Final Report

Z. Love

2022-12-16

Introduction:

As technology continues to develop and improve the field of medical research has been interested in using available medical records to understand more about disease and improve treatment outcomes. Considerable work has been done to make medical data available to researchers, including the work at MIT and Beth Israel Deaconess Medical Center (BIDMC) to release the MIMIC datasets [1,2]. The most recent release of the MIMIC data is MIMIC-IV. This release contains de-identified data from ICU and hospital patients that have been admitted to BIDC between the years of 2008 and 2019 (Fig. 1,2,3). The patient data includes physiologic data collected from monitoring, lab results, demographic data, and treatment/diagnosis data. While this data is de-identified to protect the privacy of the patients included in MIMIC-IV, additional steps are taken before an interested party can access the data. To access MIMIC-IV, one must take an online training course in patient data privacy laws and regulations as well as sign off on data use agreements before one is credentialed for MIMIC-IV access.

Over the years, many papers have been published using the MIMIC data for clinical data science. Many of these papers include the generation and deployment of prediction models [3]. For this project I will be generating a mortality prediction model with a constrained version of the MIMIC data only involving patients from the ICU. When patients enter the ICU there is a wide range of severity in disease that patients experience. It is important for doctors and medical staff to quickly assess the severity and create a treatment plan tailored to that individual. While there are a couple simple scoring models that doctors can use they have relatively low accuracy for predicting future mortality.

This project will determine if there is a viable method for predicting mortality based on the laboratory vitals collected from patients within the first 24 hours in the ICU.There are around 40 laboratory tests that are collected in the MIMIC data including: lactate, hemoglobin, and platelet levels throughout the patient's stay in the ICU (Table 1). This prediction model will give medical professionals insight into the severity of their patient's disease based on their laboratory result. Additionally, two models will have their performance evaluated and compared, logistic regression and random forest. These models were selected as they have been commonly used models in the area of mortality prediction research [4]. The better performing model out of the two will be selected for the final prediction model.

Results

The MIMIC-IV dataset is a relational database composed of many tables that contain patient data collected from a variety of sources. It is recommended by the creators of MIMIC that the data is accessed through the cloud. For this project Google's BigQuery was implemented to explore the contents of each table and to locate the laboratory test results as well as demographic patient information. Once credentialed for access using the physionet platform and linking google accounts SQL code was used to extract MIMIC data into a Google Colab notebook. Demographic and hospital stay information was extracted from the MIMIC-IV tables: mimiciv_hosp.admissions, mimiciv_icu.icustays, and mimiciv_hosp.patients. Lab results were extracted from mimiciv_hosp.labevents. Tables were merged on the shared columns: 'subject_id', 'stay_id', and 'hadm_id' (hospital admission ID). The combination of these three identifiers generate a unique patient stay in the ICU.

Once merged, the merged tables were then filtered to remove any subjects that were missing all demographic information or all laboratory values. Additionally, the subjects were removed from the study if they had more than 20% missing data across all included variables. This table was then exported as a csv file for further tidying using R. In order to reduce processing time and to improve the accuracy of the model the entire dataset was downsampled so the ratio of cases to controls was around 3:1 (controls: cases). In this study, cases include any subjects with the outcome of interest, mortality. The dataset was reduced from around 50,000 subjects to around 17,000 subjects. At this point the final dataset was generated and prepared for modeling and analysis (Fig. 4).

For logistic regression and random forest model a training and testing datasets were generated with a 80/20 split. Logistic regression modeling was performed and tested with an accuracy of 0.74, indicating that the model was able to correctly identify patients that were at risk for mortality 74% of the time. The logistic regression model also reported a sensitivity or recall score of 0.76 and specificity of 0.71 showing 76% of the time mortality was predicted correctly and 71% of the time survival was predicted correctly (Table 2). Using the 'vip' package in r the variable importance was plotted for the logistic regression model [5]. The plot shows, 'albumin_min', 'lactate_max', 'lactate_min', and 'sodium_max' as the top predictive features of mortality for MIMIC-IV ICU patients (Fig. 5). Finally, the ROC-AUC curve was plotted for the logistic regression model and reported an AUC of 0.74. This value closely matches the accuracy of the model and provides additional evidence of the predictive qualities of the model (Fig. 7).

The random forest model was then generated using the same train and test as a basis of the model. However, for the random forest model the variables were imputed for missing values using the median value of the feature. The model was then evaluated using the test set and reported an accuracy of 0.85. The random forest model predicted more mortality outcomes correctly compared with the logistic regression model. Aside from accuracy the random forest model has a sensitivity and 0.97 and specificity of 0.34 (Table 2). These values are not balanced with almost all cases of mortality correctly predicted and very few survival cases predicted correctly. Since prevalence of 0.80 is reported, it can be explained that due to the relatively low proportion of survival according to prevalence, the low specificity (ability to predict survival) does not impact the overall accuracy. The top predictive features were plotted and include: 'lactate_mean', 'lactate_min', 'lactate_max', 'phosphate_max', and 'sodium_max' (Fig. 6). Lastly, the ROC-AUC curve was generated and found an AUC of 0.66. This decrease in performance in comparison with the accuracy is reflective of the low specificity of the model (Fig. 8).

Conclusion

The goal of this study was to report the predictive properties of laboratory results of MIMIC-IV patients in the ICU on the outcome of mortality. The features were evaluated using logistic regression and random forest modeling to determine the relative importance of each feature as well as the function of the mortality prediction model overall. While the logistic regression model did have a lower overall accuracy compared to the random forest, there was more even balance in the sensitivity and specificity metrics. In the case of mortality prediction in the ICU there is greater concern in a low sensitivity as that would indicate a low ability to predict mortality which has a greater consequence than low specificity, indicating reduced survival prediction. However, it is not ideal for a model to struggle with a key metric evaluating the overall performance. Therefore, the logistic regression model is a better overall, more reliable model for mortality prediction in the ICU. Future studies of the random forest model using this data could look to determine a more optimal 'mtry' value or number of nodes to include in the model and could potentially improve the model performance.

