# ICU Analysis

John Rand

1/5/2023

## Problem Statement:

Not all patients admitted to an ICU belong in the ICU. The criteria for which patients belong in an ICU are typically subjective. It would be helpful to have a predictive model that identifies highest-risk patients during intake using factors that the medical staff can quickly access (without lengthy or complex tests). One way to identify the key factors for triaging ICU patients is to determine which factors have the highest association to mortality outcomes in the ICU and, therefore, indicate serious conditions.

## Data Description:

The dataset for 200 ICU patients was a subset obtained from Hosmer D.W., Lemeshow, S., and Sturdivant, R.X. Applied Logistic Regression. 3rd ed., 2013. Accessible from http://www.umass.edu/statdata/statdata/data/icu.txt./ It contains the following variables:
- vital.status: categorical outcome of whether the patient survived until hospital discharge (lived), or did not (died)
- age: age, measured in years
- gender: gender, either male or female
– race: either white, black, or other
– type: type of admission, either elective or emergency
– service: the type of service the patient needed upon ICU admission, either medical or surgical
– conscious: level of consciousness at ICU admission, either no coma/stupor, deep stupor, or coma
– cancer: coded yes if cancer was part of the present problem, no if otherwise
– renal: coded yes if the patient had a history of chronic renal failure, no if otherwise
– infect.prob: yes if infection was probable, no if otherwise
– cpr: yes if CPR was administered prior to ICU admission, no if otherwise
– sys: measured in mm Hg. Typical systolic blood pressure ranges from 90 - 120 mm Hg.
– hr: measured in beats/min. Typical resting heart rate ranges from 60 - 100 bpm.
– previous: yes if previously admitted to an ICU within 6 months, no if otherwise
– fracture: coded yes if patient had a long bone, multiple, neck, single area, or hip fracture; no if otherwise
– creat: creatinine levels, measured in mg/dL. Typical ranges are 0.5 - 1.0 mg/dL. Elevated creatinine levels may be a sign of renal failure.
– PO2: oxygen partial pressure, measured in mm Hg. Normal arterial oxygen concentration is between 75-100 mm Hg, levels below 60 require supplemental oxygen.
– PH: normal blood pH is typically between 7.35 and 7.45. Low blood pH is indicative of acidosis, which can have serious consequences.
– PCO2: carbon dioxide partial pressure, measured in mm Hg. Normal arterial CO2 concentration is between 35-45 mm Hg. Values higher than 45 mm Hg is indicative of respiratory failure.
– bicarb: bicarbonate level, measured in mEq/L. Low bicarbonate levels are indicative of metabolic acidosis.

