Covid Pre-Condition

Rafael Pereira

2/1/2021

Introduction

This algorithm use data to predict if a covid positive patient will need hospitalization or not based in his/her pre-conditions. The data used for this is COVID-19 patient pre-condition dataset (https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset) acquired from the Mexican government. The variable that will be predicted is the type of patient (variable's name: patient_type), 1 for outpatient (a patient who receives medical treatment without being admitted to a hospital) and 2 for inpatient (a patient who's been admitted to hospital for medical treatment).The pre-conditions (predictors) used for this are:sex of the patient (1 for female and 2 for male, variable's name: sex),the age of the patient (variable's name: age). In the next variables 1 indicates that the patient has it and 2 that the patient doesn't have it: pneumonia (variable's name: pneumonia), diabetes(variable's name: diabetes),chronic obstructive pulmonary disease (variable's name: copd),asthma (variable's name: asthma),immunosuppression (variable's name: inmsupr),hypertension (variable's name: hypertension),other diseases (variable's name: other_disease), cardiovascular diseases(variable's name:cardiovascular), obesity (variable's name: obesity),chronic kidney disease (variable's name: renal_chronic) and smoking habits (variable's name: tobacco). Only the covid positive patients will be used for the models(variable's name: covid_res) For the algorithm are used five models: linear discriminant analysis (LDA), generalized linear model (LGM), quadratic discriminant analysis (QDA), classification and regression tree (RPART) and an ensemble of the other four models.

Analisis

The first step is the installation of the libraries that will be used:

```
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")
if(!require(ggplot2)) install.packages("ggplot2", repos = "http://cran.us.r-project.org")
if(!require(gridExtra)) install.packages("gridExtra", repos ="http://cran.us.r-project.org")

library(tidyverse)
library(caret)
library(data.table)
library(ggplot2)
library(gridExtra)
```

The data is loaded from the url in cvs format, after that is read and stored in a data frame:

```
url <- "https://github.com/RPereira98/Covid-Pre-condition/raw/main/covid.zip"
dl <- tempfile()
download.file(url, dl)
```

```
unzip(dl,"covid.csv")
covid_dat <- read_csv("covid.csv")
covid_dat<-data.frame(covid_dat)
```

Only the covid positive patients are important for the models:

```
ind<- which(covid_dat$covid_res==1)
covid_dat<-covid_dat[ind,]
```

Some variables will not be used for the analysis (dates,ID of patients, pregnancy, intubation, ICU):

```
covid_dat<-covid_dat[,-c(1,4,5,6,7,10,21,22,23)]
```

The values 97, 98 and 99 are NAs, not useful:

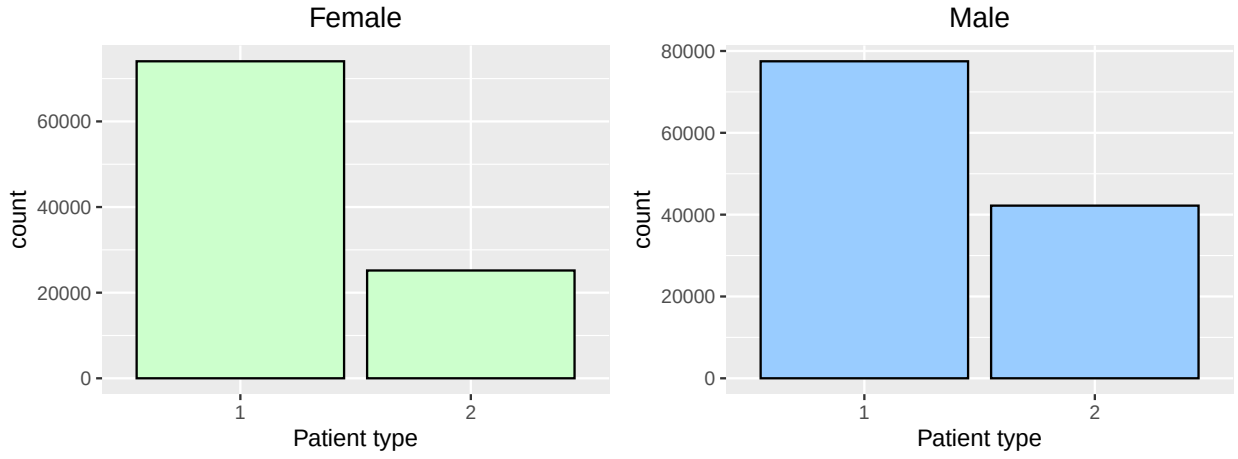
```
ind<- which((covid_dat$pneumonia%in%c(97,98,99))|(covid_dat$diabetes%in%c(97,98,99))|(covid_dat$copd%in%
covid_dat<-covid_dat[-ind,]
```