Both models found that the lactate variable is highly predictive of mortality. As well as lactate, both models selected sodium as an important variable for predicting mortality. These results are in line with current medical understanding. Changes in lactate are often noted as the onset for changes in blood pressure and heart rate which are important indicators of health [6]. Lactate levels have also been reported to correlate with the diagnosis of septic shock in critically ill patients [7]. High sodium levels have also been reported to indicate renal function decline and dehydration which both are important for patient health [8,9]. The top predictive features selected by both models are not surprising and give greater evidence that the predictive model is informative of patient health within the first 24 hours of lab results collected.

In future work, more physiologic data could be added to the model to improve the accuracy of the prediction and to determine the relative importance of laboratory results in predicting mortality compared with other data, for example, vital signs and comorbidities. As it is, the models provide insight into ICU patients and the predictive qualities of their laboratory results collected on the risk for mortality. The results of this report also indicate that logistic regression is an acceptable model for mortality prediction and should continue to be a metric in which we compare newer or more complex model's performance.

References

1. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. E215–e220.
2. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2022). MIMIC-IV (version 2.1). PhysioNet. https://doi.org/10.13026/rrgf-xw32.
3. Pang K, Li L, Ouyang W, Liu X, Tang Y. Establishment of ICU Mortality Risk Prediction Models with Machine Learning Algorithm Using MIMIC-IV Database. Diagnostics. 2022; 12(5):1068. https://doi.org/10.3390/diagnostics12051068
4. Naemi A, Schmidt T, Mansourvar M, et alMachine learning techniques for mortality prediction in emergency departments: a systematic reviewBMJ Open 2021;11:e052663. doi: 10.1136/bmjopen-2021-052663
5. Brandon M. Greenwell and Bradley C. Boehmke (2020). Variable Importance Plots—An Introduction to the vip Package. The R Journal, 12(1), 343–366. URL https://doi.org/10.32614/RJ-2020-013.
6. Okorie ON, Dellinger P. Lactate: biomarker and potential therapeutic target. Crit Care Clin. 2011 Apr;27(2):299-326. doi: 10.1016/j.ccc.2010.12.013. PMID: 21440203.
7. Dong Hyun Oh, Moo Hyun Kim, Woo Yong Jeong, Yong Chan Kim, Eun Jin Kim, Je Eun Song, In Young Jung, Su Jin Jeong, Nam Su Ku, Jun Yong Choi, Young Goo Song, June Myung Kim, Risk factors for mortality in patients with low lactate level and septic shock, Journal of Microbiology, Immunology and Infection, Volume 52, Issue 3, 2019,Pages 418-425, https://doi.org/10.1016/j.jmii.2017.08.009.
8. Braun MM, Barstow CH, Pyzocha NJ. Diagnosis and management of sodium disorders: hyponatremia and hypernatremia. Am Fam Physician. 2015 Mar 1;91(5):299-307. PMID: 25822386.
9. Cook NR, He FJ, MacGregor GA, Graudal N. Sodium and health-concordance and controversy. BMJ. 2020 Jun 26;369:m2440. doi: 10.1136/bmj.m2440. Erratum in: BMJ. 2020 Jun 29;369:m2608. He, J [corrected to He, Feng J]. PMID: 32591335; PMCID: PMC7318881.