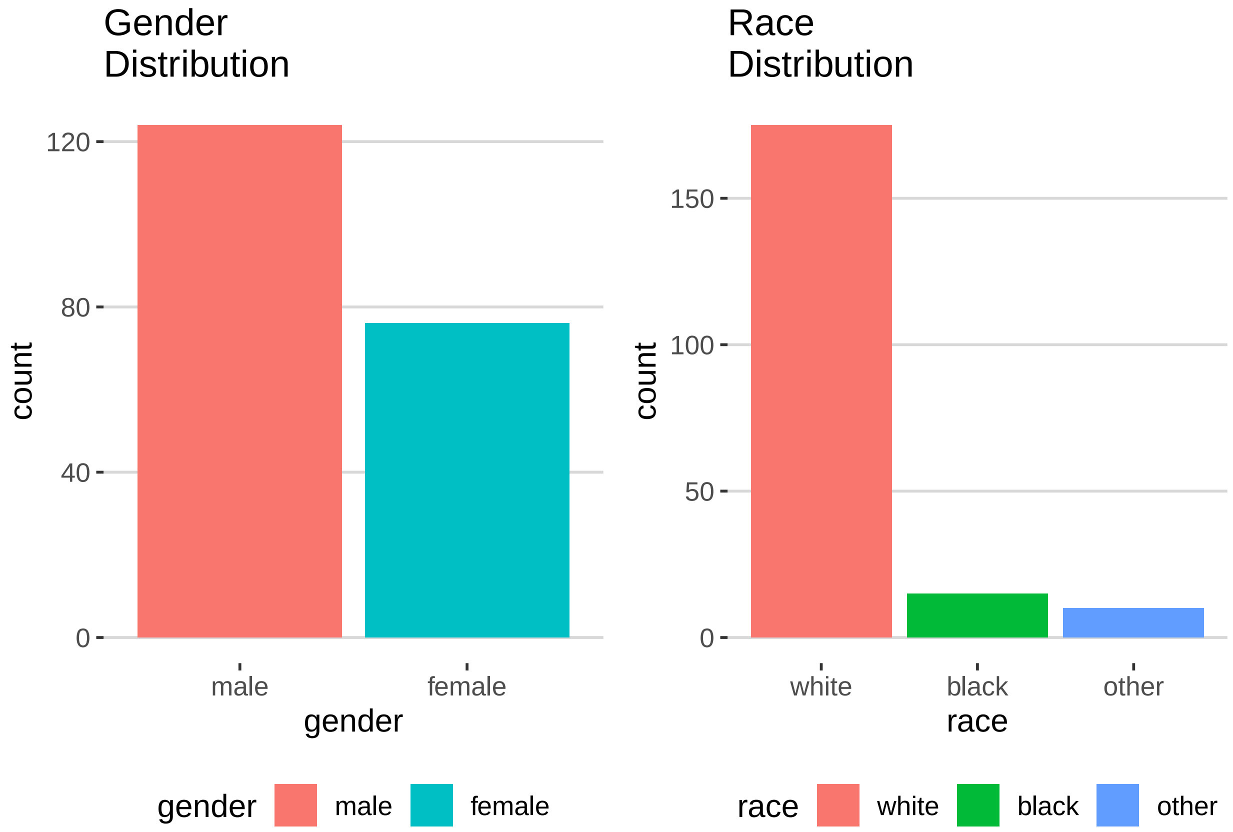
Note that columns P02, PH, PC02, and bicarb contain data from an arterial blood gas (ABG) test, which measures the amount of oxygen and carbon dioxide in the blood and also checks the blood's acid-base balance. Excess acidity can indicate severe conditions, such as kidney failure, severe infection, toxicity, or complications from diabetes. The test does not require any special preparation and takes 5-10 min. Test results are available within 15 minutes.

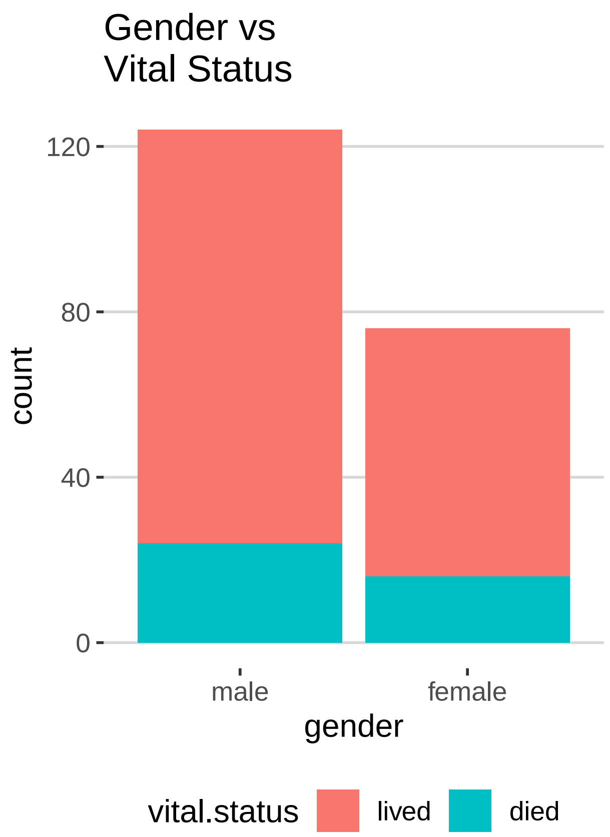
**First, let's check if the dataset has any missing values. The presence of missing values or incorrect data types and formats would require data cleaning prior to modeling.**

