Answer to Q1:

a. Import the csv dataset and name it data1 and ensure that all categorical data are treated as categories instead of integers, numeric or text string characters. Show your code.

Categorical variables converted to categories are:
group, outcome, gendera, hypertensive, atrialfibrillation, CHD.with.no.MI, diabetes, deficiencyanemias, depression, Hyperlipemia, Renal.failure, COPD

```
#Import csv
setwd("/Users/FabrianneEffendi/Documents/01 UNI/BC2407 Analytics II/AY21S2 BC2407 CBA")
data1 <- read.csv("data01.csv")

#Convert categorical data to categories
data1$group<-as.factor(data1$group)
data1$outcome<-as.factor(data1$outcome)
data1$gendera<-as.factor(data1$gendera)
data1$hypertensive<-as.factor(data1$hypertensive)
data1$atrialfibrillation<-as.factor(data1$atrialfibrillation)
data1$CHD.with.no.MI<-as.factor(data1$CHD.with.no.MI)
data1$diabetes<-as.factor(data1$diabetes)
data1$deficiencyanemias<-as.factor(data1$deficiencyanemias)
data1$depression<-as.factor(data1$depression)
data1$Hyperlipemia<-as.factor(data1$Hyperlipemia)
data1$Renal.failure<-as.factor(data1$Renal.failure)
data1$COPD<-as.factor(data1$COPD)

str(data1)
```

b. Li F, et al (2021) used the terms Derivation group and Validation group. What is the purpose of the two groups and how is this reflected in the dataset?

Derivation group is the trainset used to train the model, and is reflected as group = 1 in the dataset.
Validation group is the testset used to test the model, and is reflected as group = 2 in the dataset.

c. There are missing values in the dataset. Show a table of missing value counts that shows all those variables that has missing values, it's data type (numeric, integer, factor, character, etc.) and it's missing value count.

	NA.Count	Data.Type
outcome	1	factor
BMI	215	numeric
heart.rate	13	numeric
Systolic.blood.pressure	16	numeric
Diastolic.blood.pressure	16	numeric
Respiratory.rate	13	numeric
temperature	19	numeric
SP.O2	13	numeric
Urine.output	36	numeric
Neutrophils	144	numeric
Basophils	259	numeric
Lymphocyte	145	numeric
PT	20	numeric
INR	20	numeric
Creatine.kinase	165	numeric
glucose	18	numeric
Blood.calcium	1	numeric
PH	292	numeric
Lactic.acid	229	numeric
PCO2	294	numeric

Fig 1.1: Table of missing value count

d. Explore data1. Produce charts, tables or/and statistics to explain 3 interesting findings.

Interesting Finding #1:
ICU HF patients with lower BMI levels have a higher incidence of in-hospital mortality.

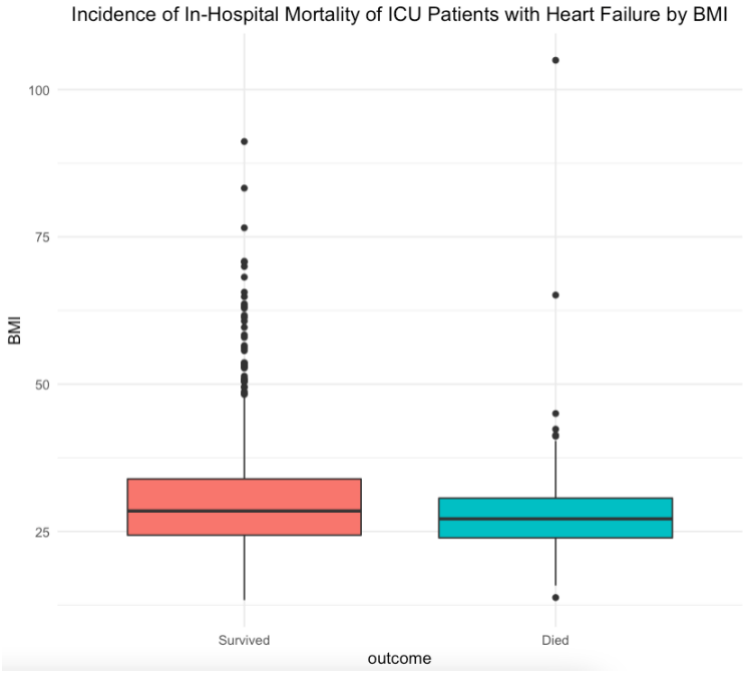


Fig 1.2: Incidence of In-Hospital Mortality of ICU Patients with Heart Failure by BMI

The median BMI of ICU heart failure patients who died in hospital is lower than that of those who survived. Furthermore, the interquartile range of the BMI of ICU heart failure patients who died in hospital is smaller than that of those who survived, with much lesser outside value outliers than that of those who survived, emphasising the low BMI concentration among those who died.

