**Discovering the factors that affects the rate of lactate level in sepsis patients: Using Machine Learning Algorithm**

**Introduction**:

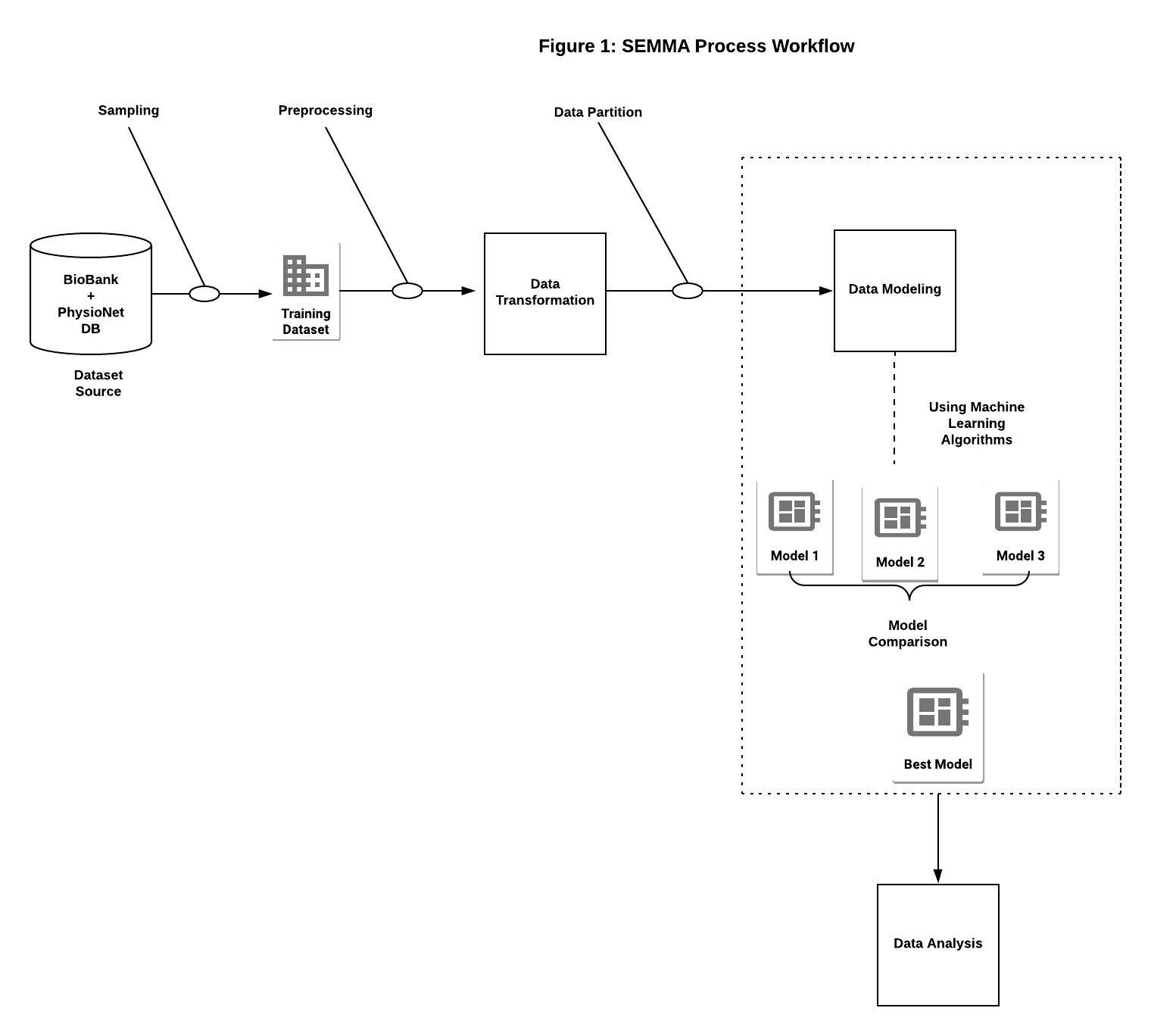
Sepsis is a life-threatening disease, caused as a result of the body's response to infection. Body's immune system overdrive to fight infection (for ex; a simple scratch) and eventually leading to multiple organ failure, shock and ultimately leading to death. Anyone whose immune system is depressed can escalate to sepsis. WHO has identified elderly people, pregnant women, neonates, hospitalized patients, and people with HIV/AIDS, liver cirrhosis, cancer, kidney disease, autoimmune diseases, and no spleen, are at higher risk.

Early symptoms of this disease include; abnormal temperature (temperature above 38 and temperature below 36 ), chills/shivering, Fast Heartbeat, Fast Breathing and Confusion/Drowsiness; And some of the severe symptoms include- diarrhea, vomiting, severe muscle pain, urine trouble, etc. These symptoms could be due to sepsis and can be only confirmed by a blood test. If the blood results show the presence of bacteria and depending on the type of bacteria and patient’s allergy to antibiotics, IV can be administered. However recent studies have shown that lactate level rate is a vital sign to diagnose sepsis.

It is very important to treat sepsis in its early stage due to its fatality. According to WHO;it is estimated that more than 30 million people are affected with sepsis worldwide every year, potentially leading to 6 million deaths**.**

Analyzing the sepsis patient’s phenotype data and applying machine learning algorithms, we can study how the lactate level rate affects the sepsis patients. Additionally, this could result in better healthcare delivery systems for the population risk of sepsis.

**Methods:**

SEMMA approach (Figure 1)is adopted for the methodology. In regards to the SEMMA evaluation system, the first process is the ***Sample*** Process. **(Need to check the dataset from Biobank and PhysioNet – to mention the number of total rows, columns, and target, Also will be mentioned if we are planning to concatenate different files to 1 file format )** In the Explore stage, each variable was evaluated for correct classification type, and adjusted if necessary. 

Many variables such as Gender(1= Male, 2=Female) can be reclassified from continuous to ordinal. Further exploration into the dataset; duplicate column information, any text verbiage columns, outliers are identified that could be hidden and excluded. For handling missing data, a subset of data to be cleaned is created and later merged to the original file. A pivot table for the identified subset is analyzed and checked for any interesting patterns. Missing numeric data are replaced by mean**(need to confirm).**

During the Modify stage, each of the variables was assessed for understanding, correlation, and duplicity

continuous variables). A correlation matrix plot and descriptive statistics are run to understand each variable and its distribution across the dataset.

**(analyze the scope of feature engineering depending on the dataset)**

In the Model stage, After partitioning the dataset**(need to confirm the data partition)**, a GridSearchCV() on at least three (**or more if needed**) different ML algorithms are run to identify the best parameters to use for a model. Three model interpretability plots (**or more if needed**) such as partial dependence plot, variable importance plot and scatterplots from the best fitting model are shown to represent the variable importance graphically.

Lastly, in Analyze stage- Error metrics table(i.e. RMSE, MAE, standard deviation, MAPE, bias), variable importance plot of different algorithms are compared.

**Results**:

Using the Biobank Data search, a total of 7157 patients affected with a different type of sepsis-related records were studied. An external dataset (Cyber ICU or non ICU) was used to validate the training datasets.

The best model was chosen from the 3 models run based on comparing the model accuracy and error metrics. From this, some of the important variables affecting the lactate levels are identified.