Appendix

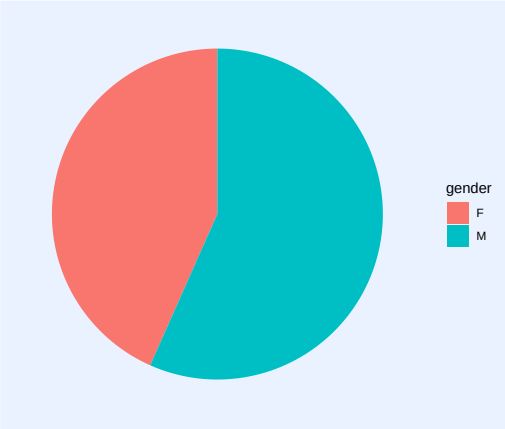

Figure 1: Percentage of each gender incldued in the MIMIC-IV derived data

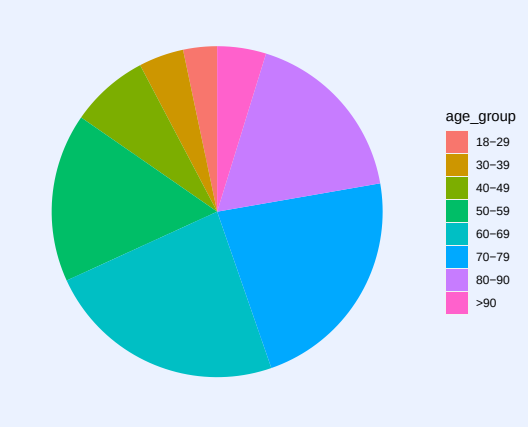

Figure 2: Percentage of each age group incldued in the MIMIC-IV derived data

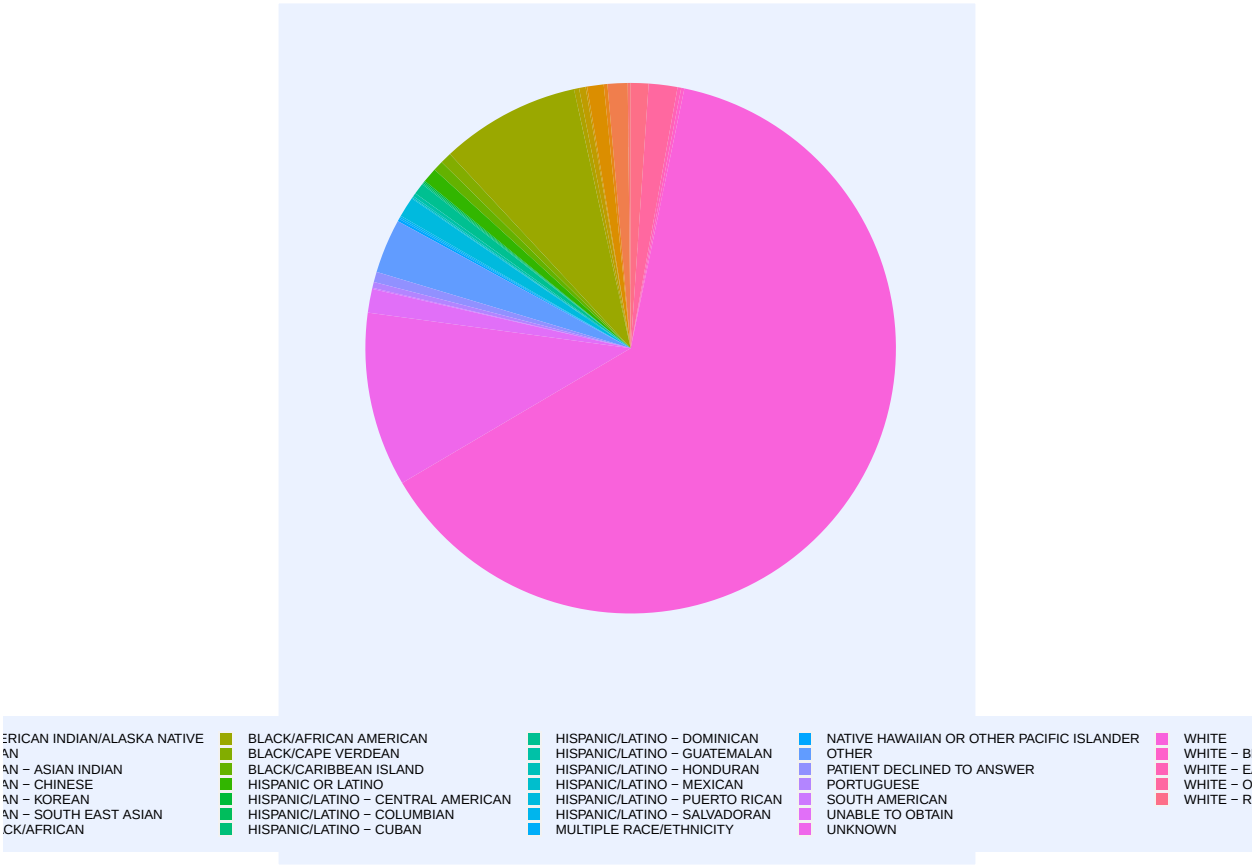

Figure 3: Percentage of each ethnicity incldued in the MIMIC-IV derived data

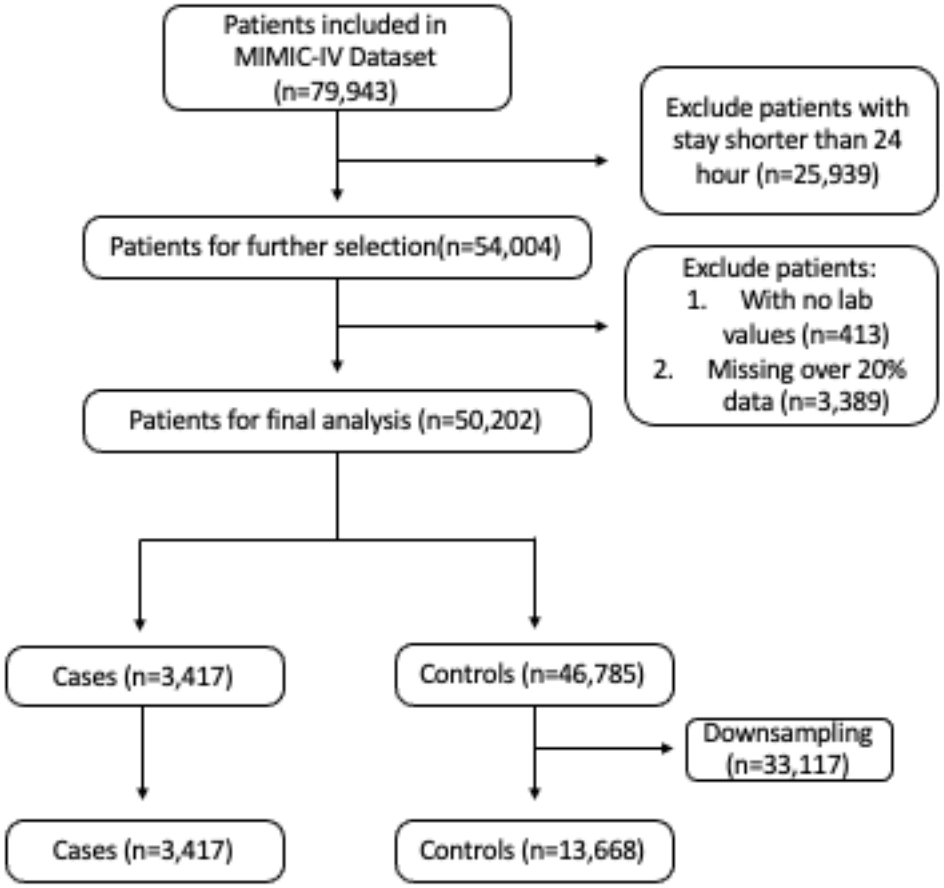

Figure 4: flowchart of study population selection from MIMIC-IV

Table 1: Overview of MIMIC-IV data

Feature	Survival	Mortality
n	13668	3417
admission_age (mean (SD))	64.99 (16.44)	69.77 (15.22)
gender = M (%)	7827 (57.3)	1861 (54.5)
los_icu (mean (SD))	4.06 (5.47)	6.13 (5.42)
aniongap_min (mean (SD))	12.92 (3.36)	15.20 (4.67)
aniongap_max (mean (SD))	15.40 (4.21)	19.26 (6.42)
albumin_min (mean (SD))	3.15 (0.66)	2.87 (0.73)
bicarbonate_min (mean (SD))	22.30 (4.54)	19.56 (6.07)
bicarbonate_max (mean (SD))	24.44 (4.17)	22.92 (5.30)
bilirubin_min (mean (SD))	1.57 (3.61)	2.87 (6.22)
bilirubin_max (mean (SD))	1.78 (3.93)	3.32 (6.89)
creatinine_min (mean (SD))	1.28 (1.37)	1.72 (1.46)
creatinine_max (mean (SD))	1.49 (1.81)	2.09 (1.72)
chloride_min (mean (SD))	102.09 (6.15)	100.86 (7.32)
chloride_max (mean (SD))	105.79 (6.93)	105.69 (8.11)
hematocrit_min (mean (SD))	30.08 (6.73)	29.63 (6.98)
hematocrit_max (mean (SD))	34.63 (6.03)	34.04 (6.60)
hemoglobin_min (mean (SD))	9.97 (2.27)	9.66 (2.34)
hemoglobin_max (mean (SD))	11.39 (2.11)	11.03 (2.24)
lactate_min (mean (SD))	1.49 (0.78)	2.46 (2.09)
lactate_max (mean (SD))	2.55 (1.84)	4.62 (4.16)
lactate_mean (mean (SD))	1.99 (1.17)	3.49 (2.97)
magnesium_min (mean (SD))	1.91 (0.39)	1.90 (0.39)
magnesium_max (mean (SD))	2.22 (0.92)	2.31 (0.96)
phosphate_min (mean (SD))	3.30 (1.12)	3.73 (1.60)
phosphate_max (mean (SD))	3.90 (1.44)	4.93 (2.12)
platelet_min (mean (SD))	192.11 (103.36)	175.86 (113.93)
