SMU
SINGAPORE MANAGEMENT UNIVERSITY

School of Computing and Information Systems

Healthcare Analytics Group Project

Updated 12 July 2024

Real-world Dataset

SingHealth Electronic Medical Records which contains

- Data from 2012-2016 from Singapore General Hospital with 90,785 patients
- All surgery patients except for Cardiac and Neurosurgery patients

Objectives:

1. Evaluate the Risk Factors
2. Development Predictive Models

Outcomes:

1. Mortality within 30 days
2. Admission to ICU within 24 hours
3. Days from surgery to Mortality (for patients with recorded death)

Chan DXH, Sim YE, Chan YH, Poopalalingam R, Abdullah HR. Development of the Combined Assessment of Risk Encountered in Surgery (CARES) surgical risk calculator for prediction of postsurgical mortality and need for intensive care unit admission risk: a single-center retrospective study. *BMJ Open*. 2018;8(3):e019427. doi:10.1136/bmjopen-2017-019427

Project Aims

1. Risk factor analysis
   - utilize the dataset to evaluate the risk factors associated with the outcomes
2. Predictive modelling
   - develop validated predictive models for the prediction of various outcomes

Project Tasks

(1) Data wrangling and preparation. Examples of competencies in this area include:

- Missing data imputation methods (try different imputation methods and compare the robustness of results
- Data quality assessment (visualize and summarize the data used in the models
- Dealing with bias, imbalanced data, etc(improve on them if possible)

(2) Risk Factor Analysis. Examples of competencies in this area include:

- Univariate and multivariate analysis
- Model quality evaluation
- Model improvement
- Model analysis and interpretation of results

(3) Predictive Model Development. Examples of competencies in this area include:

- Test out different models (machine learning/statistical models)
- Test out different validation metrics (include the most common ones)
- Hyperparameter tuning (find the best hyperparameters)
- Improve on the Training/Validation/Testing procedures (e.g., dealing with missing values, outliers, imbalanced data)

Project Tasks

- **Submit a presentation of 30 Slides** (additional information can be put into the Annex)
- **Submit all codes and presentation through eLearn**
- **You do not need to do the presentation,** hence, ensure that your slides contain all the relevant information.

Note:

- Always ensure that the information is available at the time of prediction
- Rigour and accuracy of the model prediction and validation processes
- Comparison across different models is mandatory

This is a real-world project with two real world dataset. Hence, some of the features may not have been fully tested. This will present unique opportunities to deal with novel problems that were unanticipated previously.

SGH Surgery Open Data

- **No codes are provided**
- Read the assumptions of the data discussed in Chan et al., 2018.
- Read and understand what have been done
  - You may use different models and algorithms
  - You may introduce more features. Explain the rationale of the features that you introduce
  - Improve on the training/validation/test processes where needed

Chan DXH, Sim YE, Chan YH, Poopalalingam R, Abdullah HR. Development of the Combined Assessment of Risk Encountered in Surgery (CARES) surgical risk calculator for prediction of postsurgical mortality and need for intensive care unit admission risk: a single-center retrospective study. *BMJ Open*. 2018;8(3):e019427. doi:10.1136/bmjopen-2017-019427

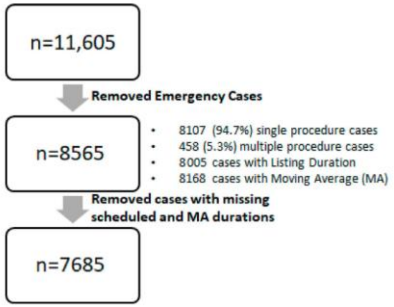
Project Guidance

- This is not a Kaggle Competition nor a coding competition (you can use other software if they are more effective
- Scores will not be based solely on prediction accuracy
- Document the assumptions and limitations
- Focus on the concepts to improve the analysis
- Coding style will NOT be evaluated. But the codes have to run to reproduce the results
- **No presentation is needed**. A **presentation slide deck** has to be submitted for this project.

Presentation Structure

Maximum slide limit (excluding Annex): 30 slides
Submit in **PPT and PDF Format**!

1. Title/ Abstract
2. Introduction/ Background
   - Literature Review
   - Data understanding
   - Project Management
3. Data and Methods:
   - Data Wrangling
   - Feature Engineering
   - Descriptive Statistics
   - Model Development and Validation
4. Results and Discussion
   - Data visualization (relevant ones)
   - Comparison of models
   - Limitations and Assumptions
   - Future work
5. Conclusion
6. References
7. Appendix



n=11,605

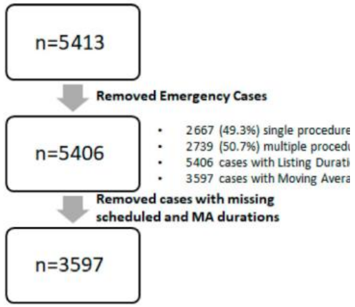
Removed Emergency Cases

n=8565

- 8107 (94.7%) single procedure cases
- 458 (5.3%) multiple procedure cases
- 8005 cases with Listing Duration
- 8168 cases with Moving Average (MA)

