R Forensics

November 9, 2020

0.1 Installing/Loading Necessary Packages

```
In [3]: library(tidyverse)
        install.packages("cowplot")
        library(cowplot)
```

```
Warning message:
"package 'cowplot' is in use and will not be installed"
```

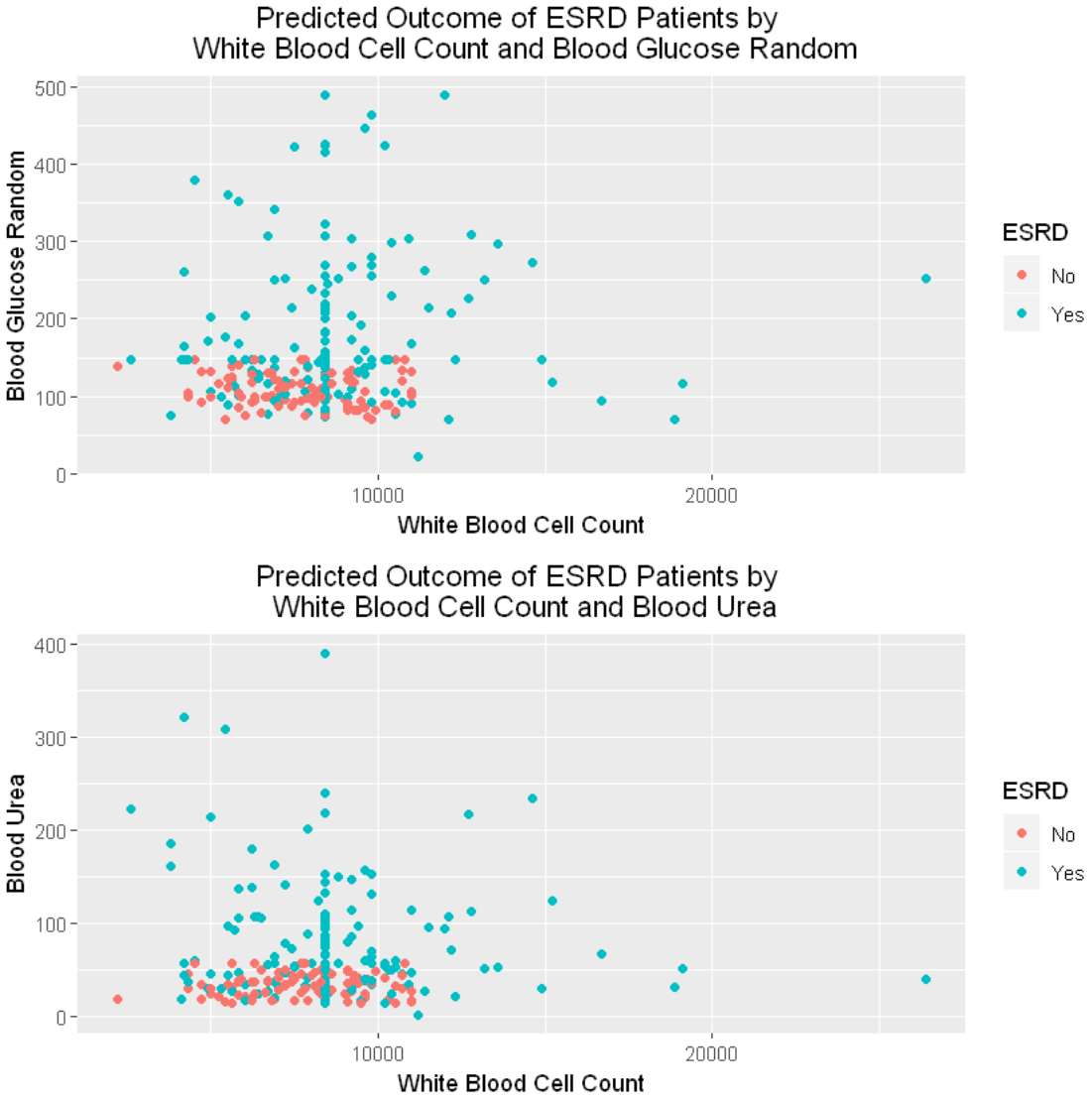
0.1.1 Obtaining the data and additional objects to be used later

```
In [8]: output <- read.csv("output.csv")

        facet_labels <- c("No", "Yes")
        names(facet_labels) <- c(0, 1)
        pred_out <- output %>% group_by(age, prediction) %>%
                    summarize(count=sum(prediction))
        class_out <- output %>% group_by(age, classification) %>%
                     summarize(count=sum(classification))
```

