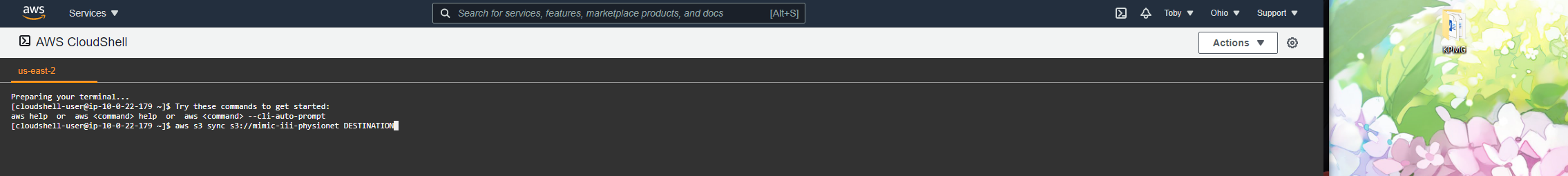
Graphical user interface, application, Teams

Description automatically generatedData on Google Cloud Platform

Running command on aws shell



Not enough cloud spaceText

Description automatically generated

Big query access sucessfully

Graphical user interface

Description automatically generated with medium confidence

Checked out the Big query tutorial

Tutorial: https://mimic.mit.edu/docs/iii/tutorials/intro-to-mimic-iii-bq/

Graphical user interface, text, application, email

Description automatically generated

Testing queriesGraphical user interface, application

Description automatically generated

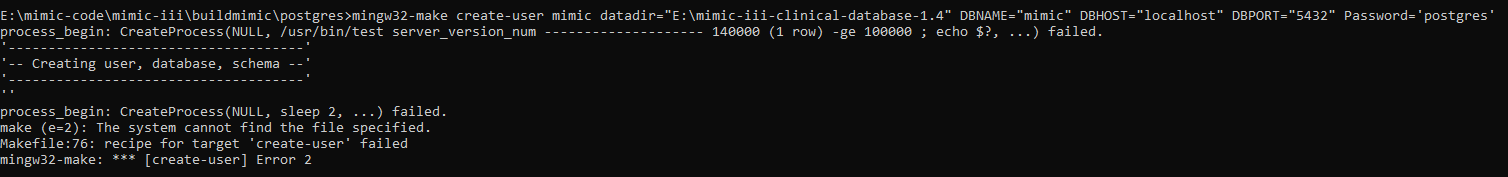
Questions:

How am does machine learning with the data work?

Do I have to arrange them into a numpy array? Add it to jupyter note book?

The database look quite diveresed and contains a lot of items, how would I know what to train for?

What is the aim of the project? Must I follow the project specification of can I improvise and change to my own will?

Mingw32 error

# Diagram Description automatically generatedPreprocessing Data

## Static features

A picture containing table

Description automatically generated

## Timeseries features

Diagram

Description automatically generated

Diagram

Description automatically generated

**Pre-processing for in-hospital mortality prediction**

In the previous part, we have extracted patients, events (mapped to variables), and outcomes (hospital mortality and length of stay on the ICU) data. Here, we will pre-process this data. The following pre-processing steps will need to be done:

1. Unit conversion: map events variables to the same metric.
2. Detect and remove or replace outliers in the event data.
3. Reorganize the events data so that we have a column for every variable, and a row for every hour and every ICU stay identifier.
4. Impute time series data.
5. Standardize the continuous event variables.
6. One-hot encode the categorical event variables.