Removed cases with missing scheduled and MA durations

n=7685



n=5413

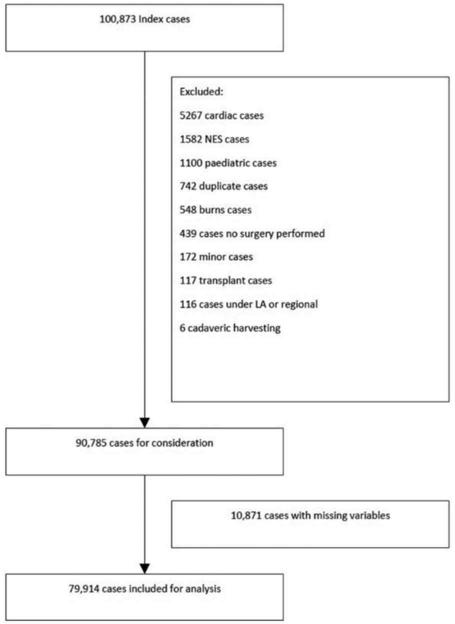
Removed Emergency Cases

n=5406

- 2667 (49.3%) single procedure
- 2739 (50.7%) multiple procedu
- 5406 cases with Listing Durati
- 3597 cases with Moving Avera

Removed cases with missing scheduled and MA durations

n=3597



100,873 Index cases

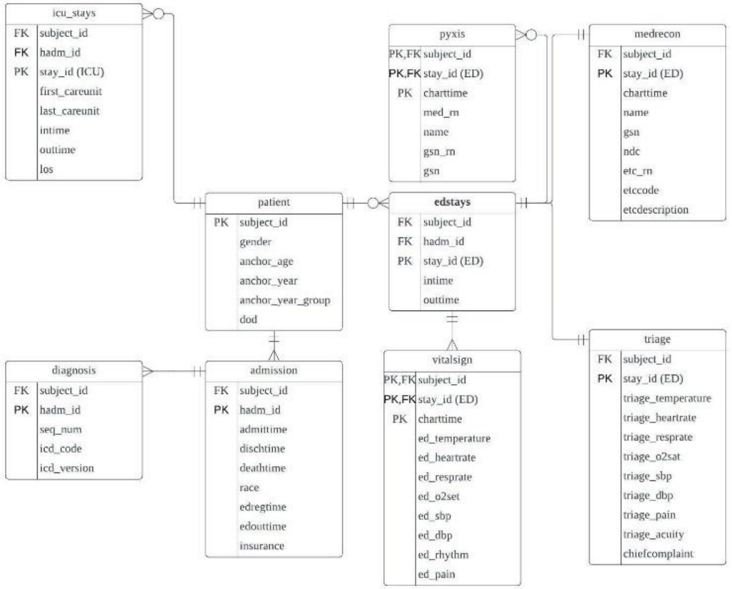
Excluded:

5267 cardiac cases
1582 NES cases
1100 paediatric cases
742 duplicate cases
548 burns cases
439 cases no surgery performed
172 minor cases
117 transplant cases
116 cases under LA or regional
6 cadaveric harvesting

90,785 cases for consideration

10,871 cases with missing variables

79,914 cases included for analysis



	icu_stays
FK	subject_id
FK	hadm_id
PK	stay_id (ICU)
	first_careunit
	last_careunit
	intime
	outtime
	los

	pyxis
PK,FK	subject_id
PK,FK	stay_id (ED)
PK	charttime
	med_rn
	name
	gsn_rn
	gsn

	medrecon
FK	subject_id
PK	stay_id (ED)
	charttime
	name
	gsn
	ndc
	etc_rn
	etccode
	etcdescription

	patient
PK	subject_id
	gender
	anchor_age
	anchor_year
	anchor_year_group
	dod

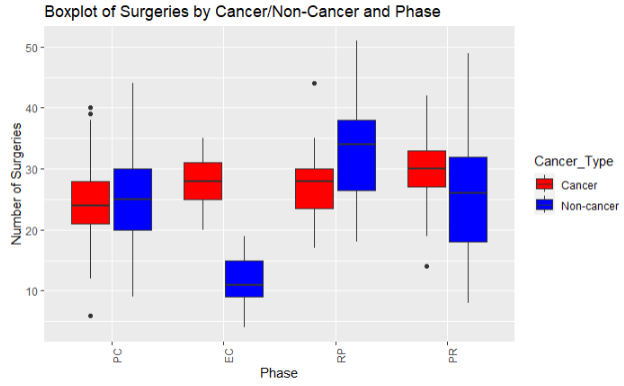
	edstays
FK	subject_id
FK	hadm_id
PK	stay_id (ED)
	intime
	outtime

	diagnosis
FK	subject_id
PK	hadm_id
	seq_num
	icd_code
	icd_version

	admission
FK	subject_id
PK	hadm_id
	admittime
	dischtime
	deathtime
	race
	edregtime
	edouttime
	insurance

	vitalsign
PK,FK	subject_id
PK,FK	stay_id (ED)
PK	charttime
	ed_temperature
	ed_heartrate
	ed_resprate
	ed_o2set
	ed_sbp
	ed_dbp
	ed_rhythm
	ed_pain

	triage
FK	subject_id
PK	stay_id (ED)
	triage_temperature
	triage_heartrate
	triage_resprate
	triage_o2sat
	triage_sbp
	triage_dbp
	triage_pain
	triage_acuity
	chiefcomplaint



Groupings

- Split into groups of 7-8 students each
- There should be 4 groups
- Identify a project manager/asst manager whose core task is to manage the projects and keep everyone on track and organize the work, documentation and codes
- Parallelize your tasks as much as possible.
- LEARN AS YOU DO!
- Use collaborative tools where needed (GitHub, GoogleDocs, etc).
- Data science is a TEAM EFFORT! Have fun!