platelet_max (mean (SD))	219.29 (109.84)	213.06 (125.11)
potassium_min (mean (SD))	3.83 (0.54)	3.88 (0.68)
potassium_max (mean (SD))	4.58 (0.85)	4.79 (0.94)
ptt_min (mean (SD))	32.27 (12.53)	36.76 (16.55)
ptt_max (mean (SD))	42.31 (28.27)	54.38 (37.52)
inr_min (mean (SD))	1.37 (0.64)	1.65 (0.93)
inr_max (mean (SD))	1.54 (0.98)	2.07 (1.59)
pt_min (mean (SD))	14.96 (6.62)	17.91 (9.58)
pt_max (mean (SD))	16.72 (9.59)	22.37 (16.87)
sodium_min (mean (SD))	136.35 (5.10)	136.00 (6.09)
sodium_max (mean (SD))	139.62 (4.71)	140.45 (6.61)
bun_min (mean (SD))	23.34 (19.43)	34.84 (25.82)
bun_max (mean (SD))	26.86 (21.97)	40.51 (28.36)
bun_mean (mean (SD))	25.09 (20.49)	37.71 (26.90)
wbc_min (mean (SD))	10.44 (6.86)	12.59 (9.97)
wbc_max (mean (SD))	13.39 (9.87)	16.76 (13.64)
wbc_mean (mean (SD))	11.89 (7.88)	14.61 (11.51)

```
## # A tibble: 76,943 x 63
##     ...1 subject_id  hadm_id  stay_id gender dod        admittime
##    <dbl>      <dbl>    <dbl>    <dbl> <chr>  <date>     <dttm>
##  1     0   19669999 20005479 32977919 F      2148-08-10 2148-06-01 12:48:00
##  2     1   17002995 20006309 31646901 M      NA         2177-12-02 01:39:00
##  3     2   10559183 20008400 36107959 M      2116-11-14 2116-01-03 14:50:00
##  4     3   18172155 20009330 36841282 M      NA         2144-01-01 00:33:00
##  5     4   18549459 20021612 34145253 F      NA         2162-05-06 17:35:00
##  6     5   12801663 20026025 30820844 M      2172-09-26 2171-10-25 06:14:00
##  7     6   13824324 20026038 36969919 F      NA         2110-09-21 01:20:00
##  8     7   16502573 20028981 38516488 M      NA         2166-12-01 18:14:00
##  9     8   13090274 20032357 35077467 M      NA         2154-06-02 23:26:00
## 10     9   13132088 20045152 34108286 M      NA         2197-07-18 22:06:00
## # ... with 76,933 more rows, and 56 more variables: dischtime <dttm>,
## #   