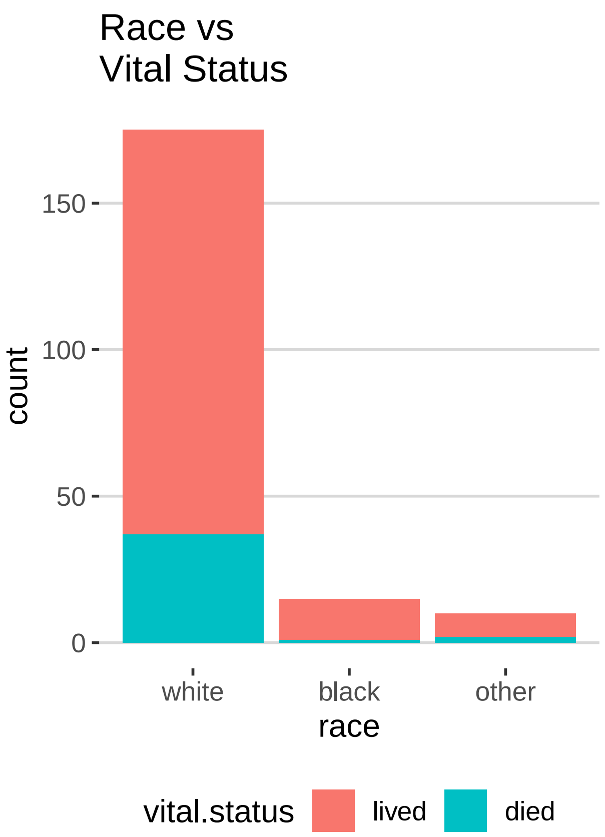
```
## integer(0)
```

**The dataset has no missing values.**

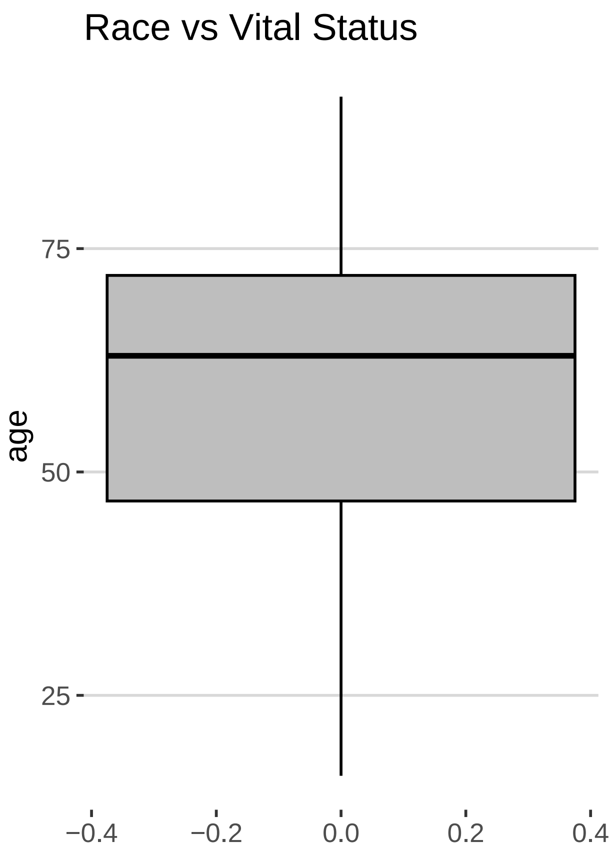
**Next, let's explore the ICU study population demographics:**