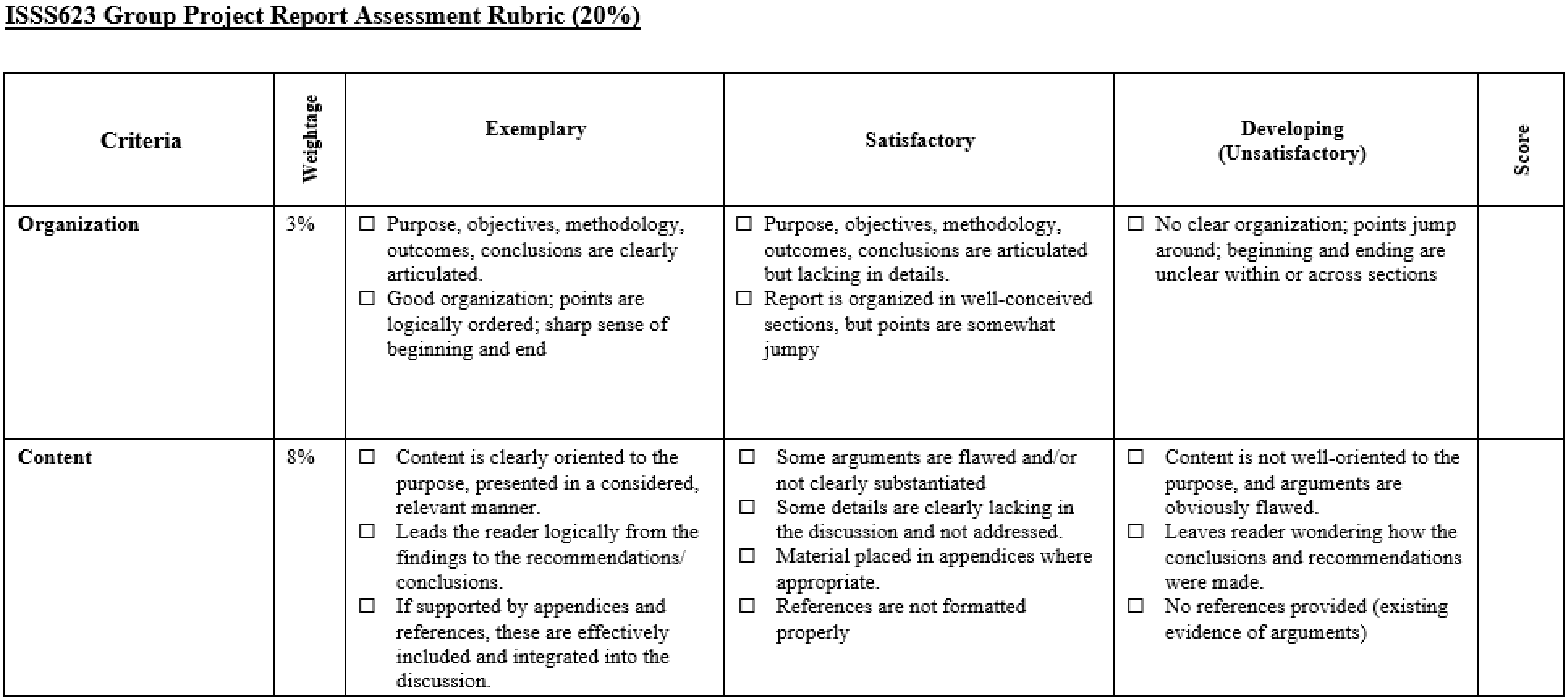
Project Assessment

1. Group Project – 40% of overall
   - **Group analytics project - 30%**
     - Presentation Slides - 20% (see report marking rubric)
     - Novelty of solution - 6%
       - Evidence of research and literature references to substantiate your proposed solutions
       - Effort level devoted to complete the group project
     - Runnable codes; completeness of submission - 2%
     - Timeliness of submission – 2%
   - **Peer Evaluation – 10%**

2. Due dates:
   - **17 August 2024 (2330 HRS)**

3. Submission of collaterals should be through eLearn, only 1 final submission per group.
   - Do not use the Course GitHub repo for your interim commits and collaborations