los_hospital <dbl>, admission_age <dbl>, race <chr>,
## #   hospital_expire_flag <dbl>, hospstay_seq <dbl>, first_hosp_stay <lgl>,
## #   icu_intime <dttm>, icu_outtime <dttm>, los_icu <dbl>, icustay_seq <dbl>,
## #   first_icu_stay <lgl>, mort_icu <dbl>, mort_hosp <dbl>, aniongap_min <dbl>,
## #   aniongap_max <dbl>, albumin_min <dbl>, albumin_max <dbl>,
## #   bicarbonate_min <dbl>, bicarbonate_max <dbl>, bilirubin_min <dbl>, ...
```

```
## # A tibble: 17,085 x 63
##     ...1 subject_id  hadm_id  stay_id gender dod    admittime
##    <dbl>      <dbl>    <dbl>    <dbl> <chr>  <date> <dttm>
##  1 71371   18730396 27843196 32293861 M      NA     2137-01-07 13:05:00
##  2 51027   14416150 22762399 37643228 M      NA     2192-04-13 02:33:00
##  3 74510   13887386 27608534 39676044 M      NA     2182-12-01 07:25:00
##  4 67355   18363524 26794552 30269184 M      NA     2125-04-22 00:28:00
##  5 10295   18056033 23500830 39947778 M      NA     2151-07-18 19:01:00
##  6 53916   15875773 20875756 39025103 F      NA     2114-04-03 11:22:00
##  7 25379   14271881 29053297 37866560 M      NA     2121-03-15 00:23:00
##  8 59620   11413236 25471688 39619449 F      NA     2185-11-20 04:00:00
##  9 53782   12325110 29214383 36440882 M      NA     2152-03-12 21:31:00
## 10 13332   11669075 22701290 31098095 M      NA     2147-01-06 18:12:00
## # ... with 17,075 more rows, and 56 more variables: dischtime <dttm>,
## #   los_hospital <dbl>, admission_age <dbl>, race <chr>,
## #   hospital_expire_flag <dbl>, hospstay_seq <dbl>, first_hosp_stay <lgl>,
## #   icu_intime <dttm>, icu_outtime <dttm>, los_icu <dbl>, icustay_seq <dbl>,
## #   first_icu_stay <lgl>, mort_icu <dbl>, mort_hosp <dbl>, aniongap_min <dbl>,
## #   aniongap_max <dbl>, albumin_min <dbl>, albumin_max <dbl>,
## #   bicarbonate_min <dbl>, bicarbonate_max <dbl>, bilirubin_min <dbl>, ...
```

