient, on that day, will be inserted into all the days prior to that cell. The same thing is done for if 2 days have empty days in between, where the second value is repeated to all the values prior to that day. Some days contained multiple entries, where the same day had one value, and the rest being ‘nan’. The ‘nan’ was removed to keep the initial value. If there were multiple values for the same day, along with ‘nan’, the ‘nan’s were deleted and the mean is calculated for all the values of that day.

Attributes that depended on the dates of other attributes:

'APPLI\_E' and 'GESDOSIS\_N' depend on the dates on START\_D and STOPP\_D.

'Clinical\_cure(1:healing-2:improvement-3:failure-9:na)' and 'Microbio\_cure(1=>7-9)' depend on Visite\_day\_ccmc\_clean\_reshape. Sometimes, 2 values are used in one cell, Average makes no sense so I took the last recorded value.

'CHECK1\_B(0:n-1:y)', 'CHECK1\_E(day)', 'CHECK2\_B', 'CHECK2\_E', 'CHECK3\_B', 'CHECK3\_E', 'CHECK4\_B', 'CHECK4\_E' sometimes have one day less, so the last day is repeated for the last day.

'Drainage(0:n-1:y)' and 'Drainage\_number' depend on 'Visite\_day\_chir\_clean\_reshape'. Sometimes, 2 values are used in one cell, Average makes no sense so I took the last recorded value.

'target\_based\_on(0:no\_patogen-1:patogen)' and 'change\_in\_pip(0:n-1:y)' depend on 'blood\_sampling\_day'.

'Pathogen\_type(unique\_values)', 'can\_cause\_sepsis(0:unlikely-1:probable-9:unknown)', 'pip\_resistance(1:sensitive-2:intermediate-3:resistant-9:not\_tested)' all depend on 'Date\_micobio\_results'.

'Antimic(0:n-1:y)', 'App\_type(1=>6)', 'total\_daily\_dose(mg)', 'pathogen\_resistance(1,2,3,8,9)' depend on 'Start\_day'.

From the main dataset, there were 3 sections that were created: continuous, discrete, and pathogens. Continuous data were continuous attributes such as ‘hearthrate\_min’ and the discrete were the attributes that had in their name the different attributes used such as ‘renal\_replacement(0:n-1:y)’, except the following three: 'Pathogen\_type(unique\_values)', 'can\_cause\_sepsis(0:unlikely-1:probable-9:unknown)', and 'pip\_resistance(1:sensitive-2:intermediate-3:resistant-9:not\_tested)'. These three were used for creating the pathogens section. The values of these three attributes were not changed.

To do the pathogens section, It was recorded all the different pathogens that are available in the data set, and it was found that there are 36 different pathogens. The data was originally stored by showing the pathogens present in the patient on a certain day, then the next attribute showed which of these caused sepsis, and finally, the last attribute showed the pip resistance of that pathogen. All this data was stored in one cell. These three attributes were broken into 108 different attributes (36\*3) as each column represents a certain pathogen, whether this pathogen can cause sepsis in that state, and finally the pip resistance of that pathogen. It was originally thought that the ‘can cause sepsis’ and ‘pip resistance’ are characteristics of the pathogen. However, it was noticed that some pathogens were in some cases able to cause sepsis, and at other times they had different pip resistance.

As for processing the sofa score, the days that contained a value for time1 and time2 (usually first 2 days) were deleted and the days that contained a value for 21 or 22 were deleted. There were sometimes some missing values of the sofa scores in the middle of the days. To fill this, a natural projection for the sofa score for the patient is created with a linear approximation using a line of best fit. The missing cells used the value of that projection for that day.

1. Dates of patients

What I found is that each patient is at the hospital for a maximum of 15 days, where days 1 to the last day are all consecutive. Some patients had a 16th day, which was typically x days after the 15th day. The values for x that I found are 2 days, 4 days, 6 days, 8 days, 9 days, 10 days, 11 days, 12 days, 13 days, 14 days, 16 days, 19 days, 20 days, 21 days, 25 days, 26 days, 57 days, and 6 months. I found one case where the 16th day is one day after the 15th. The patients that had a day that is x number of days after the last day are only some of the patients that were at the hospital for 15 days. All these values are used for the patient, days 1 to 16 as they all contain data about the patient.

1. Sofa Score 0

After post-processing the sofa scores, I had 40 sofa scores with a value of 1.0, and 27 sofa scores with a value of 0.0. Before processing, I had 6 values with a sofa score of 0.0 and 23 values with a sofa score of 1.0.