Based on the output from the Python code we see the most important factors on the dataset. Taking a look at the relationship among the top three we can confirm assumptions that higher White Blood Cell Count and Higher Blood Glucse Random or Blood Urea make it to be higher chances of having ESRD. There are still cases in the lower ranges but no patient without ESRD was seen with the higher levels as seen below.

```
In [6]: glucose <- output %>%
          ggplot(aes(x=wc, y=bgr, color=factor(prediction))) +
          geom_point() +
          ylab("Blood Glucose Random") +
          xlab("White Blood Cell Count") +
          labs(title="Predicted Outcome of ESRD Patients by \n White Blood Cell Count and Blooc
          theme(plot.title=element_text(hjust=0.5)) +
          scale_color_discrete(name="ESRD", labels=c("No", "Yes"))

        urea <- output %>%
          ggplot(aes(x=wc, y=bu, color=factor(prediction))) +
          geom_point() +
          ylab("Blood Urea") +
```

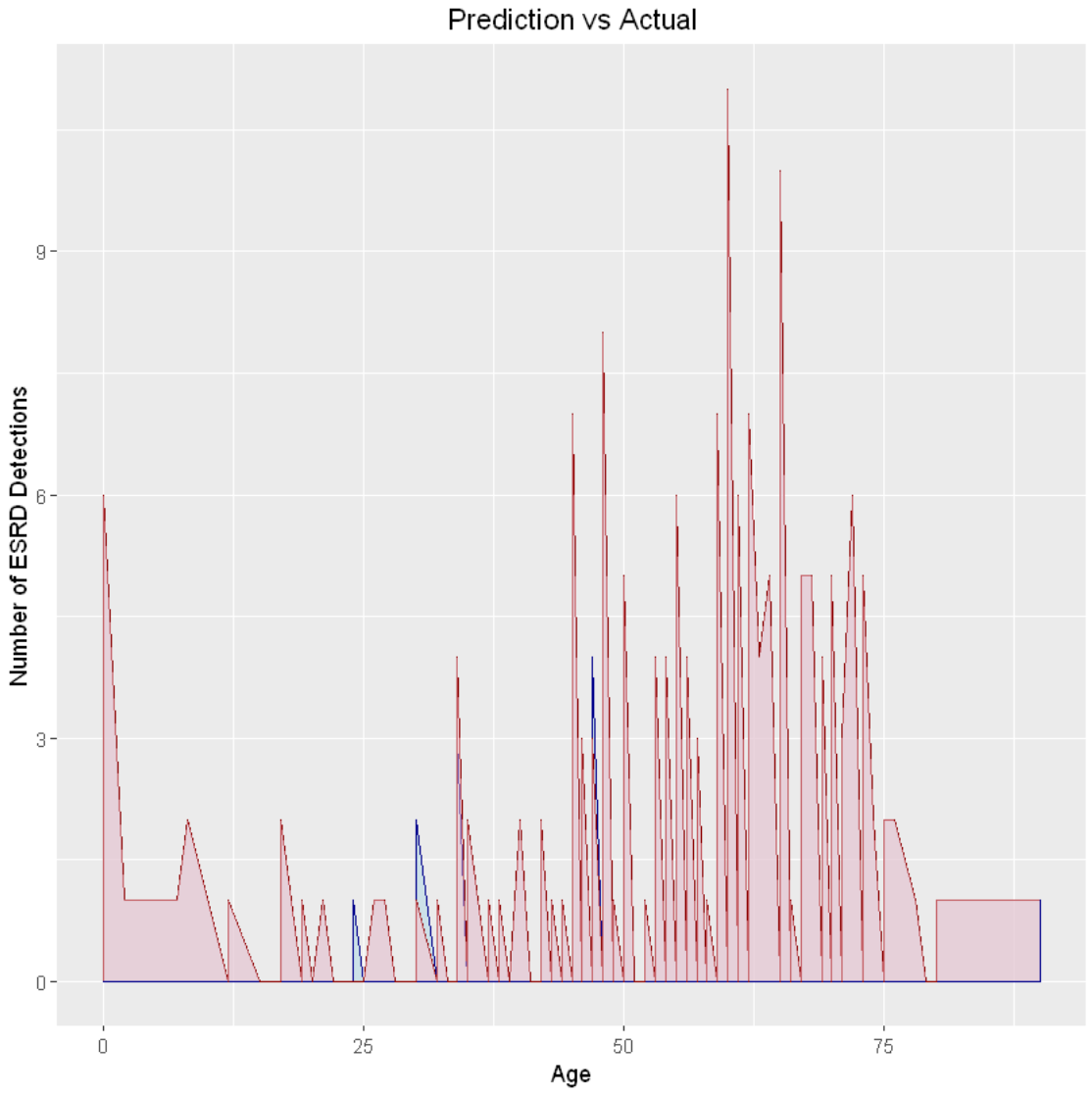
```
    xlab("White Blood Cell Count") +
    labs(title="Predicted Outcome of ESRD Patients by \n White Blood Cell Count and Blood
    theme(plot.title=element_text(hjust=0.5)) +
    scale_color_discrete(name="ESRD", labels=c("No", "Yes"))

plot_grid(glucose, urea, ncol=1)
```