Research reveals that this may due to the 'obesity paradox', which describes a survival benefit for higher BMI in patients with heart failure (Khan, et al, 2021). Even though higher BMIs are associated with greater risk of the development of heart failure, obesity is associated with lower mortality in patients with established heart failure (American College of Cardiology, 2015).

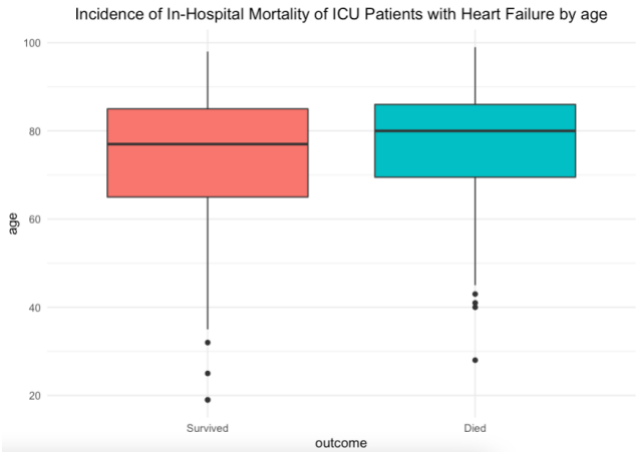


Fig 1.4: Incidence of In-Hospital Mortality of ICU Patients with Heart Failure by age

The possible survival advantage of increased body weight in HF may also be age-dependent. As people become older, BMI may remain relatively stable while muscle mass decreases and body fat increases. In a large worldwide analysis of acute HF, lower BMI was associated with increased age, worsened HF severity, and increased risk of death. Meanwhile, morbidly obese HF patients tend to be younger and present for health care earlier. Some hypothesise that because they have symptoms earlier, therapies with known survival benefits can be initiated earlier in the disease process (Hall, 2018).

Nevertheless, the lack of positive correlation between BMI and in-hospital mortality of ICU-admitted heart failure patients may simply be due to the study's relatively small sample size (Li F, et al, 2021).

Interesting Finding #2:
There is no significant difference in Left Ventricular Ejection Fraction (LVEF) between surviving and non-surviving heart failure ICU patients.

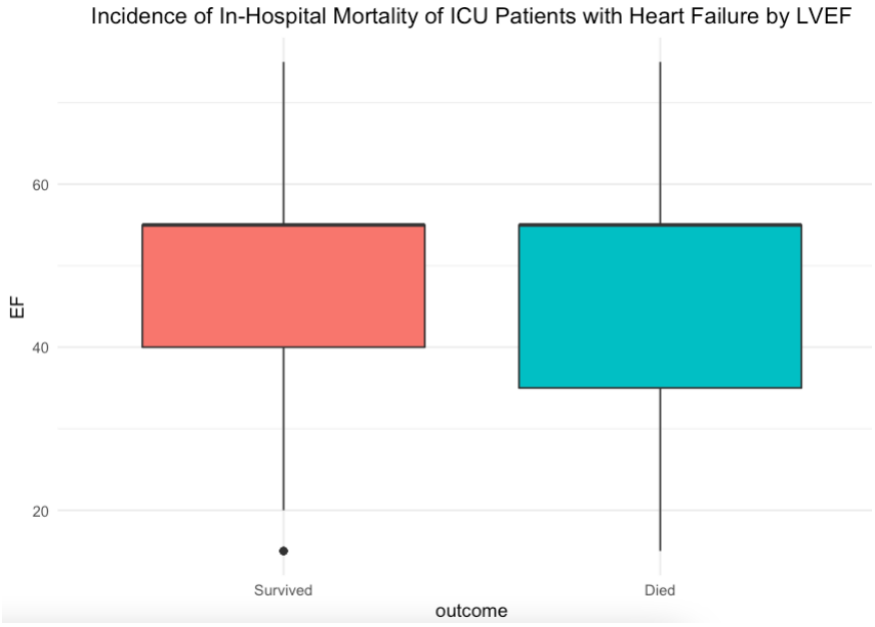


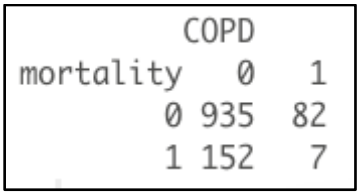
Fig 1.3: Incidence of In-Hospital Mortality of ICU Patients with Heart Failure by LVEF

The mean LVEF of survivors and non-survivors are 48.83 and 47.92 respectively.

