Report explaining the analysis

- **Explaining the goal of the task:**

  This task is based on evaluating a label, based on several data features, whether the concerned patient has Sepsis (denoted by '1') or not (denoted by '0'). The dataset given is from the eICU Collaborative Research database.

- **Explaining the dataset:**

  The database has data of numerous patients admitted to various hospitals diagnosed with different diagnosis of diseases. This database has 31 data tables each containing various aspects of medical features like physiological features, lab assessment, amongst others. The entire database has over 200 million data rows some of which are time-dependant (data streamed over period of stay).

  The type of patients we are concerned in this task are those diagnosed with Sepsis.

- **Problem statement of the task:**

  The problem statement focuses on evaluation of three parameters namely, tsuspicion, tSOFA and tsepsis. Based on these three parameters and a certain condition explained below, we have to classify if the concerned patient has Sepsis or not.

  tsuspicion

  1. Clinical suspicion of infection identified as the earlier timestamp of IV antibiotics and blood cultures within a specified duration.

  2. If antibiotics were given first, then the cultures must have been obtained within 24 hours. If cultures were obtained first, then antibiotic must have been subsequently ordered within 72 hours.

  3. Antibiotics must have been administered for at least 72 consecutive hours to be considered.

  tSOFA

The occurrence of end organ damage as identified by a two-point deterioration in SOFA score within a 24-hour period.

tsepsis

The onset time of sepsis is the earlier of tsuspicion and tSOFA as long as tSOFA occurs no more than 24 hours before or 12 hours after tsuspicion, otherwise, the patient is not marked as a sepsis patient.

Specifically, if ***tsuspicion − 24 ≤ tSOFA ≤ tsuspicion + 12,*** then

tsepsis = min(tsuspicion, tSOFA).

- **Plan of approach:**

A lot of things need to be considered while planning to analyse a database so large. To perform the analysis I have used Pandas and Numpy module of Python. My python version is 3.6 and I'm using Jupyter Notebooks.

1) The first step is to note the features that would be required and the tables that contain them. Select only those tables as the tables are quite large and unnecessary loading of datasets would affect the computing speed. Sometimes, the computer might even crash or hang up.

2) The next step is to clean the data and join the selected tables on appropriate axes so that we would get minimum amount of tables.

3) Then, I have calculated the tsuspicion as the timestamp of the drug administration. I, however, could not find the time of blood cultures so I considered only the timestamp of drug administration. I found this using the 'medication' dataset. Also, I've filtered out only the patients diagnosed with sepsis using the 'diagnosis' dataset.

4) Next step was to calculate the tSOFA. The SOFA scores can be calculated by using a simple chart which is as follows (chart was provided by the mentors):

Table 1. Sequential [Sepsis-Related] Organ Failure Assessment Score[a]

System	Score 0	1	2	3	4
Respiration					
Pao_2/Fio_2, mm Hg (kPa)	≥400 (53.3)	<400 (53.3)	<300 (40)	<200 (26.7) with respiratory support	<100 (13.3) with respiratory support
Coagulation					
Platelets, $\times 10^3/\mu L$	≥150	<150	<100	<50	<20
Liver					
Bilirubin, mg/dL (μmol/L)	<1.2 (20)	1.2-1.9 (20-32)	2.0-5.9 (33-101)	6.0-11.9 (102-204)	>12.0 (204)
Cardiovascular	MAP ≥70 mm Hg	MAP <70 mm Hg	Dopamine <5 or dobutamine (any dose)[b]	Dopamine 5.1-15 or epinephrine ≤0.1 or norepinephrine ≤0.1[b]	Dopamine >15 or epinephrine >0.1 or norepinephrine >0.1[b]
Central nervous system					
Glasgow Coma Scale score[c]	15	13-14	10-12	6-9	<6
Renal					
Creatinine, mg/dL (μmol/L)	<1.2 (110)	1.2-1.9 (110-170)	2.0-3.4 (171-299)	3.5-4.9 (300-440)	>5.0 (440)
Urine output, mL/d				<500	<200

Abbreviations: Fio_2, fraction of inspired oxygen; MAP, mean arterial pressure; Pao_2, partial pressure of oxygen.

[a] Adapted from Vincent et al.[27]

[b] Catecholamine doses are given as μg/kg/min for at least 1 hour.

[c] Glasgow Coma Scale scores range from 3-15; higher score indicates better neurological function.

5) Using this chart, the SOFA scores can be easily calculated. To get the required features, I have used the 'apacheApsVar' dataset as most of the features are readily available. However, some, like MAP(Mean Arterial Pressure) and GCS(Glasgow Coma Scale) need to be calculated. I am attaching links that will help you to know how to calculate these features.

   GCS : https://en.wikipedia.org/wiki/Glasgow_Coma_Scale
   MAP: https://www.nursingcenter.com/ncblog/december-2011/calculating-the-map

6) Using these features, I have calculated the SOFA score and labelled it as initial SOFA score.
7) Then, I have analysed the datasets containing stream data of patients in order to calculate the change in SOFA scores over a period of time. After a lot of studying those datasets, I selected the vitals dataset as it contained the streamed data about systolic and diastolic blood pressure of patients. Using them I have found the change in MAP over a period of a day. Unfortunately, this was the only feature I could extract from datasets with streaming data. Thus, I have considered the change in SOFA score using the MAP scores only.
8) Next, I found out the patients who have a deteriorated SOFA score within 24 hours of initial SOFA score. This time of deterioration is labelled as tSOFA.
9) Finally I have calculated the tSepsis of patients using the formula as stated above.
10) I stored this dataset by the name of 'finalLabels.csv'.