UNIVERSITI MALAYA

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

SEMESTER 1 2024/2025

WQD7005 DATA MINING (OCC 1)

ASSIGNMENT

LECTURER: PROF DR TEH YING WAH

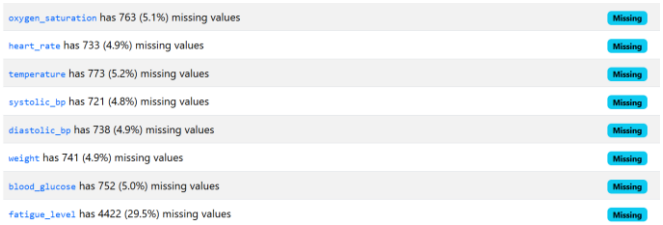
NAME: YEE SEE MARN

MATRIC NUMBER: 23102510

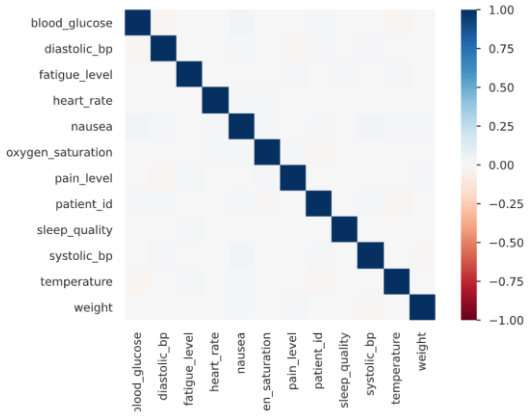
AI-Assisted Summary Report and Visualization

Key findings before pre-processing:

- According to the YData Profiling Report, there are missing values for variables oxygen_saturation (5.1%), heart_rate (4.9%), temperature (5.2%), systolic_bp (4.8%), diastolic_bp (4.9%), weight (4.9%), blood_glucose (5.0%) and fatigue_level (29.5%).



- According to the statistical summaries for each variable,
    - oxygen_saturation: Mean is 97%, with a relatively small standard deviation which is 1.50, suggesting generally good oxygen levels. Range is from 91.4 to 102.3.
    - heart_rate: Mean is 74.46 bpm, with a standard deviation of 9.95. Range is from 37 to 113.
    - temperature: Mean is 36.8 degrees, with a small standard deviation. Range is from 35 to 38.3.
    - systolic_bp: Mean is 119.47 mmHg, with a standard deviation of 9.98. Range is from 79 to 159.
    - diastolic_bp: Mean is 79.43 mmHg, with a standard deviation of 6.98. Range is from 51 to 106.
    - weight: Mean is 69.95, with a standard deviation of 14.94. Range is from 12 to 122.8.
    - blood glucose: Mean is 99.56, with a standard deviation of 19.87. Range is from 12 to 173.
- There are some outliers or extreme values for oxygen_saturation, heart_rate, temperature, systolic_bp, diastolic_bp, weight, blood_glucose that need further investigation.
- The heatmap show that the correlation between the variables is insignificant.



- “Sleep”, “of”, “feeling” are some of the frequent words for clinical note columns but does not show any informative insights. Sentiment analysis can be conducted later to obtain more valuable insights of patients.

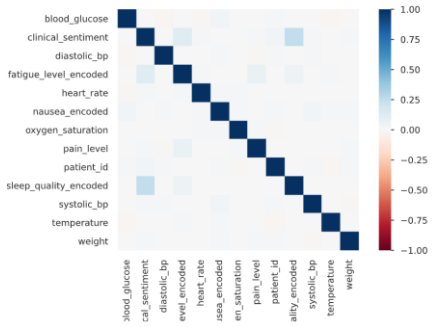
Steps of pre-processing:

- For missing values of vital signs variables, mean imputation is implemented and normalize by using min-max scaler.
- For missing values of categorical variables, LLMs model which is flan-t5-base is implemented to impute textual data. It is a fine-tuned version of the T5 (Text-to-Text Transfer Transformer) model that developed by google. After the imputation, the categorical variables undergo encode by using LabelEncoder in scikit-learn.
- There are no missing values for clinical note columns, no imputation is needed. Sentiment analysis pipeline is implemented through distilbert-base-uncased-finetuned-sst-2-english model, which is a SLMs. The results show that there are 8403 positive and 6597 negative sentiments.

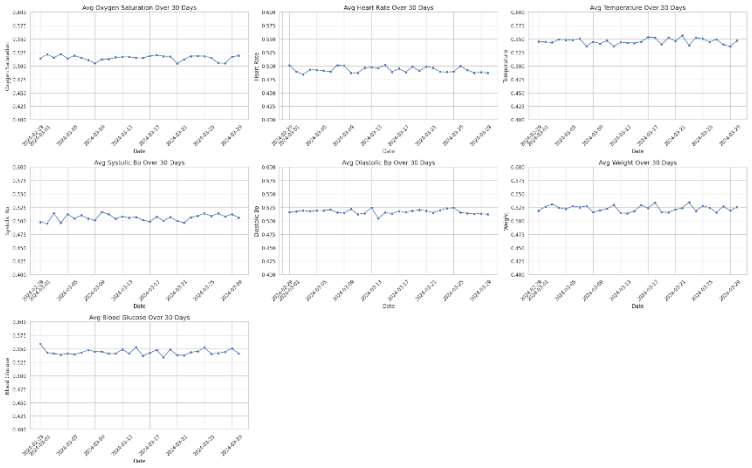
Key findings after pre-processing:

- No missing values.
- Vital signs variables are normalized and scaled consistently, categorical variables are encoded, and clinical notes undergo sentiment analysis and able to investigate correlations with numerical variables.
- From the heatmap,

- Slightly positive correlation between sleep_quality and fatigue_level and is possible clinically as worse sleep may increase fatigue.
- Moderate positive correlation between sleep_quality and clinical_sentiment and this show reasonable since the most frequent word in the clinical notes is sleep.
- Slight positive correlation between fatigue_level and clinical_sentiment.



- From the time series graph,
  - Most of the vitals variables are stable on average, with a few days showing significant spikes. Thus, it is important to identify what caused those happen and further investigation are needed.



- By leveraging LLMs, some suggestions are given for further analysis,
  - More details sentiment analysis of clinical notes such as using NLP techniques to extract important features.
  - Implement clustering techniques to group patients with similar characteristics and identify potential risk factors.