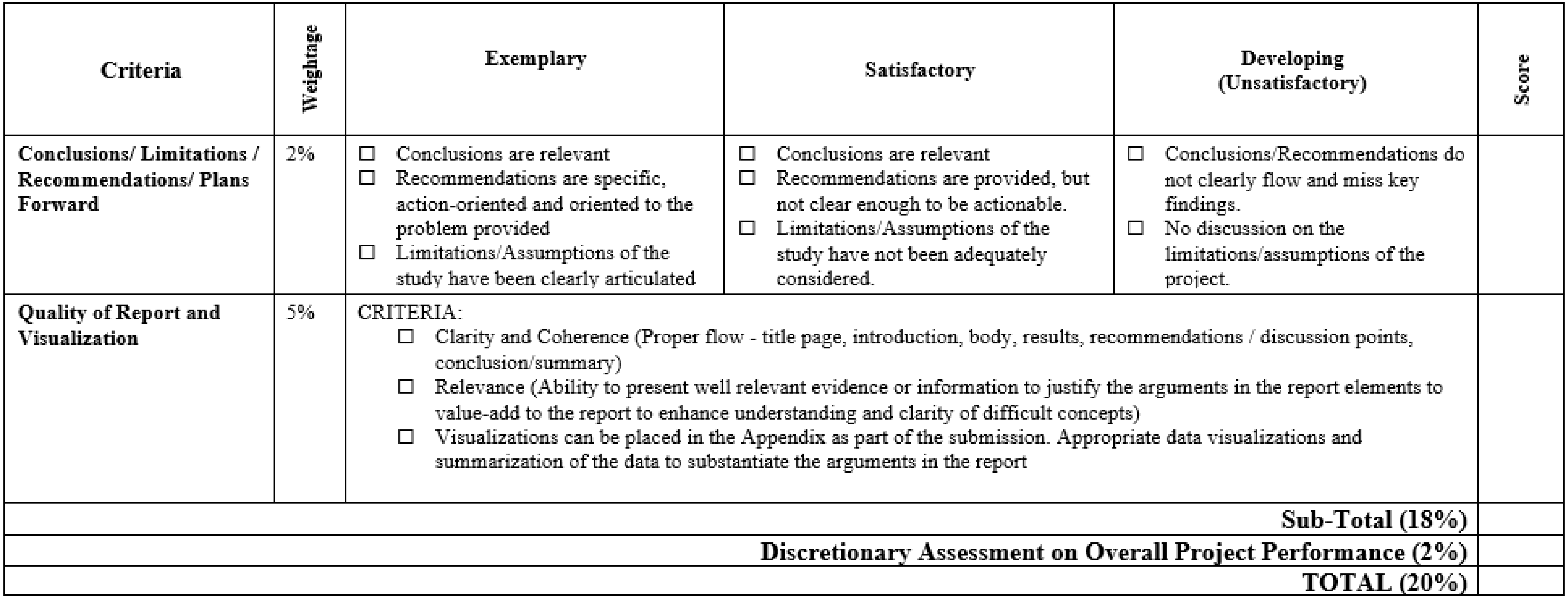
Report (Slides) Rubric (20%)



ISSS623 Group Project Report Assessment Rubric (20%)

Criteria	Weightage	Exemplary	Satisfactory	Developing (Unsatisfactory)	Score
Organization	3%	□ Purpose, objectives, methodology, outcomes, conclusions are clearly articulated. □ Good organization; points are logically ordered; sharp sense of beginning and end	□ Purpose, objectives, methodology, outcomes, conclusions are articulated but lacking in details. □ Report is organized in well-conceived sections, but points are somewhat jumpy	□ No clear organization; points jump around; beginning and ending are unclear within or across sections	
Content	8%	□ Content is clearly oriented to the purpose, presented in a considered, relevant manner. □ Leads the reader logically from the findings to the recommendations/ conclusions. □ If supported by appendices and references, these are effectively included and integrated into the discussion.	□ Some arguments are flawed and/or not clearly substantiated □ Some details are clearly lacking in the discussion and not addressed. □ Material placed in appendices where appropriate. □ References are not formatted properly	□ Content is not well-oriented to the purpose, and arguments are obviously flawed. □ Leaves reader wondering how the conclusions and recommendations were made. □ No references provided (existing evidence of arguments)	

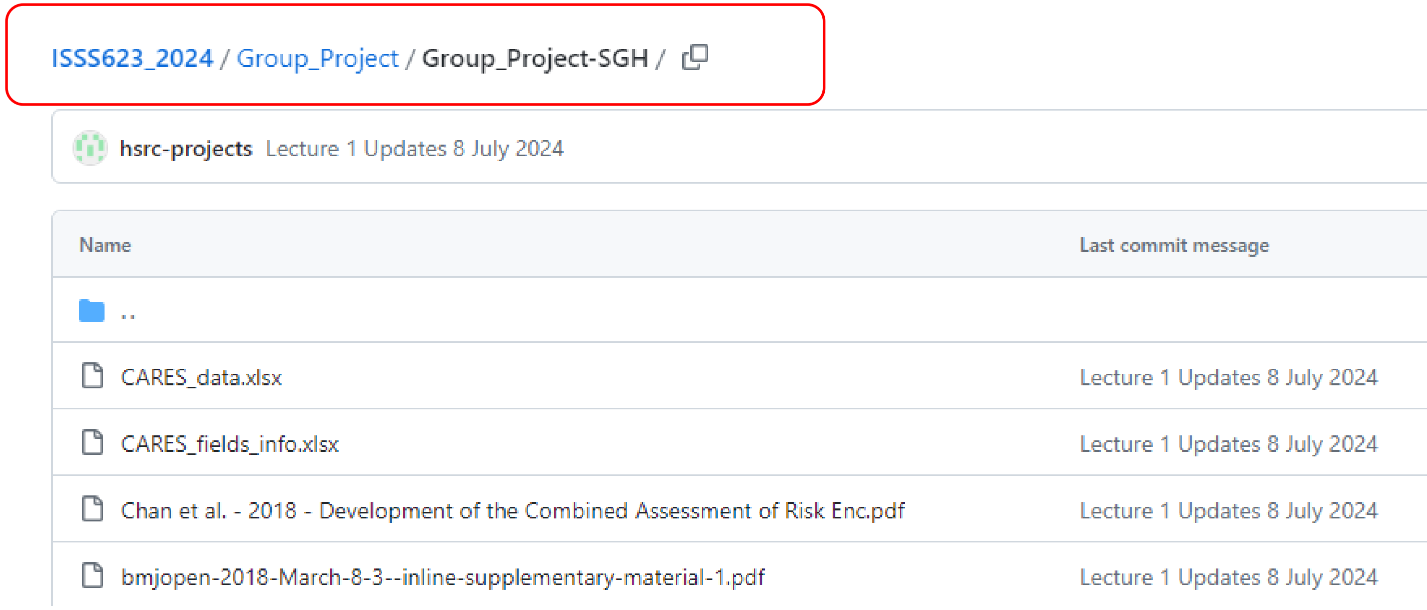
Report (Slides) Rubric (20%)