This is interesting as LVEF is usually reported to be a predictor of in-hospital mortality of ICU-admitted heart failure patients. In the 2009 paper by Peterson, et al, it was reported that "those who died were more likely to have a prior heart failure diagnosis and an LVEF <40%." Furthermore, another study claims that "Among HF patients in sinus rhythm, higher LVEFs were associated with a linear decrease in mortality up to an LVEF of 45%." (Curtis, et al)

Nevertheless, the result may just be a result of the limitation of the dataset, attributing to the study's relatively small sample size of participants and the duration of hospitalisation in a relatively shorter time.

Interesting Finding #3:
ICU HF patients with chronic obstructive pulmonary disease (COPD) have a lower incidence of in-hospital mortality.



mortality	COPD 0	COPD 1
0	935	82
1	152	7

COPD	Alive	Died
No (n, %)	935 (91.94%)	152 (95.60%)
Yes (n, %)	82 (8.06%)	7 (4.40%)

Fig 1.4: Incidence of In-Hospital Mortality of ICU Patients with Heart Failure and COPD

7.87% of ICU HF patients with chronic obstructive pulmonary disease died in hospital, while 14% of ICU HF patients with no chronic obstructive disease died in hospital.

This is contrary to expected as typically, one would associate multi-disease with higher chance of mortality. Furthermore, severe COPD can cause heart failure in your heart's lower right chamber, or ventricle (WebMD, 2020), exacerbating the heart failure situation. Yet, the results prove otherwise. This might be due to the complex relationship between COPD and heart failure that has yet to be explored, and probably a topic worthy of exploration in future studies.

Answer to Q2:

a. Create a copy of data1 and named it data2. In data2, replace all missing values by the median if the variable is continuous or the mode if the variable is categorical.

```
#Create a copy of data1 and named it data2.
data2 <- data1

#Replace all missing values by median if continuous variable or the mode if categorical variable.

#Replace missing values of continuous variables with median
data2$BMI[is.na(data2$BMI)] <- median(data2$BMI, na.rm = TRUE)
data2$heart.rate[is.na(data2$heart.rate)] <- median(data2$heart.rate, na.rm = TRUE)
data2$Systolic.blood.pressure[is.na(data2$Systolic.blood.pressure)] <- median(data2$Systolic.blood.pressure, na.rm = TRUE)
data2$Diastolic.blood.pressure[is.na(data2$Diastolic.blood.pressure)] <- median(data2$Diastolic.blood.pressure, na.rm = TRUE)
data2$Respiratory.rate[is.na(data2$Respiratory.rate)] <- median(data2$Respiratory.rate, na.rm = TRUE)
data2$temperature[is.na(data2$temperature)] <- median(data2$temperature, na.rm = TRUE)
data2$SP.O2[is.na(data2$SP.O2)] <- median(data2$SP.O2, na.rm = TRUE)
data2$Urine.output[is.na(data2$Urine.output)] <- median(data2$Urine.output, na.rm = TRUE)
data2$Neutrophils[is.na(data2$Neutrophils)] <- median(data2$Neutrophils, na.rm = TRUE)
data2$Basophils[is.na(data2$Basophils)] <- median(data2$Basophils, na.rm = TRUE)
data2$Lymphocyte[is.na(data2$Lymphocyte)] <- median(data2$Lymphocyte, na.rm = TRUE)
data2$PT[is.na(data2$PT)] <- median(data2$PT, na.rm = TRUE)
data2$INR[is.na(data2$INR)] <- median(data2$INR, na.rm = TRUE)
data2$Creatine.kinase[is.na(data2$Creatine.kinase)] <- median(data2$Creatine.kinase, na.rm = TRUE)
data2$glucose[is.na(data2$glucose)] <- median(data2$glucose, na.rm = TRUE)
data2$Blood.calcium[is.na(data2$Blood.calcium)] <- median(data2$Blood.calcium, na.rm = TRUE)
data2$PH[is.na(data2$PH)] <- median(data2$PH, na.rm = TRUE)
data2$Lactic.acid[is.na(data2$Lactic.acid)] <- median(data2$Lactic.acid, na.rm = TRUE)
data2$PCO2[is.na(data2$PCO2)] <- median(data2$PCO2, na.rm = TRUE)

#Mode function
getMode <- function(v){
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v,uniqv)))]
}

#Replace missing values of categorical variable with mode
data2$outcome[is.na(data2$outcome)] <- getMode(data2$outcome)
```

Show the code used to check for missing value count and the output.

```
> sum(is.na(data2))
[1] 0
> sum(is.null(data2))
[1] 0
```

b. Produce a trainset from data2 using group = 1 and named it trainset. Remove group and ID from the trainset and show the proportion of cases who died vs alive.

```
> trainset <- subset(data2, group==1)
> trainset = subset(trainset,select=-c(group,ID))
> #died: outcome=1; alive: outcome=0
> table(trainset$outcome)

  0   1
709 116
```

