**MScBMI 33200 – Machine Learning for Biomedical Informatics**

**Assignment I**

**<Insert NAME>**

Directions

1. Follow instructions below for each question
2. You can use either R or Python for completing the assignment
3. Upload your answer sheet along with your code (as HTML or PDF) in a separate file. R users can Knit an R markdown into HTML/PDF. Python users can use IPython or Jupyer notebooks and convert them into HTML/PDF (see [instructions](https://stackoverflow.com/questions/15998491/how-to-convert-ipython-notebooks-to-pdf-and-html/25942111) )

Section 1: EMR Bots 30-day Readmission study

Q1) Using the datasets provided for the readmission study, fill out the following Table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Patient Admissions with 30-day readmissions  (n= 128) | Patient Admissions without 30-day readmissions  (n= 36,015) | *P-value* |
| Age, years, mean (sd) | | 44.33 (18.04) | 41.74 (18.06) | 0.10566 |
| Gender | Male, n (%) | 64 (50.00%) | 17,203 (47.77%) | 0.61353 |
|  | Female, n (%) | 64 (50.00%) | 18,812 (52.23%) |
| Race | Black, n (%) | 55 (42.97%) | 17,700 (49.15%) | 0.26216 |
|  | White, n (%) | 33 (25.78%) | 8,251 (22.91%) |
|  | Asian, n (%) | 19 (14.84%) | 4,682 (13.00%) |
|  | Unknown, n (%) | 21 (16.41%) | 5,382 (14.94%) |

Q2) Using the lab data, create a dataset of the **last-observed lab values** for each encounter. Merge this with the patient encounter and outcome tables to create a feature dataset. Make sure you get rid of unwanted columns such as patient ids, date of visits, and timestamps of lab values. Print the summary of the dataset below (you can insert a screenshot).

Ans.

Q3) Using the lab data, create a dataset of the **mean of lab values** for each encounter. Merge this with the patient encounter and outcome tables to create a feature dataset. Make sure you get rid of unwanted columns such as patient ids, date of visits, and timestamps of lab values. Print the summary of the dataset below (you can insert a screenshot).

Ans.

Q4) Split the data longitudinally in training (years <= 2004) and testing (years > 2004). Fill out table below using both count and % in bracket. Percentages are to be calculated within each column with “Total” as the denominator. Basically, you want to make sure that the outcome rate is comparable between the training and the test dataset.

|  |  |  |
| --- | --- | --- |
|  | Training (<= 2004) | Test (> 2004) |
| Patients who were re-admitted within 30 days | n (XX%) |  |
| Patients who were not re-admitted within 30 days |  |  |
| Total |  |  |

Section 2: GUSTO 30-day Mortality Prediction

Q1) Using the datasets provided for the GUSTO study, fill out the following Table 1.

Ans.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Patients who died within 30 days (n=231) | Patients who were alive at 30 days  (n= 3,430) | *P-value* |
| Age, years, mean (sd) | | 71.08 (10.72) | 60.16 (11.56) | 3.5536\*10^-43 |
| Gender | Male, n (%) | 140 (60.61%) | 2,581 (75.25%) | 7.9198\*10^-7 |
|  | Female, n (%) | 91 (39.39%) | 849 (24.75%) |
| Group | sample2, n (%) | 20 (8.66%) | 239 (6.97%) | 0.39140 |
|  | sample4, n (%) | 52 (22.51%) | 733 (21.37%) |
|  | sample5, n (%) | 24 (10.39%) | 405 (11.81%) |
|  | west, n (%) | 135 (58.44%) | 2,053 (59.85) |

Q2) Split the dataset into training (groups = sample2 or sample4 or sample5) and testing (group = west). Fill out table below using both count and % in bracket. As before, use “Total” as the denominator in order to ensure that the outcome rate between training and test is consistent.

Ans.

|  |  |  |
| --- | --- | --- |
|  | Training | Test |
| Patients who died within 30 days | 96 (XX%) | 135 |
| Patients who were alive at 30 days | 1,377 | 2,053 |
| Total |  |  |

Q3) Insert a summary of the train data and test data. Ensure all variables are properly represented as either numeric or categorical. Use the data dictionary as your guide to format the features. Save your workspaces for future use.

Ans.

Section 3: Reading

[1] Steyerberg et al. 2010 Assessing the performance of prediction models: a framework

for some traditional and novel measures, Epidemiology. 2010 January ; 21(1): 128–138

(see [paper](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3575184/)).