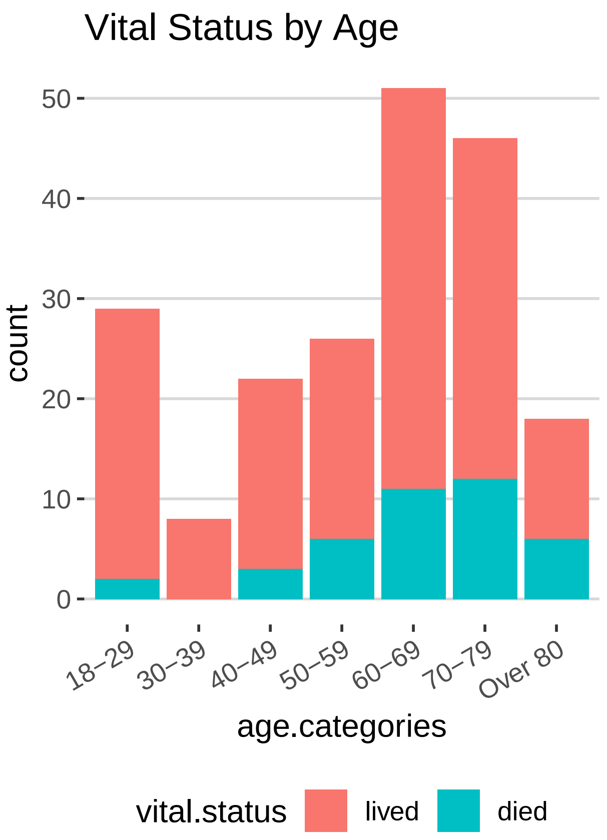


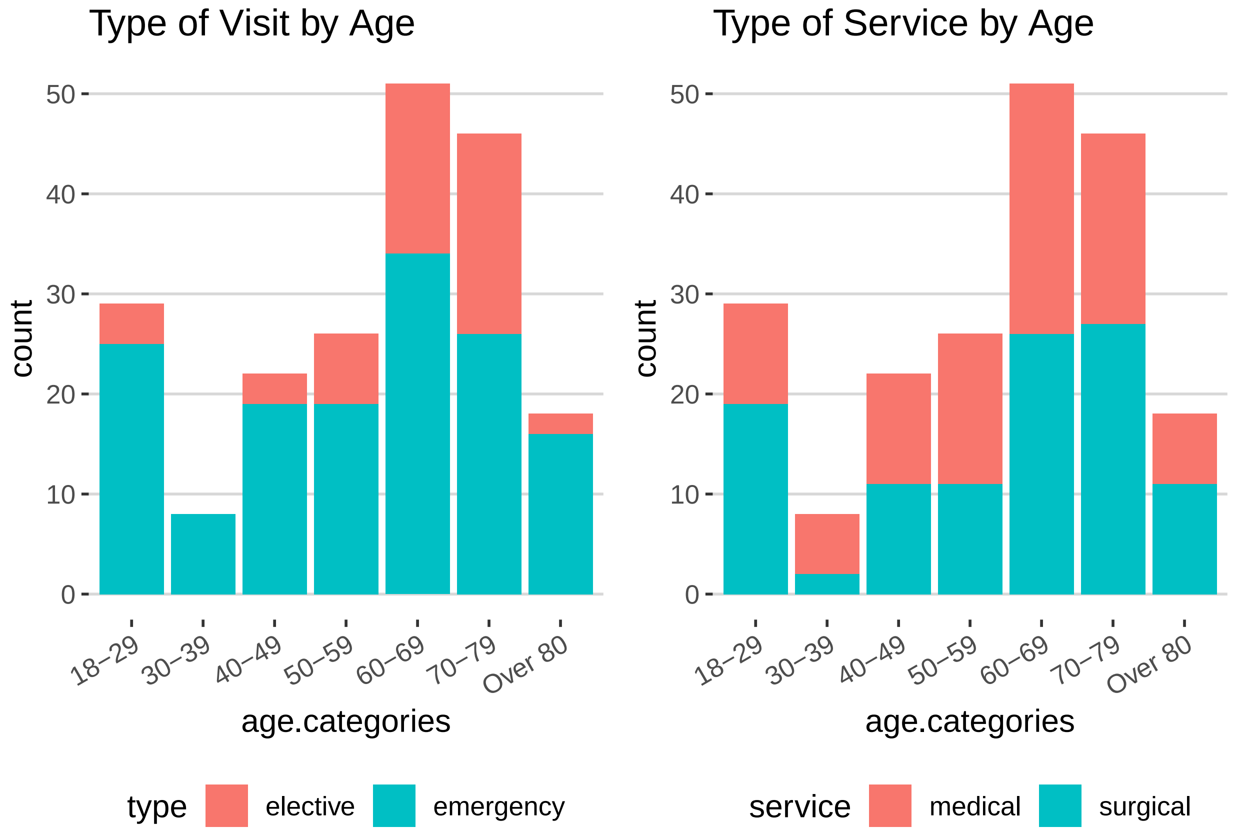




The data sample of 200 ICU patients included more males (62%, 124), and primarily consisted of white patients (87.5%, 175). Only 7.5% (15) of the population were African American, and 5% (10) were other races. Overall, the survival rate was 80% (160). Among males the survival rate was The median age of patients is 63, the mean is 57, with an age range of (16-92). 50% of patients are ages 49-70, 25% of the patients are ages 70-88 (upper quartile), and 25% of the patients are ages 18-49 (lower quartile).