In the dataset, died: outcome = 1; alive: outcome = 0

Proportion of cases who died	14.06% (116)
Proportion of cases alive	85.94% (709)

c. Produce a testset from data2 using group = 2 and named it testset. Remove group and ID from the testset and show the proportion of cases who died vs alive.

```
> testset <- subset(data2, group==2)
> testset = subset(testset,select=-c(group,ID))
> #died: outcome=1; alive: outcome=0
> table(testset$outcome)

  0   1
309  43
```

In the dataset, died: outcome = 1; alive: outcome = 0

Proportion of cases who died	12.22% (43)
Proportion of cases alive	87.78% (307)

Answer to Q3:

Briefly explain (in bullet points) how you would compute the GWTG predicted outcome on the dataset.

The following 7 predictor variables would be used for GWTG: age, admission systolic blood pressure, admission BUN, admission serum sodium, admission heart rate, non-black race and the presence of COPD, with the points tabulated based on the GWTG-HF risk score table below.

Systolic BP	Points
50-59	28
60-69	26
70-79	24
80-89	23
90-99	21
100-109	19
110-119	17
120-129	15
130-139	13
140-149	11
150-159	9
160-169	8
170-179	6
180-189	4
190-199	2
≥200	0

BUN	Points
≤9	0
10-19	2
20-29	4
30-39	6
40-49	8
50-59	9
60-69	11
70-79	13
80-89	15
90-99	17
100-109	19
110-119	21
120-129	23
130-139	25
140-149	27
≥150	28

Sodium	Points
≤130	4
131	3
132	3
133	3
134	2
135	2
136	2
137	1
138	1
≥139	0

Age	Points
≤19	0
20-29	3
30-39	6
40-49	8
50-59	11
60-69	14
70-79	17
80-89	19
90-99	22
100-109	25
≥110	28

Heart Rate	Points
≤79	0
80-84	1
85-89	3
90-94	4
95-99	5
100-104	6
≥105	8

Black Race	Points
Yes	0
No	3

COPD	Points
Yes	2
No	0

Total Score	Probability of Death
0-33	<1%
34-50	1-5%
51-57	>5-10%
58-61	>10-15%
62-65	>15-20%
66-70	>20-30%
71-74	>30-40%
75-78	>40-50%
≥79	>50%

Fig 3: GWTG-HR Risk Score Table

- Create an overall for loop to tabulate the points, total score and probability of death for each row in the trainset.
- For each predictor variable (except for non-black) race, use if-else statements to hard code the points matching the corresponding ranges in Figure 3.1.
    a. For example, if (COPD==1){points=2} else {points=0}
- For each row, add up all the points from each predictor variable + 3 to create the total score.
    a. The reason for +3 is to account for the non-black race, in which race is assumed to be non-black in the data for GWTG Risk Scoring Model as race/ethnicity is not available in the data but Li F., et al (2021) shows that non-blacks is the great majority
- Write an if-else if statement to correspond the total score to the probability of death:
    a. If the total score is 79 and above, outcome = 1 (as probability of death >50%).
    b. If the total score is below 79, outcome = 0 (as probability of death ≥50%).
- Below is an example of the if-else if statement:

```
heart.rate<-round(heart.rate)
if (heart.rate<=79){
  points=0
}else if ((heart.rate>=80) && (84>=heart.rate)){
  points=1
}else if ((heart.rate>=85) && (89>=heart.rate)){
  points=3
}else if ((heart.rate>=90) && (94>=heart.rate)){
  points=4
}else if ((heart.rate>=95) && (99>=heart.rate)){
  points=5
}else if ((heart.rate>=100) && (104>=heart.rate)){
  points=6
}else{
  points=8
}
gwtg.data$points.heart.rate[x]=points
```

- The probability of in-hospital mortality can be estimated for an individual patient by summing points assigned to the value of each predictor for a total point score with the range of 0 to 100.
- Predict the outcome of the testset using the trained model

Answer to Q4:

Briefly explain (in bullet points) how you would compute the Nomogram predicted outcome on the dataset.

To compute the Nomogram predicted outcome, the following 2 methods were weighed:

- **Method #1:** Use in-hospital mortality prediction formula using XGBoost-selected variables
- **Method #2:** Convert the developed risk in-hospital mortality nomogram into code

Method #2 was finally decided upon to maintain the definition of a Nomogram - a pictorial representation of a formula where each variable is listed separately, with a corresponding number of points assigned to a given magnitude of the variable. The cumulative point score for all the variables is then matched to a scale of outcome.