There are more outpatients than inpatients, the proportion of outpatients is:

```
mean(covid_dat$patient_type==1)
```

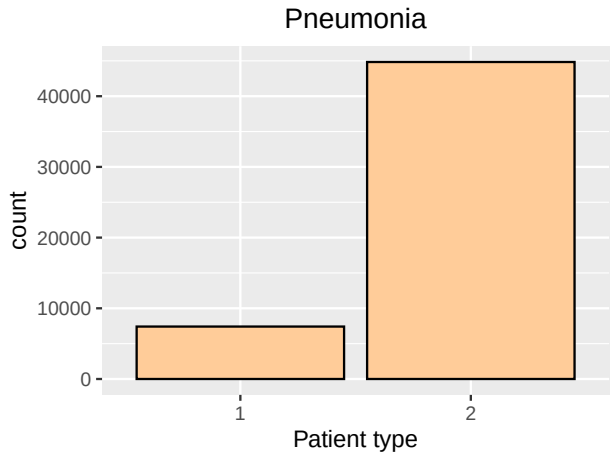
```
## [1] 0.6921728
```

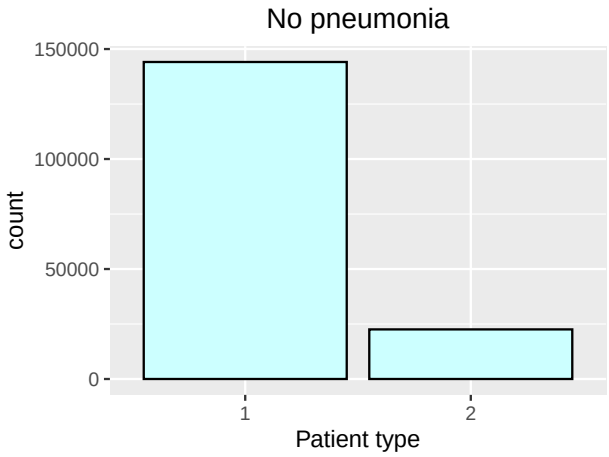
Doing a visual inspection of the data: First parameter to analyze: Sex



Both sexes have more outpatients than inpatients,with males having a bigger proportion of inpatients than females.

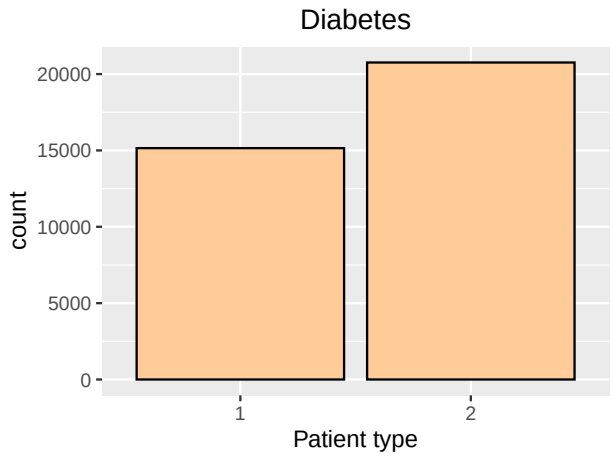
Parameter: Pneumonia

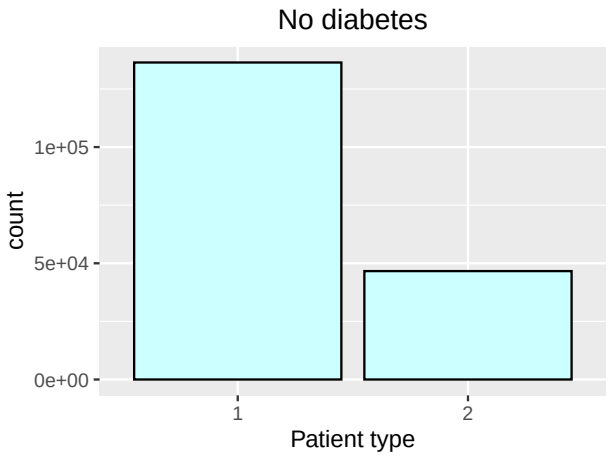




More patients with pneumonia are inpatients than outpatients, and more patients without pneumonia are outpatients than inpatients.