Race vs Vital Status
age
75
50
25
−0.4
−0.2
0.0
0.2
0.4


Vital Status by Age
count
50
40
30
20
10
0
18−29
30−39
40−49
50−59
60−69
70−79
Over 80
age.categories
vital.status
lived
died



Roughly 75% of ICU visits are classified as emergency visits, as compared to 25% of elective visits. However, predominantly younger patients of ages 18-49 (89%, 59) were in the ICU for emergency reasons. Relatively fewer patients of ages 50-Over 80 (68%, 141) were admitted for emergency reasons, and 32% had elective visits. About the same number of patients received medical interventions (46%) as surgical interventions, regardless of age. However, more deaths occurred for the older patients (50-Over 80) than for the younger patients (18-49).

### Correlation test on ABG variables to determine if they are correlated with vital.status

```
##        P-Values
## PO2      0.2430
## PH       0.3179
## PCO2     1.0000
## bicarb   0.1812
```

### Linear regression on the variables that the stepwise function selected influence a patient's vital status significantly

```
##
## Call:
## lm(formula = vital.status ~ cancer + sys + previous + type +
##     conscious + age.categories, data = icu)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94306 -0.21744 -0.06961  0.04308  0.98607
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.0034074  0.2565169   0.013 0.989415
## cancer          0.2111531  0.0880962   2.397 0.017491 *
## sys            -0.0012160  0.0007562  -1.608 0.109465
## previous        0.1029369  0.0690904   1.490 0.137886
## type            0.2477955  0.0614693   4.031 7.98e-05 ***
## conscious       0.3349251  0.0545409   6.141 4.59e-09 ***
## age.categories  0.0458910  0.0133814   3.429 0.000739 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3406 on 193 degrees of freedom
## Multiple R-squared:  0.3002, Adjusted R-squared:  0.2784
## F-statistic:  13.8 on 6 and 193 DF,  p-value: 4.993e-13
```

### Linear regression of the variables the paper noted as the most crucial to predicting a patients vital status

```
##
## Call:
## lm(formula = vital.status ~ cancer + sys + type + conscious +
##     age.categories + cpr + infect.prob, data = icu)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93410 -0.20867 -0.06701  0.03707  0.98827
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.0858933  0.2477778   0.347 0.729231
## cancer          0.1828261  0.0873715   2.093 0.037707 *
## sys            -0.0011275  0.0007752  -1.454 0.147462
## type            0.2253627  0.0625508   3.603 0.000401 ***
## conscious       0.3154037  0.0582752   5.412 1.84e-07 ***
## age.categories  0.0447249  0.0136738   3.271 0.001271 **
## cpr             0.0543762  0.1072966   0.507 0.612888
## infect.prob     0.0446454  0.0520289   0.858 0.391914
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3425 on 192 degrees of freedom
## Multiple R-squared:  0.2961, Adjusted R-squared:  0.2705
## F-statistic: 11.54 on 7 and 192 DF,  p-value: 3.302e-12
```

The variables that the stepwise function selected had a higher R-Squared value than the variables from the paper when running a linear regression model. Indicating that on this subset of the data, the variables cancer, sys, previous, type, conscious and age groups make a better model than the variables the paper selected as most important.

### Random forest classification used to generate predictive model of a patient's vital status

```
##
## Call:
##  randomForest(formula = vital.status ~ ., data = train, proximity = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 16.3%
## Confusion matrix:
##     1  2 class.error
## 1 103  5   0.0462963
## 2  17 10   0.6296296
```