Table 2: Evalutation Metrics for Predictive Models

Model	Accuracy	Sensitivity	Specificity	Prevalence
Random Forest	0.8486977	0.9669811	0.3555219	0.8065555
Logistic Regression	0.7471840	0.7555911	0.7167630	0.7834793

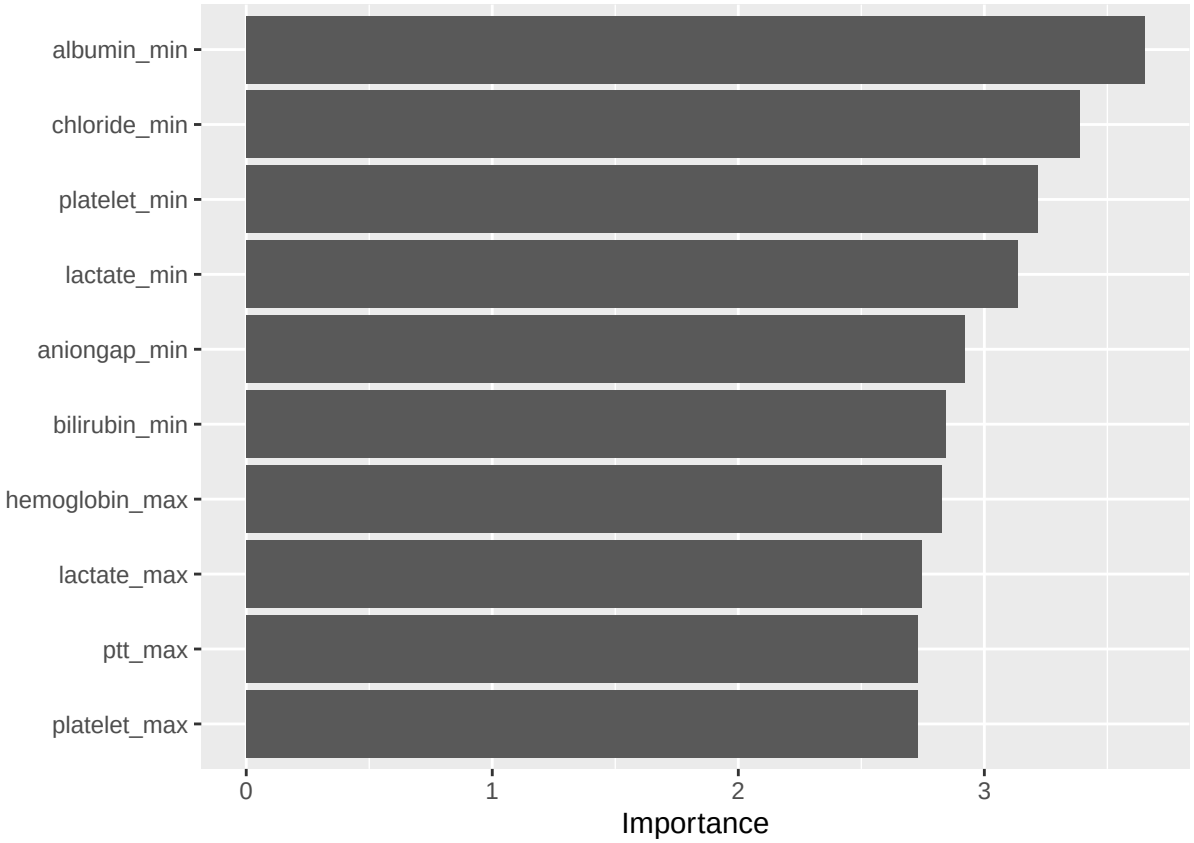

Figure 5: Top predictive features selected by the logsitic regression model

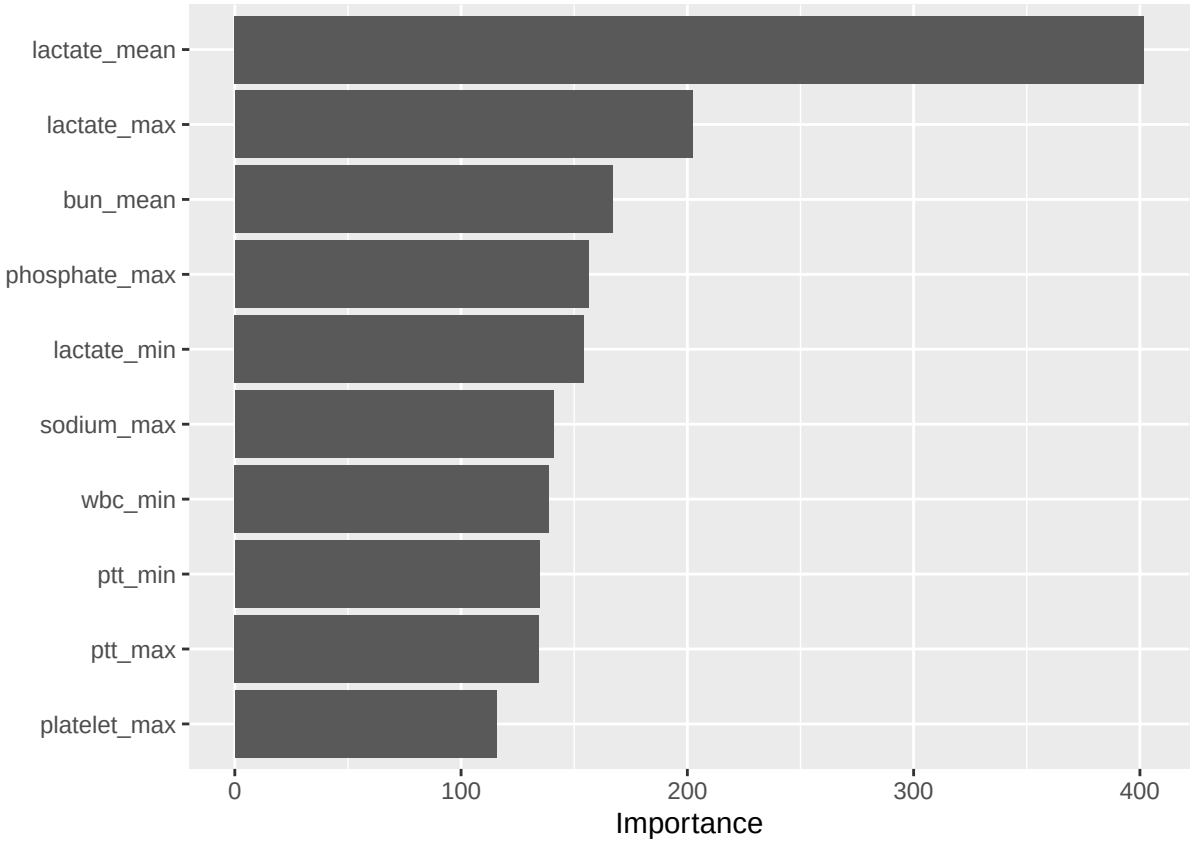

Figure 6: Top predictive features selected by the random forest model

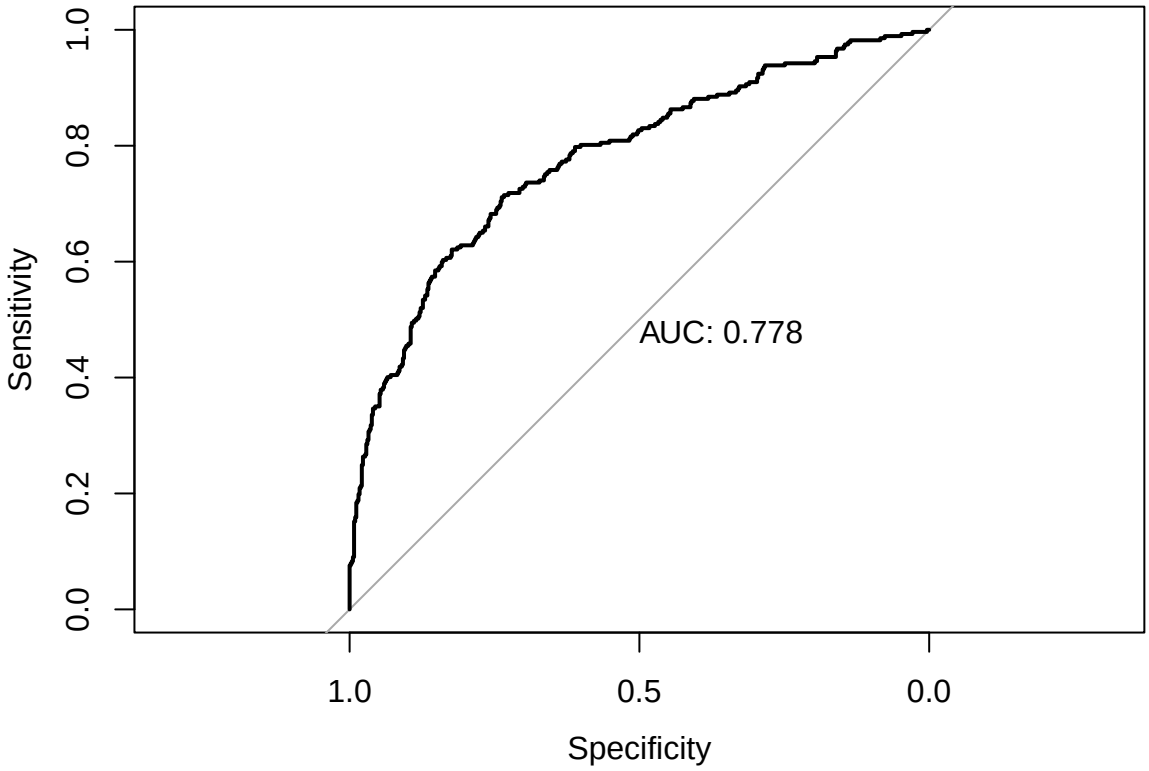

Figure 7: ROC-AUC Curve for logisitic regression model

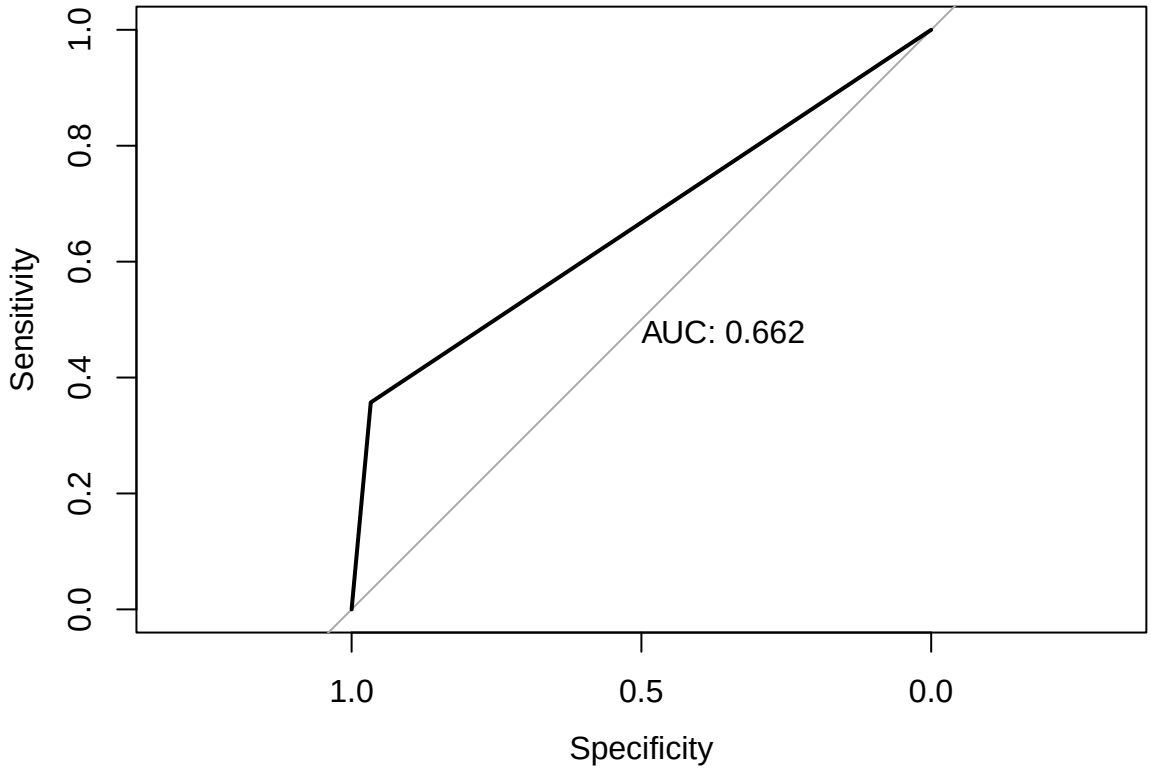

Figure 8: ROC-AUC Curve for random forest model

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)

library(readr)
library(tidyverse)
library(caret)
library(dslabs)
library(vip)
library(randomForest)
library(mlr)
library(pROC)
library(tableone)
# Load ICU data from MIMIC-IV
ICU_df <- read_csv("ICU_df.csv")
# define the control and case subjects
control <- ICU_df %>% filter(mort_icu == 0)
case <-  ICU_df %>% filter(mort_icu == 1)

#downsample dataset
control <- sample_n(control, 13668)
ICU_df <-rbind(control,case)
# Generate pie chart of proportion of genders included in data
ICU_df |>
  group_by(gender) |>
```

```
  count() |>
  ggplot(aes(x = "", y = n, fill = gender)) +
  geom_col() +
  coord_polar(theta ="y") +
  theme(axis.text = element_blank(),
        axis.ticks = element_blank(),
        axis.title = element_blank(),
        panel.grid = element_blank(),
        panel.background = element_rect(fill = "#ebf2ff"),
        plot.background = element_rect(fill = "#ebf2ff"),
        legend.background = element_rect(fill = "#ebf2ff"))

ICU_age <- ICU_df |>
  mutate(age_group = cut(ICU_df$admission_age, c(18, 29, 39, 49, 59, 69, 79, 90, Inf), c("18-29","30-39