Criteria	Weightage	Exemplary	Satisfactory	Developing (Unsatisfactory)	Score
Conclusions/ Limitations / Recommendations/ Plans Forward	2%	□ Conclusions are relevant □ Recommendations are specific, action-oriented and oriented to the problem provided □ Limitations/Assumptions of the study have been clearly articulated	□ Conclusions are relevant □ Recommendations are provided, but not clear enough to be actionable. □ Limitations/Assumptions of the study have not been adequately considered.	□ Conclusions/Recommendations do not clearly flow and miss key findings. □ No discussion on the limitations/assumptions of the project.	
Quality of Report and Visualization	5%	CRITERIA: □ Clarity and Coherence (Proper flow - title page, introduction, body, results, recommendations / discussion points, conclusion/summary) □ Relevance (Ability to present well relevant evidence or information to justify the arguments in the report elements to value-add to the report to enhance understanding and clarity of difficult concepts) □ Visualizations can be placed in the Appendix as part of the submission. Appropriate data visualizations and summarization of the data to substantiate the arguments in the report			
Sub-Total (18%)					
Discretionary Assessment on Overall Project Performance (2%)					
TOTAL (20%)					

Group Project Repository

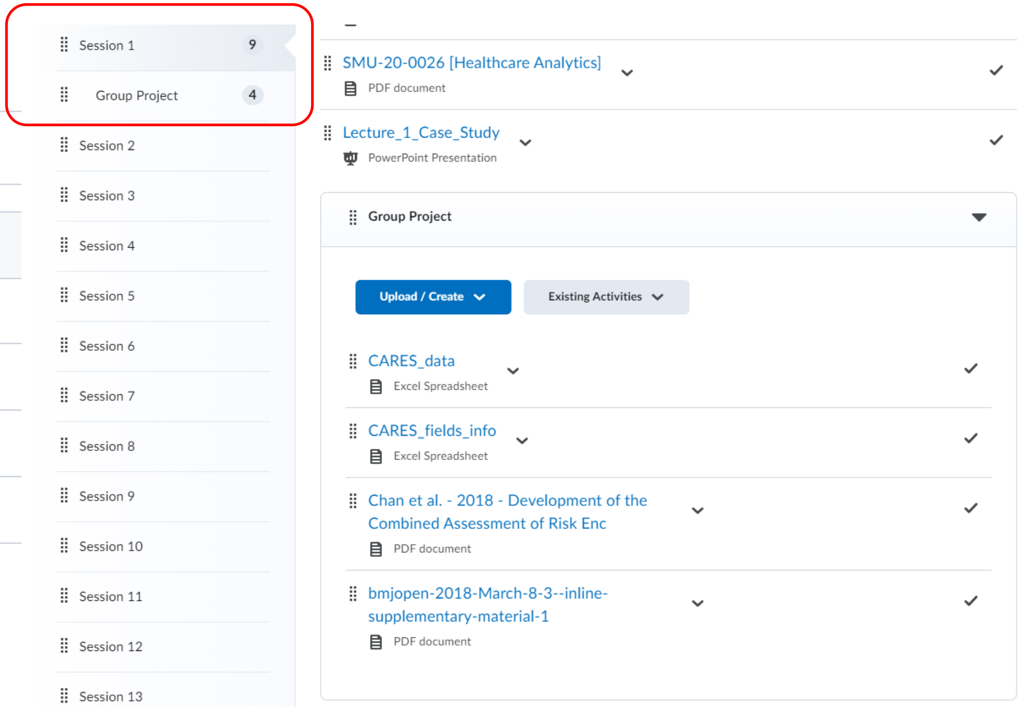
- Data and dictionary are downloadable here:
  - eLearn -> Content -> Session 1 -> Group Project
  - Github Repo: https://github.com/ISSS623-AHA/ISSS623_2024/tree/main/Group_Project/Group_Project-SGH



ISSS623_2024 / Group_Project / Group_Project-SGH /

hsrc-projects Lecture 1 Updates 8 July 2024

Name	Last commit message
..	
CARES_data.xlsx	Lecture 1 Updates 8 July 2024
CARES_fields_info.xlsx	Lecture 1 Updates 8 July 2024
Chan et al. - 2018 - Development of the Combined Assessment of Risk Enc.pdf	Lecture 1 Updates 8 July 2024
bmjopen-2018-March-8-3--inline-supplementary-material-1.pdf	Lecture 1 Updates 8 July 2024