The AI model consistently performed in the 98%-99% Accuracy with an average ROC AUC score of 97.92%. We can see the comparison of predictions vs the actual classifications, showing the small blue parts where the model was incorrect.
We can also see that the highest grouping of patients with ESRD can be found between the ages of 40 and 75.

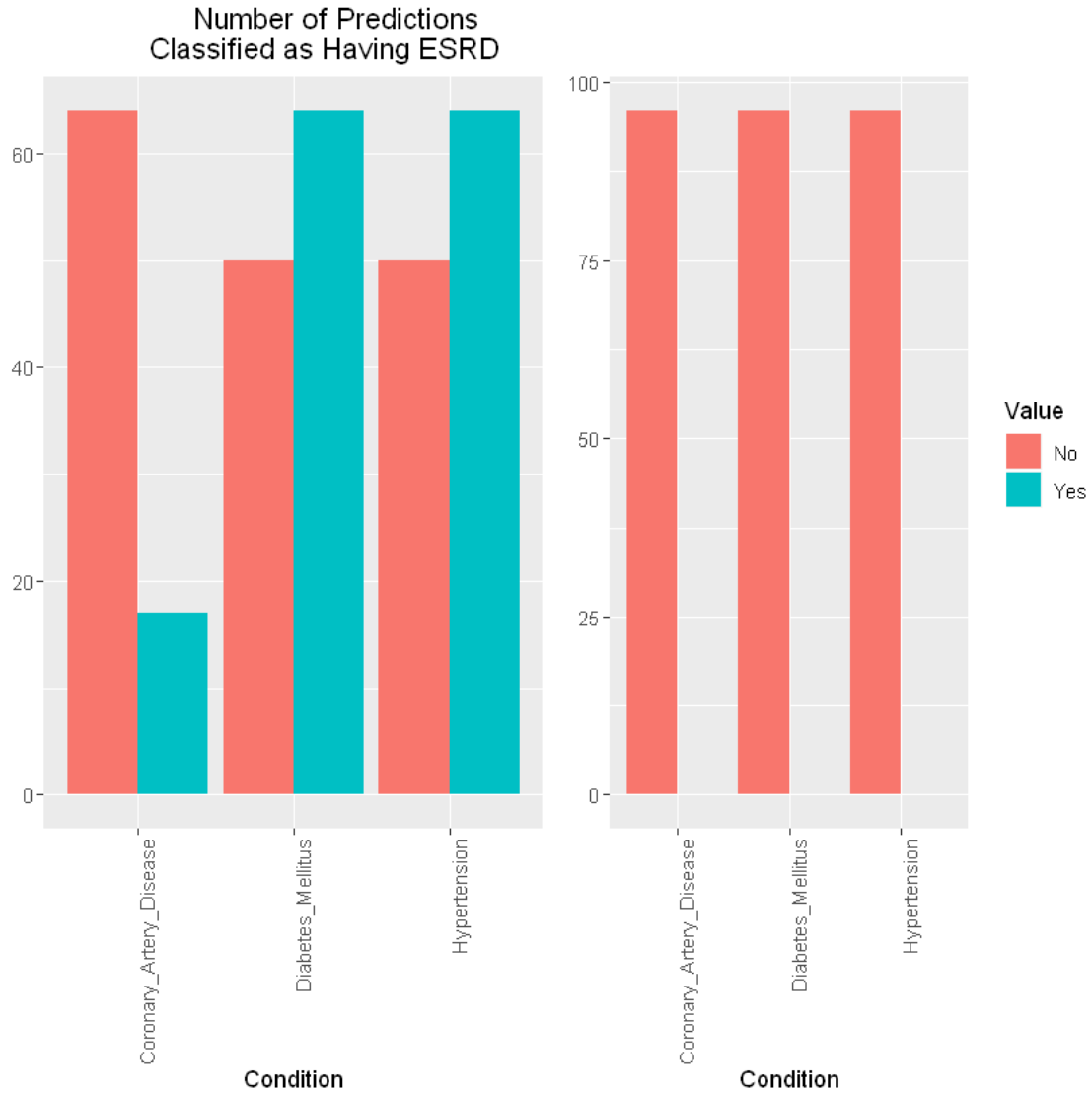
```
In [9]: ggplot(pred_out, aes(x=age, y=count)) +
          geom_line() +
```

```
geom_area(color="darkblue",
          fill="lightblue",
          alpha = 0.5) +
geom_line(class_out, mapping=aes(x=age, y=count), color="darkred") +
geom_area(class_out, mapping=aes(x=age, y=count), fill="pink", alpha=0.5)+
labs(title="Prediction vs Actual") +
xlab("Age") +
ylab("Number of ESRD Detections") +
theme(plot.title=element_text(hjust=0.5))
```