The Nomogram will use the following 6 predictor variables: Anion gap (Anion.gap), Lactate (Lactic.acid), Calcium (Blood.calcium), BUN (Urea.nitrogen), chronic renal disease (Renal.failure) and diastolic blood pressure (Diastolic.blood.pressure).

Below are the details of the Nomogram implementation, with reference to the Nomogram diagram.

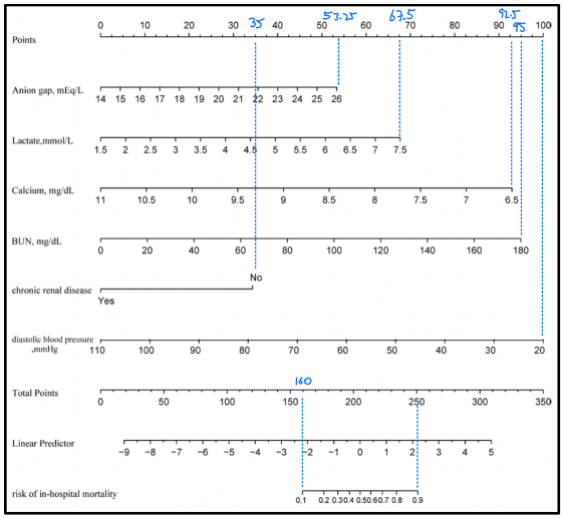


Fig 4 Annotated Nomogram

- Draw a line from the end of the respective predictor's axis upward to the Points axis, with each predictor corresponding to a specific point on the Points axis.
- Mark the specific point on the Points axis for each predictor, noting that the above Points axis has measurements of 2.5 intervals and 1.25 half-intervals. Since the start of each predictor's axis aligns to the 0 mark of the Points axis, the predictor's corresponding length of the Points scale is equal to the specific point drawn.
    - E.g. Length of corresponding Points scale for Anion gap = 53.25 - 0 = 53.25
- Calculate the range of each predictor by subtracting the axis start value from the end value
    - E.g. Range of anion gap = end value - start value = 26 - 14 = 12
- Divide the length of corresponding Points scale by the range of each predictor axis to get the calibrated scale
    - E.g. Calibrated scale = (53.25/12)

- Calculate the points by multiplying the calibrated scale with (dataset value - axis start value)
    - E.g. Point of specific variable = (anion gap - 14)*(53.25/(26-14))
- To account for values that exceeds beyond the predictors' axis range in the nomogram:
    - If value < range, Point = 0
    - If value > range, Point = corresponding Point of the maximum range
    - This is because each predictor variable has a weight contributing to the total points, so it makes more sense for the Points to be capped within that range, rather than extrapolate beyond it.
- Implement the above on the trainset using an overall for loop, with embedded if-else if statements for each predictor variable
- An example of how the if-else if loop is implemented is shown below

```
if (aniongap>26){
  points=53.25+points
}
else if (aniongap>=14){
  points=(aniongap-14)*(53.25/(26-14))+points
}
```

- Sum the total points of all predictor variables and assign them the corresponding risk
    - if Points $\geq$ 205, outcome = 1, else, outcome = 0
- Predict the outcome of the testset using the trained model

Answer to Q5:

Show the table of trainset errors (false positive rate, false negative rate, overall error) from Logistic Regression with Backward Elimination3, Random Forest with default settings, GWTG and Nomogram. Briefly state your findings.

	Model	FPR	FNR	Err
1	Logistic Reg (BE)	0.02	0.60	0.10
2	Random Forest	0.00	0.90	0.13
3	GWTG	0.00	1.00	0.14
4	Nomogram	0.05	0.91	0.17

Fig 5 Trainset errors

Trainset Summary:

- **FPR:** Random forest & GWTG < Logistic Reg (BE) < Nomogram
- **FNR:** Logistic Reg (BE) < Random Forest < Nomogram < GWTG
- **Err:** Logistic Reg (BE) < Random Forest < GWTG < Nomogram

Findings:

- Logistic regression with backward elimination has the lowest overall trainset error of 10% and false negative rate of 60%
- Random Forest and GWTG has the lowest false positive rate of 0%
- Nomogram has the highest overall error rate of 17% and false positive rate of 5%
- GWTG has a false negative rate of 100%, indicating that all actual 1s were predicted as 0s
- In general, all the models have very low false positive rates with reasonably low overall error, but very high false negative rates

Answer to Q6:

Show the table of testset errors (false positive rate, false negative rate, overall error) from Logistic Regression with Backward Elimination, Random Forest with default settings, GWTG and Nomogram. Your table structure should look like the above table too. Briefly state your findings.