Session 1 9
Group Project 4
Session 2
Session 3
Session 4
Session 5
Session 6
Session 7
Session 8
Session 9
Session 10
Session 11
Session 12
Session 13

SMU-20-0026 [Healthcare Analytics]
PDF document

Lecture_1_Case_Study
PowerPoint Presentation

Group Project

CARES_data
Excel Spreadsheet

CARES_fields_info
Excel Spreadsheet

Chan et al. - 2018 - Development of the Combined Assessment of Risk Enc
PDF document

bmjopen-2018-March-8-3--inline-supplementary-material-1
PDF document

References

- Data: **Chan DXH, Sim YE, Chan YH, Poopalalingam R, Abdullah HR.** Data from: Development of the Combined Assessment of Risk Encountered in Surgery (CARES) surgical risk calculator for prediction of post-surgical mortality and need for intensive care unit admission risk – a single-centre retrospective study. Published online February 9, 2018:13437289 bytes. doi:10.5061/DRYAD.V142481

- Prior study: **Chan DXH, Sim YE, Chan YH, Poopalalingam R, Abdullah HR.** Development of the Combined Assessment of Risk Encountered in Surgery (CARES) surgical risk calculator for prediction of postsurgical mortality and need for intensive care unit admission risk: a single-center retrospective study. *BMJ Open*. 2018;8(3):e019427. doi:10.1136/bmjopen-2017-019427

Sample Analysis

For Self-directed Learning

Real-world Dataset (Self-directed Learning)

Self-directed Learning

- Sample codes for model training and feature engineering are given for the MIMIC-IV dataset
- Student can understand at the sample codes as a Group
- This may help in the Group Project

Medical Information Mart for Intensive Care IV **(MIMIC-IV-ED)** database is a publicly accessible clinical database

- Data from 2018-2019 of deidentified health-related data from patients who were admitted to the critical care units of the Beth Israel Deaconess Medical Center (BIDMC)
- Comprehensive information for hospital stays of over 60,000 patients

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2022). MIMIC-IV (version 2.0). *PhysioNet*. https://doi.org/10.13026/7vcr-e114.
https://mimic.mit.edu/docs/about/

Real-world Dataset (Self-directed Learning)

Objective of MIMIC-IV Dataset

- Develop models for predicting the following outcomes:
  1. Hospitalization;
  2. ED readmission within 72 hours, 30 days;
  3. Critical Outcomes (ICU transfers within 12 h or mortality)
- Initial codes are provided. Take this as a learning opportunity based on these existing codes
- Read the assumptions in https://mimic.mit.edu/docs/iv/modules/
- Read and understand what have been done to generate additional features in “master_dataset_2.csv”

MIMIC-IV Modules Used

- Modules that are useful:
  - **ed** - data from the emergency department
  - **core** - provide demographics for the patient, a record for each hospitalization, and a record for each ward stay within a hospitalization. Contains three tables: patients, admissions, and transfers.
  - **icu** - ICU level data. These are the event tables, and are identical in structure to MIMIC-III (chartevents, etc)
  - **hosp** - contains data derived from the hospital wide EHR. These measurements are predominantly recorded during the hospital stay, though some tables include data from outside the hospital as well (e.g. outpatient laboratory tests in labevents).

Note: The core module has been merged with the hosp module to simplify the schema in MIMIC-IV v2.0 was released on June 12, 2022. The admissions, patients, and transfers tables are now in the hosp module.

ED Outcome Predictions – Master Dataset

- Tables used to derive the Master Dataset (“master_dataset_2.csv”):
  - CORE module- patient, admission
  - ED module - edstays, medrecon, pyxis, triage, vitalsign
  - HOSP module – diagnosis_icd
  - ICU module - icustays
- Outcomes to Predict:
  - Hospitalization – Inpatient admission following ED admission
  - Inpatient mortality - Either death or ICU within 12 hours
  - ED 72 hours reattendance – Patient return to ED withn 72 hours

Key linkage parameters across all tables:

- **Subject_id** - unique identifier which specifies an individual patient.
- **Stay_id** – ed table identifier which uniquely identifies a single emergency department stay for a single patient. Stay_id in each module may be different
- **Hadm_id** - If the patient was admitted to the hospital after their ED stay, hadm_id will contain the hospital identifier .
  - The hadm_id may be used to link the ED stay with the hospitalization in MIMIC-IV.
  - If hadm_id is NULL, the patient was not admitted to the hospital after their ED stay