**1. Unit conversion**

Variables can have different measurement units (for example: weight in oz, kg or lbs). We first need to map all variables to the same metric. The variable listed in the table below need to be mapped. The conversion indicates how to convert the variable, and the ‘check’ column indicates the check that must be true before conversion is needed.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Unit** | **Check** | **Conversion** |
| Weight | oz | - | x / 16.0 \* 0.45359237 |
| Weight | lbs | - | x\*0.45359237 |
| Fraction inspired oxygen | - | x > 1 | x/100.0 |
| Oxygen saturation | - | x <= 1 | x\*100.0 |
| Temperature | f | x > 79 | (x - 32) \* 5.0/9) |
| Height | in | - | x\*2.54 |

**2. Detect and remove/replace outliers**

We need to detect and remove or replace outliers. For this, two different types of outliers are used:

1. For each variable, there is an upper and lower threshold for detecting unusable outliers. If the outlier falls outside of the threshold, it is treated as missing (set to NaN).
2. There is also a physiologically valid range of measurements, if a non-outlier falls outside this range, it will be replaced with the nearest valid value.

The table below indicates the imputation values, along with the two different outlier ranges.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Impute value** | **Unusable outliers** | **Valid range** | **Modelled as** |
| Capillary refill rate | 0.0 | 0.0 – 1.0 | 0.0 – 1.0 | categorical |
| Diastolic blood pressure | 59.0 | 0.0 – 375.0 | 0.0 – 375.0 | continuous |
| Fraction inspired oxygen | 0.21 | 0.2 – 1.1 | 0.21 – 1.0 | continuous |
| Glascow coma scale eye opening | 4 | 1.0 – 4.0 | 1.0 – 4.0 | categorical |
| Glascow coma scale motor response | 6 | 1.0 – 6.0 | 1.0 – 6.0 | categorical |
| Glascow coma scale total | 11 | 3.0 – 15.0 | 3.0 – 15.0 | categorical |
| Glascow coma scale verbal response | 4 | 1.0 – 5.0 | 1.0 – 5.0 | categorical |
| Glucose | 128.0 | 0.0 – 2200.0 | 33.0 – 2000.0 | continuous |
| Heart rate | 86 | 0.0 – 390.0 | 0.0 – 350.0 | continuous |
| Height | 170.0 | 0.0 – 275.0 | 0.0 – 240.0 | continuous |
| Mean blood pressure | 77.0 | 0.0 – 375.0 | 14.0 – 330.0 | continuous |
| Oxygen saturation | 98.0 | 0.0 – 150.0 | 0.0 – 100.0 | continuous |
| Respiratory rate | 19 | 0.0 – 330.0 | 0.0 – 300.0 | continuous |
| Systolic blood pressure | 118.0 | 0.0 – 375.0 | 0.0 – 375.0 | continuous |
| Temperature | 37.0 | 14.2 – 47.0 | 26.0 – 45.0 | continuous |
| Weight | 81.8 | 0.0 – 250.0 | 0.0 – 250.0 | continuous |
| pH | 7.4 | 6.3 – 10.0 | 6.3 – 8.4 | continuous |

**3. Reorganize data**

We will reorganize the events data in such a way that we have a column for every variable, and a row for every hour and every ICU stay identifier. Hence, we will need to group the variable values from the same hours. We summarize this by taking the mean, median, standard deviation and count across all events that need to be grouped together.

For a single ICU stay and only two variables, an example of the structure is shown below. Moreover, we need to make sure that there is a row for every hour until the patient is discharged from the ICU. The Nan values in the ‘count’ columns can be replaced with zeroes, indicating that these variables were not measured for this ICU stay during the specific hour.

Table

Description automatically generated

**4. Imputation of time series**

If a variable was not measured during a specific hour for a specific ICU stay, the mean and median values will be NaN. We can forward fill these values, meaning that we replace them by the most recent measurement. If there are no measurements for a specific variable at all, there are two situations possible:

1. If the variable was measured for this subject, but only after the current hour, we replace the NaN value with the average or median for this specific subject/ICU stay/hospital admission combination.
2. If the variable was never measured for this subject, we replace the NaN value with the global average or global median.

**5. Standardization of continuous variables**

We need to standardize the continuous variables. The table under step 2 lists all the variables that we have extracted, and whether they will be modelled as a continuous or categorical variable.

**5. One-hot encoding the categorical variables**

The categorical variables (see table step 2) need to be one-hot encoded. This means that we will get extra columns in our dataset, namely one for each category for each categorical variable.