	Model	FPR	FNR	Err
1	Logistic Reg (BE)	0.04	0.67	0.12
2	Random Forest	0.00	0.91	0.11
3	GWTG	0.00	1.00	0.12
4	Nomogram	0.03	0.86	0.13

Fig 6: Testset errors

Testset Summary:

- **FPR:** Random forest & GWTG < Nomogram < Logistic Reg (BE)
- **FNR:** Logistic Reg (BE) < Nomogram < Random Forest < GWTG
- **Err:** Random Forest < Logistic Reg (BE) & GWTG < Nomogram

Findings:

- Random Forest has the lowest overall error of 11%, but a very high false negative rate of 91%
- Random Forest and GWTG has the lowest false positive rate of 0%
- Logistic regression with backward elimination has the lowest false negative rate of 67%
- Nomogram has the highest overall error rate of 13%
- GWTG has a false negative rate of 100%, indicating that all actual 1s were predicted as 0s.
    - This might be due to the fact that GWTG was derived from a generic population of in-hospital heart failure patients, rather than specifically ICU patients
- In general, all the models have very low false positive rates with reasonably low overall error, but very high false negative rates

Answer to Q7:

The testset errors might be skewed to one side as the trainset is unbalanced. Balance the trainset by sampling from the majority4 to obtain 50-50 distribution of alive vs death in the trainset.

Below code was used to obtain a trainset with 116 '0' outcome and 116 '1' outcome.

```
# Random sample from majority class and combine with minority to form new trainset
majority <- subset(trainset, outcome == 0)
minority <- subset(trainset, outcome == 1)

# Randomly sample the row numbers to be in trainset. Same sample size as minority cases.
set.seed(22)
chosen <- sample(seq(1:nrow(majority)), size = nrow(minority))

# Subset the original trainset based on randomly chosen row numbers.
majority.chosen <- majority[chosen,]

# Combine two data tables by appending the rows
trainset.bal <- rbind(majority.chosen, minority)

## Check trainset is balanced.
summary(trainset.bal)
```

Show the table of testset errors (false positive rate, false negative rate, overall error) from Logistic Regression with Backward Elimination and Random Forest with default settings. Briefly state your findings.

	Model	FPR	FNR	Err
1	Logistic Reg (BE)	0.30	0.33	0.31
2	Random Forest	0.29	0.23	0.28

Fig 7: Testset errors

Testset Summary:

- **FPR:** Random Forest < Logistic Reg (BE)
- **FNR:** Random Forest < Logistic Reg (BE)
- **Err:** Random Forest < Logistic Reg (BE)

Findings:

- Random Forest has the lowest overall error, false positive rate and false negative rate compared to logistic regression
- Sampling has caused random forest's error margins to decrease significantly, but caused logistic regression's error margins to increase

Answer to Q8:

Li F., et al (2021) used the top 20 variable importance from XGBoost as predictors into logistic regression. Extract the top 20 variable importance (permutation approach) from Random Forest trained on balanced trainset and fit them as predictors into logistic regression.

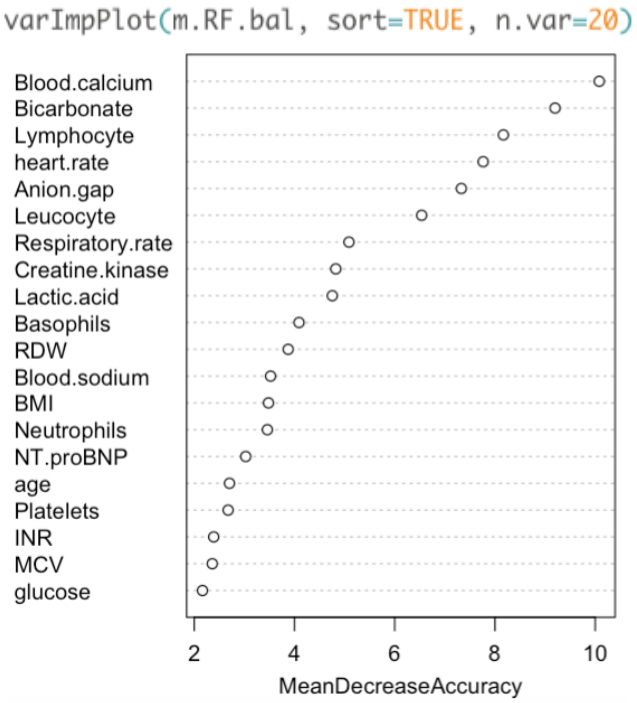


Fig 8.1: Top 20 Predictors by Variable Importance using Random Forest

Append the testset results and show the table.