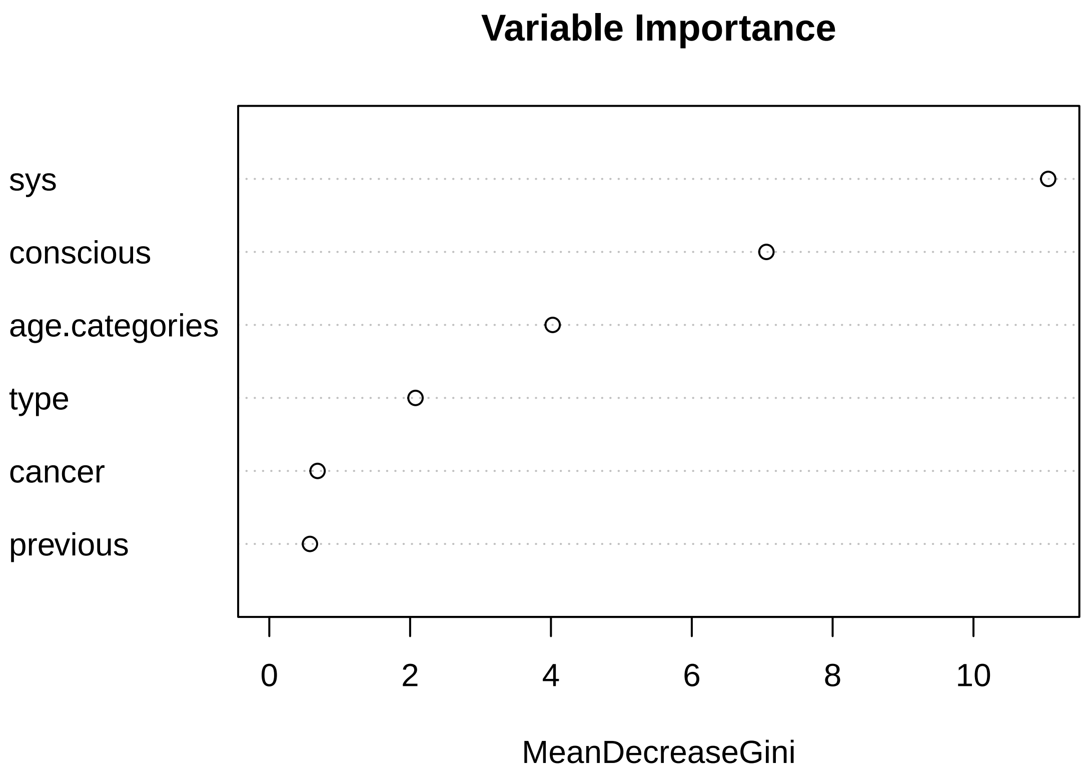
### Accuracy of random forest classification on train data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2
##          1 108  13
##          2   0  14
##
##                Accuracy : 0.9037
##                  95% CI : (0.841, 0.9477)
##     No Information Rate : 0.8
##     P-Value [Acc > NIR] : 0.0008992
##
##                   Kappa : 0.6328
##
##  Mcnemar's Test P-Value : 0.0008741
##
##             Sensitivity : 1.0000
##             Specificity : 0.5185
##          Pos Pred Value : 0.8926
##          Neg Pred Value : 1.0000
##              Prevalence : 0.8000
##          Detection Rate : 0.8000
##    Detection Prevalence : 0.8963
##       Balanced Accuracy : 0.7593
##
##        'Positive' Class : 1
##
```

### Accuracy of random forest classification on test data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2
##          1 52  9
##          2  0  4
##
```

```
##                Accuracy : 0.8615
##                  95% CI : (0.7534, 0.9347)
##     No Information Rate : 0.8
##     P-Value [Acc > NIR] : 0.137172
##
##                   Kappa : 0.4156
##
##  Mcnemar’s Test P-Value : 0.007661
##
##             Sensitivity : 1.0000
##             Specificity : 0.3077
##          Pos Pred Value : 0.8525
##          Neg Pred Value : 1.0000
##              Prevalence : 0.8000
##          Detection Rate : 0.8000
##    Detection Prevalence : 0.9385
##       Balanced Accuracy : 0.6538
##
##        ’Positive’ Class : 1
##
```