Running the above steps might take approximately 1-2 hours, so it is advised to save the data once it has been pre-processed.

**Patients extraction of MIMIC-III**

This part describes the data extraction and pre-processing steps of the MIMIC-III data for in-hospital mortality prediction. This is a binary classification task to predict in-hospital mortality based on the first 48 hours of an ICU stay. First, the patients and their corresponding (time-series) variables need to be extracted.

**1. Study cohort selection**

We will exclude patients who are younger than 15 years old and exclude ICU stays that took less than a day or more than 10 days. Additionally, we exclude patients who are on the neonatal intensive care unit (NICU). The study cohort selection procedure is visualized below.

Diagram

Description automatically generated

For each unique ICU stay identifier, we are interested in the patient’s corresponding hospital admission identifier, subject identifier, time of admission to the hospital and ICU, time of discharge of the hospital and ICU, death time (if died in hospital) and date of death, and length of stay. These variables will be needed later on during the prediction task. We will need to join the *Patients*, *Icustays* and *Admissions* tables. We compute a patient’s age from the ICU intime and patient’s date of birth, and filter out any patients younger than 15 years old. We also remove readmissions. The SQL query diagram is shown below.

Diagram

Description automatically generated

**2. Extraction of vital data and mapping to variables**

We will extract charted events and lab events that we will afterwards map to 17 different variables. The table below lists the charted and lab events that we want to filter on and their corresponding variable.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Chartevents item ID** | **Labevents item ID** |
| Capillary refill rate | 3348, 115, 8377 | - |
| Diastolic blood pressure | 8368, 220051, 225310, 8555, 8441, 220180, 8502, 8440, 8503, 8504, 8507, 8506, 224643 | - |
| Fraction inspired oxygen | 3420, 223835, 3422, 189, 727, 190, 7570 | - |
| Glasgow coma scale eye opening | 184, 220739 | - |
| Glasgow coma scale motor response | 454, 223901 | - |
| Glasgow coma scale total | 198 | - |
| Glasgow coma scale verbal response | 723, 223900 | - |
| Glucose | 807, 811, 1529, 3745, 225664, 220621, 226537 | 50931, 50809, 51478 |
| Heart rate | 211, 220045 | - |
| Height | 226707, 226730, 1394 | - |
| Mean blood pressure | 52, 220052, 225312, 224, 6702, 224322, 456, 220181, 3312, 3314, 3316, 3322, 3320 | - |
| Oxygen saturation | 834, 8498, 220227, 646, 220277 | 50817 |
| Respiratory rate | 618, 220210, 3603, 224689, 614, 651, 224422, 615, 224690 | - |
| Systolic blood pressure | 51, 220050, 225309, 6701, 455, 220179, 3313, 3315, 442, 3317, 3323, 3321, 224167, 227243 | - |
| Temperature | 3655, 677, 676, 223762, 3654, 678, 223761 | - |
| Weight | 763, 224639, 226512, 3580, 3693, 3581, 226531, 3582, 581, 3723, 3583, 3692, 580, 733, 225124, 4183, 226846 | - |
| pH | 3839, 1673, 780, 1126, 223830, 4753, 4202, 860, 220274 | 50820, 50831 |

We need to extract the events corresponding to the item IDs that are listed in the table above. We would also like to get the unique subject, hospital admission and ICU stay identifiers. Moreover, we need the time of registering the event, the value, and the measurement unit (‘valueuom’). To get all this information, we first get the correct information from the *Chartevents* and *Labevents* tables separately. We also join the *Icustays* table with both of these tables to filter on the ICU stay IDs that were selected previously (step 1). Afterwards, we can concatenate the rows of the two results to get all events needed to extract the variables above.

Diagram

Description automatically generated

Once we have the events, we can also get a dictionary, indicating a short description of each item ID. We can use this to check once more if we have selected the right events. The descriptions can be found in the *D\_items* table, and we filter on the item IDs selected in the previous query. The query diagram can be found below. Afterwards, we can extract the variables using the table above.

Diagram

Description automatically generated