	Model	FPR	FNR	Err
1	Logistic Reg (BE)	0.3	0.33	0.31
2	Random Forest	0.29	0.23	0.28
3	RF VarImpt into Logistic Reg	0.32	0.4	0.33

Fig 8.2: Testset results

Testset Summary:

- **FPR:** Random Forest < Logistic Reg (BE) < RF VarImpt into Logistic Reg
- **FNR:** Random Forest < Logistic Reg (BE) < RF VarImpt into Logistic Reg
- **Err:** Random Forest < Logistic Reg (BE) < RF VarImpt into Logistic Reg

Is this model superior than stand-alone logistic regression (with backward elimination) or Random Forest? Briefly state your findings.

The model is worse than stand-alone logistic regression or Random Forest, with the highest false positive rate, false negative rate and overall error. Simply taking the top 20 predictors of the Random Forest as the final inputs of the logistic regression model decreases the accuracy as the top 20 predictors are artificially limited by the cap of 20, instead of letting the model naturally choose the best number of predictor variables.

Answer to Q9:

A hospital in Singapore is thinking of using a risk scoring system to assess ICU patient mortality. What is your recommendation?

The models developed in this assignment are specific to in-hospital mortality in ICU patients with heart failure. Hence, there is a need for a more generalised methodology if a risk score system were to be implemented in the Singapore hospital to assess ICU patient mortality. The models developed in this assignment would be leveraged as a case analysis for generalisability.

Therefore, my recommendation is a multi-tiered risk scoring system, in which the 1st tier is a Generic ICU Mortality Risk Score, while the 2nd tier is an Organ System-specific Mortality Risk Score. This would cater to the different needs and use cases in assessing the risk of ICU patient mortality, such as in aiding in the prioritisation of which patient to attend to in the ICU, especially in matters relating to life and death.

For situations of high urgency and immediate attention warranting a quick way of assessing risk without first knowing the disease of the patient, the Generic ICU Mortality Risk Score can be used. Furthermore, in situations where multiple organ failure is likely, the Generic ICU Mortality Risk Score may be used. Meanwhile, for situations where the specific source of disease is known, the Organ System-specific Mortality Risk Score will be utilised, allowing for more catered evaluations of risk, similar to how Li F., et al (2021) developed the Nomogram specific for predicting in-hospital mortality in ICU patients with heart failure. The Organ System-specific Mortality Risk Score would be developed for key organ systems associated as key risk factors for death, for example a study shows that main risk factors for death in the intensive care unit include the central nervous system failure and cardiovascular failure (MEDICA Magazine, 2022).

The models to create such a robust risk scoring system may be developed using a hybrid machine learning model approach, where the best of the different machine learning models are incorporated as each machine learning model has their own strengths. Even though through the analysis of this assignment, Nomogram appears to be the worst performing model compared to the other models, this is due to the limitations of the methodology in 'imitating' a Nomogram in the hard coding of calculations phase (elaborated in question 10), and its massive merits should not be discounted. Hence, I would recommend an approach similar to the XGBoost - Logistic Regression - Nomogram approach as suggested by Li F. et al. (2021). This is because XGBoost is highly flexible, allowing users to define custom optimisation objectives and evaluation criteria. Furthermore, when tuned with the right parameters, XGBoost yields results of very high accuracy due to its optimisation techniques such as parallel Processing, regularisation, tree pruning and built-in-cross-validation.

When developing the Nomogram, Li F., et al (2021) filtered patients in the data sample by ensuring that only (i) ICU admission patients, (ii) patients with NT-proBNP, (iii) patients with records of echocardiography, are included in the dataset. Drawing parallels from this, we can adopt a similar methodology to create robust Organ System-Specific mortality risk scores, in which patients are filtered based on key tests taken relating to the organ-related disease e.g. NT-proBNP and echocardiography for cardiovascular failures.

The implementation of the multi-tiered risk scoring system would be able to cater to the complexities and nuances of ICU care, and provide a much more accurate and reliable risk scoring assessment of the ICU patient mortality.

Answer to Q10:

Suggest other ideas that may improve the accuracy of the model.

Suggestions to improve general accuracy across models

To increase the general accuracy of the model such that it can better predict a wider variety of ICU heart failure situations, the sample size can be increased to include patients' records from the ICUs of other hospitals, instead of limiting it to a single hospital. Furthermore, more recent data can be added in the training of the model instead of being limited to the 2001-2012 data extracted from MIMIC database, during which the treatment of heart failure might have changed greatly, resulting in decreased accuracy of the model when applied to current situations. Having more data would reduce reliance on assumptions and weak correlations, hence improving accuracy across models.