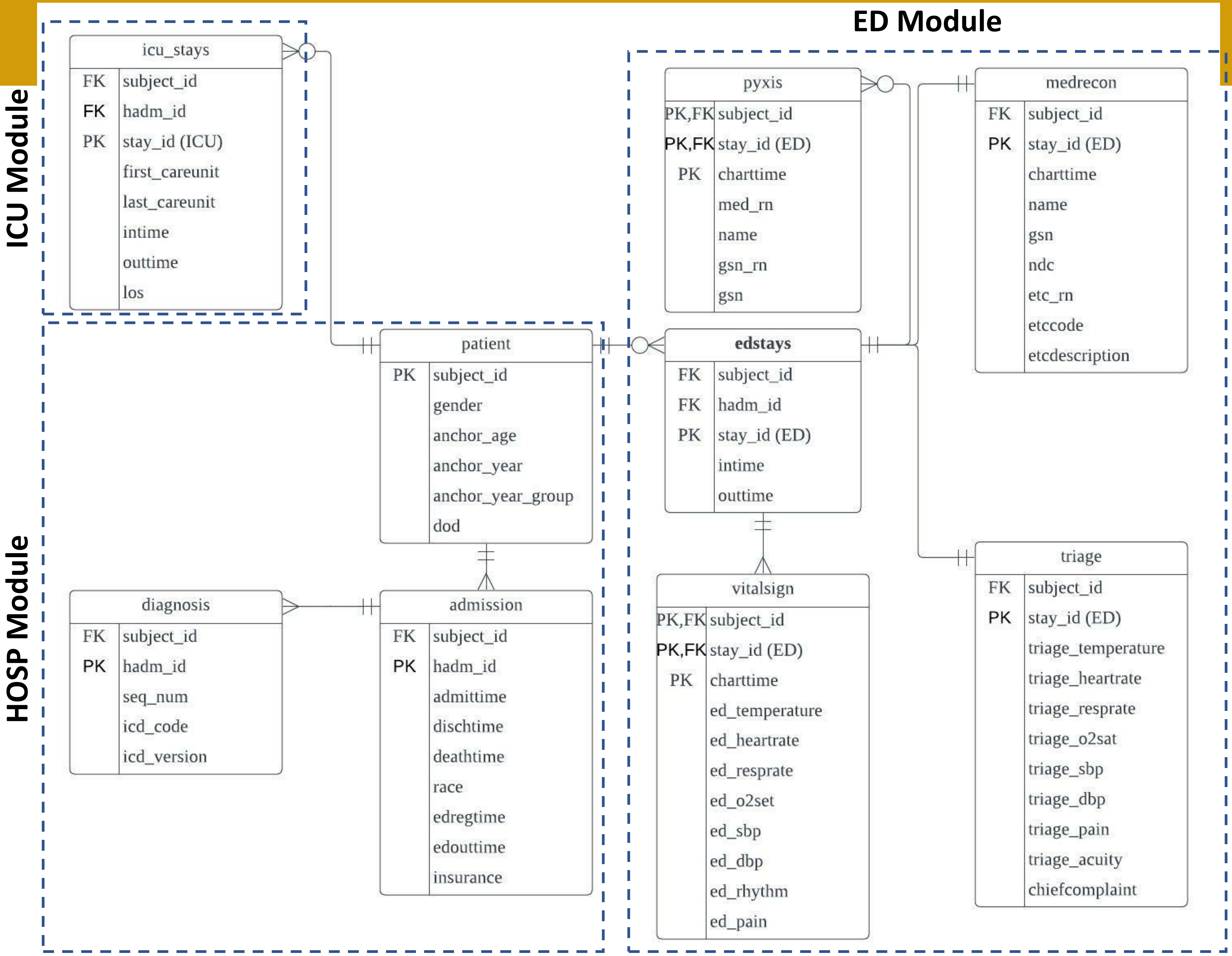
Reference: https://arxiv.org/ftp/arxiv/papers/2111/2111.11017.pdf

Model Comparison

- Comparators for existing scores (the codes to generate these scores are provided in the GitHub):
  - NEWS:
    - Royal College of P. National early warning score (NEWS) 2. Standardising the assessment of acute-illness severity in the NHS. 2017
    - https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2
    - Neurological scores (AVPU) not available (may be inaccurate)
  - MEWS:
    - Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified early warning score in medical admissions. QJM. 2001;94(10):521-526.
    - (https://academic.oup.com/qjmed/article/94/10/521/1558977?login=true)
    - Neurological scores (AVPU) not available (may be inaccurate)
  - REMS:
    - Olsson T, Terent A, Lind L. Rapid Emergency Medicine score: a new prognostic tool for in-hospital mortality in nonsurgical emergency department patients. J Intern Med. 2004;255(5):579-587.
    - https://www.emergencymed.org.il/wp-content/uploads/2019/10/Rapid-Emergency-Medicine-score_-a-new-prognostic-tool-for-in-hospital-mortality-in-nonsurgical-emergency-department-patients.pdf
    - Neurological scores (AVPU) not available (may be inaccurate)

- Machine Learning algorithms in the sample codes:
  - Logistic regression
  - Random Forest
  - Gradient Boosting
  - Multilayer Perceptron

Database Schema for MIMIC-IV Tables Used



ICU Module

icu_stays	
FK	subject_id
FK	hadm_id
PK	stay_id (ICU)
	first_careunit
	last_careunit
	intime
	outtime
	los

HOSP Module

patient	
PK	subject_id
	gender
	anchor_age
	anchor_year
	anchor_year_group
	dod

diagnosis	
FK	subject_id
PK	hadm_id
	seq_num
	icd_code
	icd_version

admission	
FK	subject_id
PK	hadm_id
	admittime
	dischtime
	deathtime
	race
	edregtime
	edouttime
	insurance

ED Module

pyxis	
PK,FK	subject_id
PK,FK	stay_id (ED)
PK	charttime
	med_rn
	name
	gsn_rn
	gsn

medrecon	
FK	subject_id
PK	stay_id (ED)
	charttime
	name
	gsn
	ndc
	etc_rn
	etccode
	etcdescription