ICU_age |>
  group_by(age_group) |>
  count() |>
  ggplot(aes(x = "", y = n, fill = age_group)) +
  geom_col() +
  coord_polar(theta ="y") +
  theme(axis.text = element_blank(),
        axis.ticks = element_blank(),
        axis.title = element_blank(),
        panel.grid = element_blank(),
        panel.background = element_rect(fill = "#ebf2ff"),
        plot.background = element_rect(fill = "#ebf2ff"),
        legend.background = element_rect(fill = "#ebf2ff"))
demo_pie <- ICU_df |>
  group_by(race) |>
  count() |>
  ggplot(aes(x = "", y=n, fill = race)) +
  geom_col() +
  coord_polar(theta ="y") +
  theme(axis.text = element_blank(),
        axis.ticks = element_blank(),
        axis.title = element_blank(),
        panel.grid = element_blank(),
        panel.background = element_rect(fill = "#ebf2ff"),
        plot.background = element_rect(fill = "#ebf2ff"),
        legend.background = element_rect(fill = "#ebf2ff"),
        legend.position = "bottom",
        legend.text = element_text(size=5),
        legend.key.size = unit(0.2, "cm"))
demo_pie
knitr::include_graphics("FlowChart.png")
myVars <- c("admission_age", "gender", "los_icu", "aniongap_min", "aniongap_max", "albumin_min", "bicarb

carVars <- c("gender")

table <- CreateTableOne(vars= myVars, data = ICU_df, factorVars = carVars, strata = "mort_icu")
table1 <- print(table, noSpaces=TRUE, printToggle = FALSE)
#write.csv(table1, file = "TableOne.csv")
```

```
TableOne <- read_csv("TableOne.csv")
knitr::kable(as.data.frame(TableOne|>select(-p, - test)|>rename(" " = "...1")),
      col.names = c("Feature", "Survival", "Mortality"),
      caption = "Overview of MIMIC-IV data")
Joined_SQL <- read_csv("df_labs.csv")
print(Joined_SQL)
print(as.tibble(ICU_df))
set.seed(2007)
y <- ICU_df$mort_icu
test_index <- createDataPartition(y, times = 1, p = 0.2, list = FALSE)
test_set <- ICU_df[test_index,]
train_set <- ICU_df[-test_index,]
# Define model predictors 'x' and outcomes 'y'
y_train <- as.factor(train_set$mort_icu)
y_test <- as.factor(test_set$mort_icu)

x <- ICU_df |>
  select(aniongap_min, aniongap_max, albumin_min, bicarbonate_min, bicarbonate_max, bilirubin_min, bili

x_train <- train_set |>
  select(aniongap_min, aniongap_max, albumin_min, bicarbonate_min, bicarbonate_max, bilirubin_min, bili
x_train_impute <- impute(x_train, classes = list(numeric = imputeMedian()))

x_test <- test_set |>
  select(aniongap_min, aniongap_max, albumin_min, bicarbonate_min, bicarbonate_max, bilirubin_min, bili
x_test_impute <- impute(x_test, classes = list(numeric = imputeMedian()))
# Generate logistic regression model
glm_fit <- glm(mort_icu ~ aniongap_min + aniongap_max + albumin_min + bicarbonate_min + bicarbonate_max
                     data = train_set,
                     family = "binomial")
# Generate random forest mmodel
control <- trainControl(method="cv", number=5)
grid <- data.frame(mtry = c(1, 5, 10, 25, 30, 35, 40))
train_rf <- caret::train(x_train_impute$data, y_train,
                  method = "rf",
                  ntree = 150,
                  trControl = control,
                  tuneGrid = grid,
                  nSamp = 5000)

fit_rf <- randomForest(x_train_impute$data, y_train, mtry=train_rf$bestTune$mtry)
# Evaluate both models using confusion matrix
p_hat_logit <- predict(glm_fit, test_set, type = "response")
y_hat_logit <- factor(ifelse(p_hat_logit > 0.5, 1, 0))

cm_log <- confusionMatrix(as.factor(test_set$mort_icu), y_hat_logit)
log_Accuracy <- cm_log$overall[["Accuracy"]]
log_Sensitivity<- cm_log$byClass["Sensitivity"]
log_Specificity <- cm_log$byClass["Specificity"]
log_Prevalence <- cm_log$byClass["Prevalence"]

y_hat_rf <- predict(fit_rf, x_test_impute$data)
```

```r
cm_rf <- confusionMatrix(y_hat_rf, y_test)
rf_Accuracy <- cm_rf$overall[["Accuracy"]]
rf_Sensitivity<- cm_rf$byClass["Sensitivity"]
rf_Specificity <- cm_rf$byClass["Specificity"]
rf_Prevalence <- cm_rf$byClass["Prevalence"]

Model <- c("Random Forest", "Logistic Regression")
Accuracy <- c(rf_Accuracy, log_Accuracy)
Sensitivity <- c(rf_Sensitivity, log_Sensitivity)
Specificity <- c(rf_Specificity, log_Specificity)
Prevalence <- c(rf_Prevalence, log_Prevalence)
df <- data.frame(Model, Accuracy, Sensitivity, Specificity, Prevalence)
knitr::kable(df,
      col.names = c("Model", "Accuracy", "Sensitivity", "Specificity", "Prevalence"),
      caption = "Evalutation Metrics for Predictive Models")
# Plot variable importance of logisitic regression model
vip(glm_fit)
# Plot variable importance of random forest model
vip(fit_rf)
# Plot ROC-AUC Curve for logisitc regression model
test_roc = roc(test_set$mort_icu ~ p_hat_logit, plot = TRUE, print.auc = TRUE)
# Plot ROC-AUC Curve for random forest model
y_hat_rf_prob <- predict(fit_rf, x_test_impute$data, type = "response")
test_roc = roc(y_test ~ as.numeric(y_hat_rf_prob), plot = TRUE, print.auc = TRUE)
```