Additionally, if such data is available, having an additional variable that classifies the severity of heart failure might aid in increasing accuracy of the model. Clinically, heart failure is categorised in 4 stages (Stage A, B, C and D). The data selected by Li F. et al. (2021) are all patients with NT-proBNP record. Instead of merely using this record as a selection criteria, the test record may be further used to find out the severity of the heart failure condition. This is because greater severity of heart failure correlates to higher in–hospital mortality risk.

Suggestions to improve Nomogram model

To improve the accuracy of the Nomogram model, the equation method should be used instead of hard coding the pictorial Nomogram. That is, using the in-hospital mortality prediction formula using XGBoost-selected variables to predict in-hospital mortality for ICU patients:

$$log\ odds\ of\ mortality = 4.62536 + 0.24559 \times anion\ gap + 0.61542 \times lactate - 1.04993 \times calcium + 0.02687 \times BUN - 1.76330 \times CKD - 0.05633 \times DBP$$

By definition, Nomogram is a pictorial representation of a formula where each predictor corresponds to a specific point by drawing a line straight upward to the Points axis. One of the main uses of such Nomograms is usually for quick diagnostics by medical practitioners during a patient consultation as it makes prognostic model results easier to read. However, when developing such models, such pictorial representations may come at the expense of accuracy which is limited to the subdivision of the axis range. This is evident by its consistently worst overall error rate in train and test sets, contrary to what is expected. The Nomogram's XGBoost-based algorithm should supposedly have superior accuracy compared to the other 3 models in this assignment due to the fact that XGBoost uses Gradient Boosting, a method of building an ensemble of classification trees in an additive and sequential manner, where each subsequent new tree grown learns from its predecessor through a gradient descent algorithm.

Suggestions to improve Random Forest model

To improve the accuracy of the random forest model (the model with lowest overall test error in question 6), optimisation techniques such as Bayesian Optimisation may be used to fine-tune the hyperparameters of the random forest model such as maximum features, number of trees, Out-Of-Bag score and other features.

Appendix A: List of assumptions on the data or analysis

[If necessary, you may add your assumptions into the list.]

1. Race is assumed to be non-black in the data for GWTG Risk Scoring Model as race/ethnicity is not available in the data but Li F., et al (2021) shows that non-blacks is the great majority (approx. 86%).
2. Nomogram model in this assignment is to be developed using the line method and not the formula method.

Bibliography

American College of Cardiology. (2015). *The Obesity Paradox in Heart Failure: Why Does It Clinically Matter?* https://www.acc.org/latest-in-cardiology/articles/2015/05/06/10/22/http%3a%2f%2fwww.acc.org%2flatest-in-cardiology%2farticles%2f2015%2f05%2f06%2f10%2f22%2fthe-obesity-paradox-in-heart-failure

Curtis, J. P., Sokol, S. I., Wang, Y., Rathore, S. S., Ko, D. T., Jadbabaie, F., Portnay, E. L., Marshalko, S. J., Radford, M. J., & Krumholz, H. M. (2003). The association of left ventricular ejection fraction, mortality, and cause of death in stable outpatients with heart failure. *Journal of the American College of Cardiology*, *42*(4), 736–742. https://doi.org/10.1016/s0735-1097(03)00789-7

Hall, M. E. (2018). Body Mass Index and Heart Failure Mortality. *JACC. Heart Failure*, *6*(3), 243–245. https://doi.org/10.1016/j.jchf.2017.12.013

Khan, A., Iterson, E. H. V., & Laffin, L. J. (2021). The obesity paradox in heart failure: What is the role of cardiorespiratory fitness? *Cleveland Clinic Journal of Medicine*, *88*(8), 449–458. https://doi.org/10.3949/ccjm.88a.20098

Li, F., Xin, H., Zhang, J., Fu, M., Zhou, J., & Lian, Z. (2021). Prediction model of in-hospital mortality in intensive care unit patients with heart failure: Machine learning-based, retrospective analysis of the MIMIC-III database. *BMJ Open*, *11*(7), e044779. https://doi.org/10.1136/bmjopen-2020-044779

MEDICA Magazine. (2022). *Main Cause of Death in Intensive Care Units*. https://www.medica-tradefair.com/en/News/Archive/Main_Cause_of_Death_in_Intensive_Care_Units

Peterson, P. N., Rumsfeld, J. S., Liang, L., Albert, N. M., Hernandez, A. F., Peterson, E. D., Fonarow, G. C., Masoudi, F. A., & American Heart Association Get With the Guidelines-Heart Failure Program. (2010). A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. *Circulation. Cardiovascular Quality and Outcomes*, *3*(1), 25–32. https://doi.org/10.1161/CIRCOUTCOMES.109.854877

WebMD. (2020, December 7). *What's the Link Between COPD and Heart Failure?* https://www.webmd.com/lung/copd/copd-and-heart-failure