With Hypertension, Diabetes, and Heart Disease being some of the leading causes of ESRD. We can expect seeing that patients not having these conditions are not diagnosed with ESRD. Which with the opposite we see that patients with these conditions have a higher possibility to be diagnosed with ESRD.

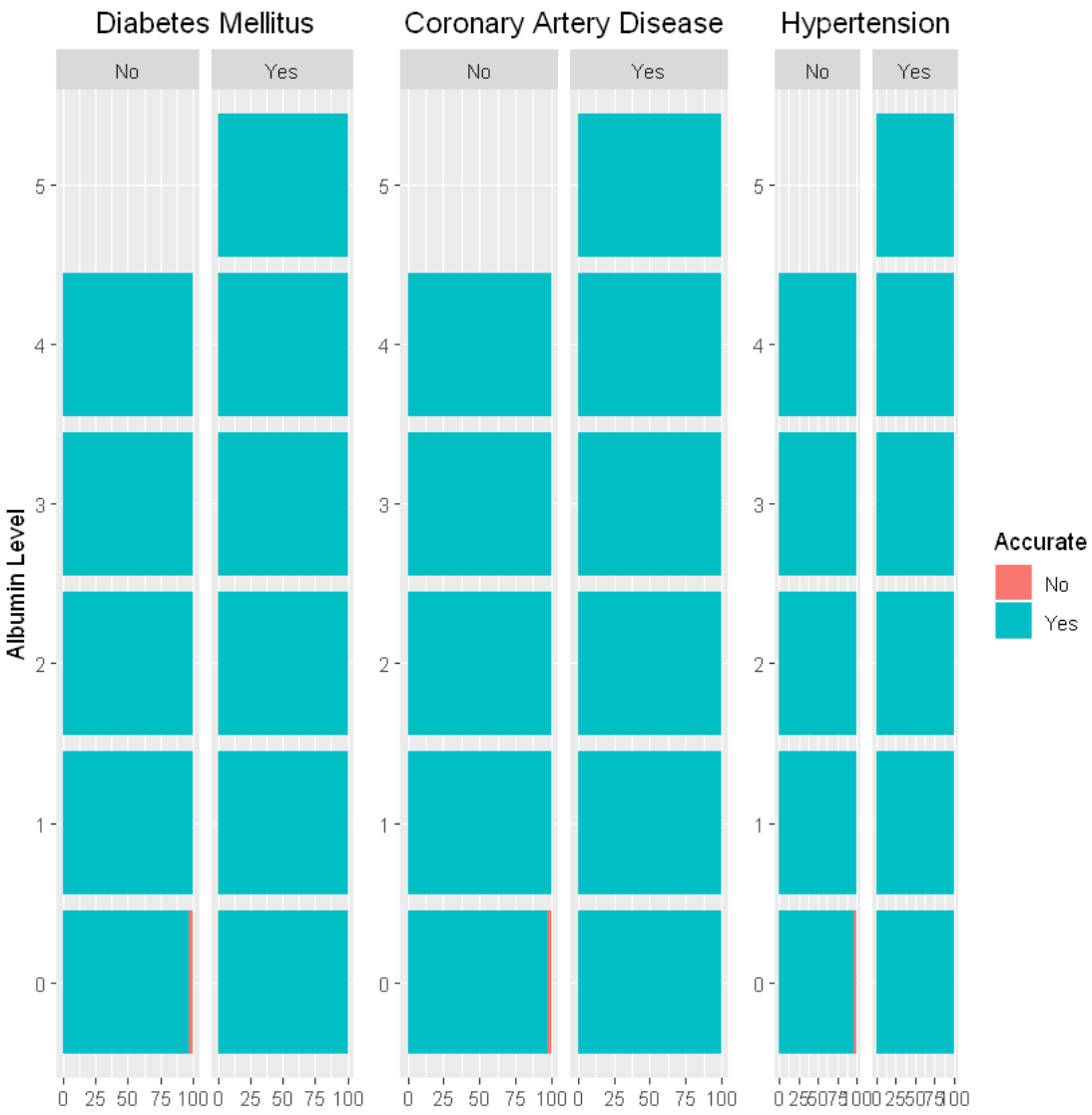
In [20]:
```
esrdyes <- output %>%
  group_by(htn, dm, cad) %>%
  summarize(esrd_yes=sum(prediction), esrd_no=n()-esrd_yes) %>%
  mutate(Hypertension=ifelse(htn==0, "No", "Yes"),
         Diabetes_Mellitus=ifelse(dm==0, "No", "Yes"),
         Coronary_Artery_Disease=ifelse(cad==0, "No", "Yes")) %>%
  gather("Condition", "Value", 6:8) %>%
  ggplot(aes(x=Condition, y=esrd_yes, fill=Value)) +
  geom_bar(stat="identity", position="dodge") +
  theme(axis.title.y=element_blank(),
        plot.title=element_text(hjust=0.7),
        legend.position = "none",
        axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title="Number of Predictions\n Classified as Having ESRD")

esrdno <- output %>%
  group_by(htn, dm, cad) %>%
  summarize(esrd_yes=sum(prediction), esrd_no=n()-esrd_yes) %>%
  mutate(Hypertension=ifelse(htn==0, "No", "Yes"),
         Diabetes_Mellitus=ifelse(dm==0, "No", "Yes"),
         Coronary_Artery_Disease=ifelse(cad==0, "No", "Yes")) %>%
  gather("Condition", "Value", 6:8) %>%
  ggplot(aes(x=Condition, y=esrd_no, fill=Value)) +
  geom_bar(stat="identity", position="dodge") +
  theme(axis.title.y=element_blank(),
        plot.title=element_text(hjust=0.5),
        axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title="\n")

plot_grid(esrdyes, esrdno, nrow=1)
```



