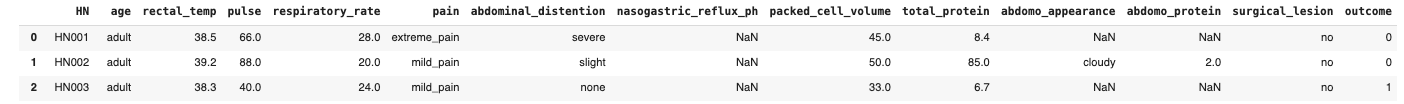
**Part 2) Machine Learning (10%) [Horse Survival Dataset]**

In this exam, we aim to implement a binary classification model to predict as to whether or not a horse can be survived (the variable “outcome”) either “died” (0) or “lived” (1); “lived” is the target level. Note that the data is obtained and modified from the Kaggle on “Horse Survival Dataset”. The sample data and the data dictionary are shown below.



Details of some variables are listed below:

* HN (horse number) is an index.
* There are 4 categorical variables: 'pain', 'abdominal\_distention', 'age', 'surgical\_lesion'.

You must submit the following items:

* File.ipynb – a source code.
  + It **cannot** be graded if we **cannot** map your answer to the question, so please add a comment to identify the questions, e.g., Q1, Q2, etc.
  + All answers must be obtained from your code and shown or written in this notebook file.
  + For the notebook file, the output must be shown **(DO NOT clear the output).**
* The file must be renamed as “{student\_id}\_{firstname}\_Part2\_ML”, e.g., 6030133421\_Chaiyatad\_Part2\_ML.ipynb

# Tasks

1. Preprocess the data.
   * Upload the data “horse\_2.csv” to your notebook and then load it to the variable “df”
   * Remove any patients with missing target (if any)
   * Drop variables that missing value > 50%
   * Q1) How many variables have been removed?
2. Impute missing values.
   * You must use command “SimpleImputer”
   * Impute numerical variables with mean and round (no decimal points)
   * Impute categorical variables with most\_frequent
   * You must show the imputed values of all missing variables.
   * Q2) What is the imputed value of the variable “pulse”?
3. Create a dummy code for the nominal variables ('pain', 'abdominal\_distention', 'age', 'surgical\_lesion') with dropping the first level. Please ensure that the source categorical variables should not be in the data frame anymore. You must use command “OneHotEncoder”.
   * You must show information (Hint: df.info()) of the table.
   * Q3) How many dummy variables (columns) do you get from this step?
4. Train/Test split
   * To prevent an error propagation to remaining questions. Use the provided cleaned data by loading “cleaned\_horse\_2.csv” to the variable “cleaned\_df”. Set the variable “HN” to be index.
   * Split train/test **with stratification** by setting testing size to be 30% and random seed to be 2024.
5. [Model1] Create a random forest with the parameters below.

* n\_estimators = 100
* mins\_samples\_leaf = 7
* max\_depth = 5
* random\_state = 2024
  + Q4) Which variable is the most important feature?

1. Evaluate the model on the testing data. Show a classification report with 4 digits and show a confusion matrix with an option label=[“died”, “lived”] (row and column are actual and prediction, respectively).
   * Q5) What is the macro F1?
   * Q6) What is the number of correctly classified as “lived”?
2. Use the model to predict a horse with “HN089”. What is the probability to be lived (round 4 decimal points)?
   * Q7) What is the probability that this horse will live?
3. [Model2] Perform MLflow on the random forest with the possible parameters as below. Each combination of those parameters must be logged along with the metric “accuracy” and the model. Search the winner model with the highest accuracy.

* n\_estimators = [100, 200, 300]
* mins\_samples\_leaf = [7, 10]
* max\_depth = [5, 7]
* random\_state = 2024
  + Q8) What are the parameters (n\_estimators, max\_depth, min\_samples\_leaf) of the winner?
  + Q9) What is the accuracy of the winner?

1. Load the winner model and evaluate the model on the testing data. Show a classification report with 4 digits.
   * Q10) What is the macro F1?

**Multiple choices**

Q1) How many variables that have been removed?

1. 1
2. 2
3. 3
4. 4

Q2) What is the imputed value of the variable “pulse”?

1. 72
2. 70
3. 38
4. 30

Q3) How many dummy variables (columns) do you get from this step?

1. 8
2. 9
3. 10
4. 11

Q4) Which variable has the highest weight?

1. abdominal\_distention\_none
2. surgical\_lesion\_yes
3. packed\_cell\_volume
4. pulse

Q5) What is the macro F1?

1. 0.7390
2. 0.7556
3. 0.7617
4. 0.7667

Q6) What is the number of correctly classified as “lived”?

1. 9
2. 14
3. 27
4. 40

Q7) What is the probability that this horse will live?

1. 0.8335
2. 0.1665
3. 0.8883
4. 0.1117

Q8) What are the parameters (n\_estimators, max\_depth, min\_samples\_leaf) of the winner?

1. n\_estimators=100, max\_depth=5, min\_samples\_leaf=10
2. n\_estimators=100, max\_depth=5, min\_samples\_leaf=5
3. n\_estimators=200, max\_depth=7, min\_samples\_leaf=5
4. n\_estimators=200, max\_depth=7, min\_samples\_leaf=10

Q9) What is the accuracy of the winner?

1. 0.7444
2. 0.7556
3. 0.7685
4. 0.7778

Q10) What is the macro F1?

1. 0.7444
2. 0.7556
3. 0.7685
4. 0.7778