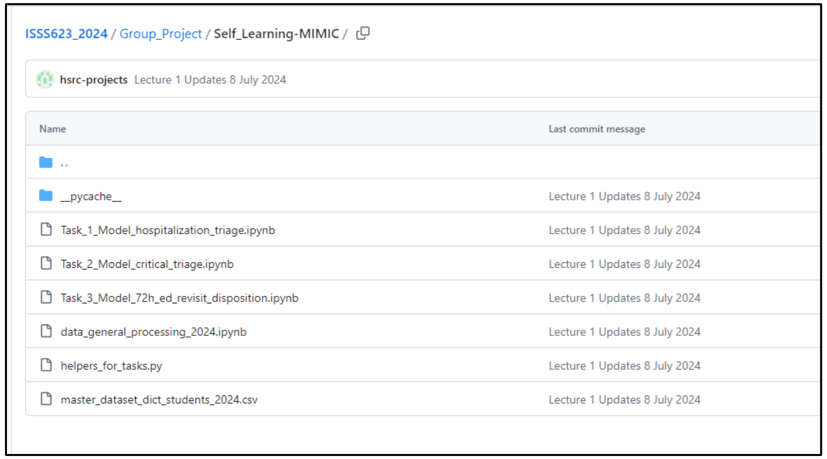
edstays	
FK	subject_id
FK	hadm_id
PK	stay_id (ED)
	intime
	outtime

vitalsign	
PK,FK	subject_id
PK,FK	stay_id (ED)
PK	charttime
	ed_temperature
	ed_heartrate
	ed_resprate
	ed_o2set
	ed_sbp
	ed_dbp
	ed_rhythm
	ed_pain

triage	
FK	subject_id
PK	stay_id (ED)
	triage_temperature
	triage_heartrate
	triage_resprate
	triage_o2sat
	triage_sbp
	triage_dbp
	triage_pain
	triage_acuity
	chiefcomplaint

Sample Project Repository

- Data and dictionary are downloadable here:
  - Github Repo: https://github.com/ISSS623-AHA/ISSS623_2024/tree/main/Group_Project/Self_Learning-MIMIC

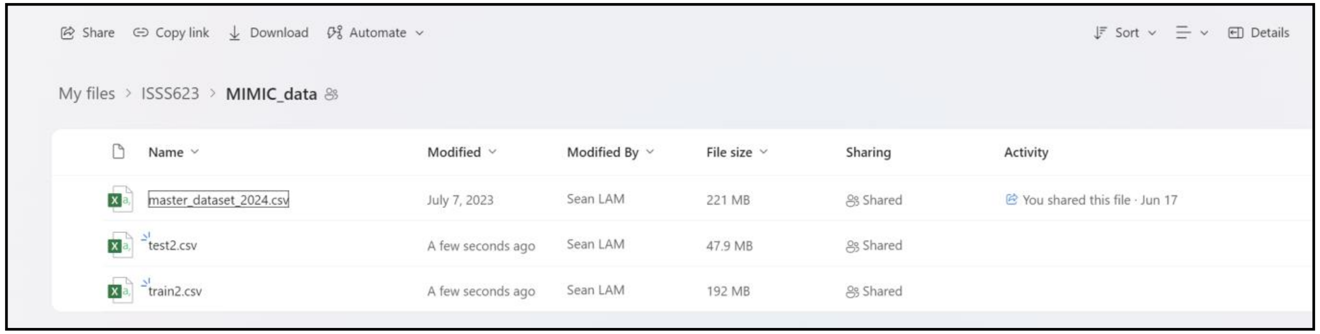


ISSS623_2024 / Group_Project / Self_Learning-MIMIC /

hsrc-projects Lecture 1 Updates 8 July 2024

Name	Last commit message
..	
__pycache__	Lecture 1 Updates 8 July 2024
Task_1_Model_hospitalization_triage.ipynb	Lecture 1 Updates 8 July 2024
Task_2_Model_critical_triage.ipynb	Lecture 1 Updates 8 July 2024
Task_3_Model_72h_ed_revisit_disposition.ipynb	Lecture 1 Updates 8 July 2024
data_general_processing_2024.ipynb	Lecture 1 Updates 8 July 2024
helpers_for_tasks.py	Lecture 1 Updates 8 July 2024
master_dataset_dict_students_2024.csv	Lecture 1 Updates 8 July 2024

- MIMIC-IV wrangled data is downloadable from here:
  - https://smu-my.sharepoint.com/:f:/g/personal/seanlam_smu_edu_sg/EpZ0qWsfTwlKuroJK9bmaJABNG-ahozUEMLr3HVfMv3dKA?e=3812v9



My files > ISSS623 > MIMIC_data

Name	Modified	Modified By	File size	Sharing	Activity
master_dataset_2024.csv	July 7, 2023	Sean LAM	221 MB	Shared	You shared this file · Jun 17
test2.csv	A few seconds ago	Sean LAM	47.9 MB	Shared	
train2.csv	A few seconds ago	Sean LAM	192 MB	Shared	

References

- MIMIC-IV reference:
  - Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2022). MIMIC-IV (version 2.0). *PhysioNet*. https://doi.org/10.13026/7vcr-e114.
  - https://mimic.mit.edu/
- Prior study:
  - https://paperswithcode.com/paper/benchmarking-predictive-risk-models-for