Lastly we can see the accuracy being almost close to 100% on all comparing the Albumin Level vs each of the major medical diseases present.

```
In [21]: htn_plt <- output %>% group_by(al, htn) %>% summarize(yes=(sum(accurate)/n())*100, no=
           ggplot(aes(factor(al), y=percent, fill=accurate)) +
           geom_bar(stat="identity") +
           facet_grid(~htn, labeller = labeller(htn = facet_labels)) +
           coord_flip() +
           labs(title="Hypertension") +
           xlab("Albumin Level") +
           theme(plot.title=element_text(hjust=0.5),
                 axis.title.x=element_blank(),
                 axis.title.y=element_blank()) +
           scale_fill_discrete(name="Accurate", labels=c("No", "Yes"))
```

```
dm_plt <- output %>% group_by(al, dm) %>% summarize(yes=(sum(accurate)/n())*100, no=1
  ggplot(aes(factor(al), y=percent, fill=accurate)) +
  geom_bar(stat="identity") +
  facet_grid(~dm, labeller = labeller(dm = facet_labels)) +
  coord_flip() +
  labs(title="Diabetes Mellitus") +
  xlab("Albumin Level") +
  theme(plot.title=element_text(hjust=0.5),
        axis.title.x=element_blank(),
        legend.position = "none")+
  scale_fill_discrete(name="Accurate", labels=c("No", "Yes"))

cad_plt <- output %>% group_by(al, cad) %>% summarize(yes=(sum(accurate)/n())*100, no=
  ggplot(aes(factor(al), y=percent, fill=accurate)) +
  geom_bar(stat="identity") +
  facet_grid(~cad, labeller = labeller(cad = facet_labels)) +
  coord_flip() +
  labs(title="Coronary Artery Disease") +
  xlab("Albumin Level") +
  theme(plot.title=element_text(hjust=0.5),
        axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        legend.position = "none")+
  scale_fill_discrete(name="Accurate", labels=c("No", "Yes"))

plot_grid(dm_plt,cad_plt, htn_plt, nrow=1)
```


Diabetes Mellitus
Coronary Artery Disease
Hypertension
No
Yes
Albumin Level
0
1
2
3
4
5
0 25 50 75 100
Accurate
No
Yes