### Random forest classification on variables from original report

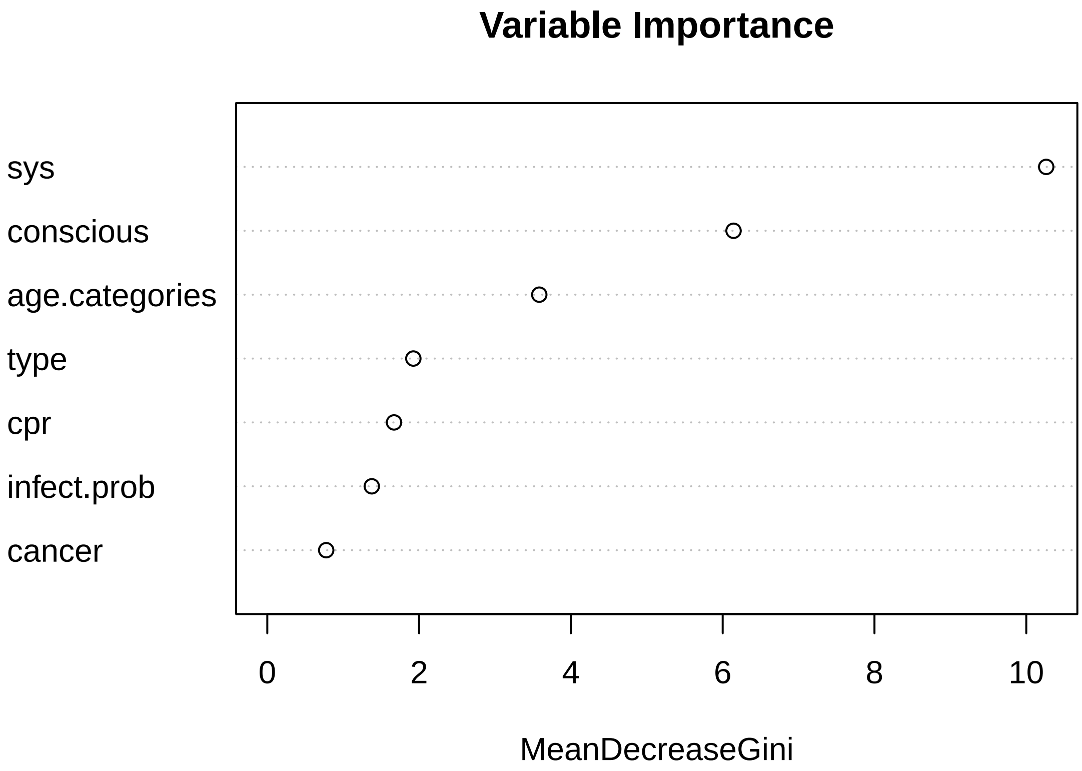
```
##
## Call:
##  randomForest(formula = vital.status ~ ., data = train.2, proximity = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 19.26%
## Confusion matrix:
##     1 2 class.error
## 1 101 7  0.06481481
## 2  19 8  0.70370370
```

### Accuracy of random forest classification on train data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2
##          1 108  13
##          2   0  14
##
##                Accuracy : 0.9037
##                  95% CI : (0.841, 0.9477)
##     No Information Rate : 0.8
##     P-Value [Acc > NIR] : 0.0008992
##
##                   Kappa : 0.6328
##
##  Mcnemar’s Test P-Value : 0.0008741
##
##             Sensitivity : 1.0000
##             Specificity : 0.5185
##          Pos Pred Value : 0.8926
##          Neg Pred Value : 1.0000
##              Prevalence : 0.8000
##          Detection Rate : 0.8000
##    Detection Prevalence : 0.8963
##       Balanced Accuracy : 0.7593
##
##        ’Positive’ Class : 1
##
```

### Accuracy of random forest classification on test data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2
##          1 52 10
##          2  0  3
##
##                Accuracy : 0.8462
##                  95% CI : (0.7352, 0.9237)
##     No Information Rate : 0.8
##     P-Value [Acc > NIR] : 0.222902
##
##                   Kappa : 0.3243
##
##  Mcnemar's Test P-Value : 0.004427
##
##             Sensitivity : 1.0000
##             Specificity : 0.2308
##          Pos Pred Value : 0.8387
##          Neg Pred Value : 1.0000
##              Prevalence : 0.8000
##          Detection Rate : 0.8000
##    Detection Prevalence : 0.9538
##       Balanced Accuracy : 0.6154
##
##        'Positive' Class : 1
##
```



## Insights

From the paper this report was based on, there was some slight differences in what was considered most important when determining a patient's vital status. This was done using a stepwise regression with vital.status as the dependant variable. The stepwise regression noted 6 variables of interest when determining a patient's vital status, those being systolic blood pressure, conscious, age group, type of visit, cancer, and if the patient was admitted within the past 6 months.

Running a random forest classification on a subset of the data with those 6 variables gave about an 86% accuracy when determining a patient's vital status.

The variables the paper noted as most important when classiying a patient's vital status were systolic blood pressure, conscious, age group, type of visit, cpr prior to admission, probable infection, and cancer.

Running a random forest classification on the variables the paper noted as most important on the same subset of the data gave about an 85% accuracy when determining a patient's vital status.

Between the two tests there is a slight difference in the overall accuracy. It's interesting to note that on a subset of the data the variables best for determining a patient's vital status differ from the orginal paper. This could be due to the fact that it is a subset of the data and note the whole sample size.