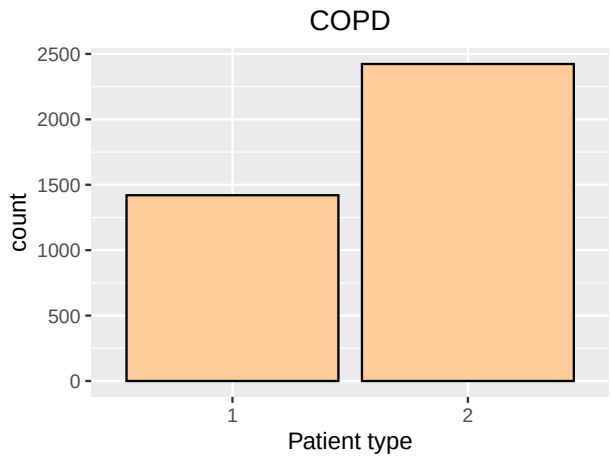
Parameter: Diabetes

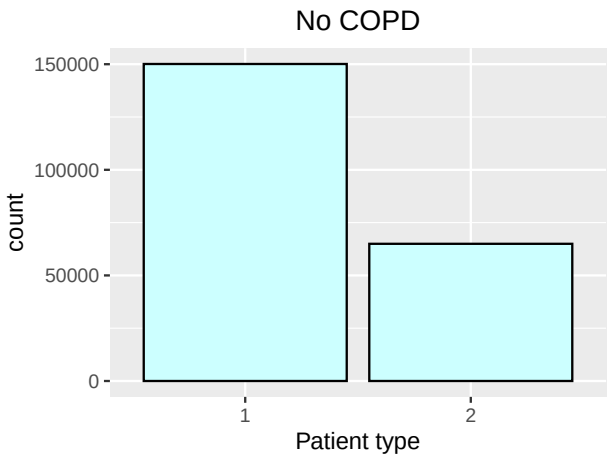




There are slightly more patients with diabetes that are inpatients than outpatients, and there are more patients without diabetes that are outpatients than inpatients.

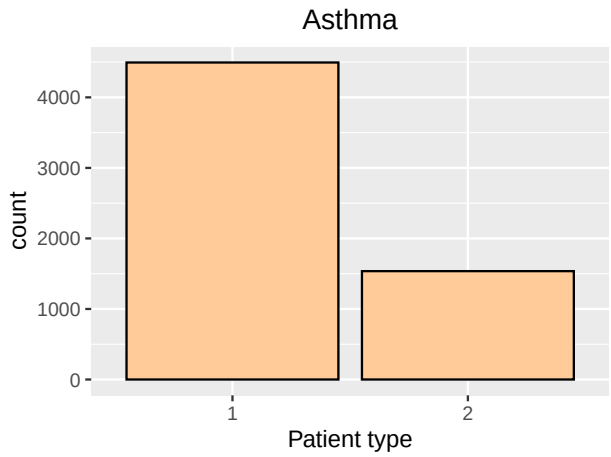
Parameter: Chronic obstructive pulmonary disease

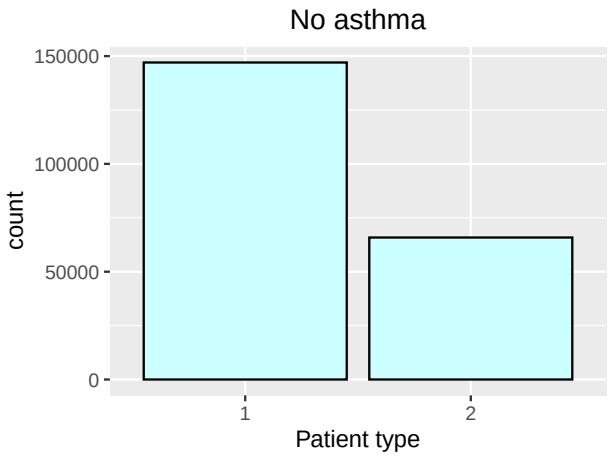




There are more patients with COPD that are inpatients than outpatients and there are more outpatients without COPD than inpatients without COPD.

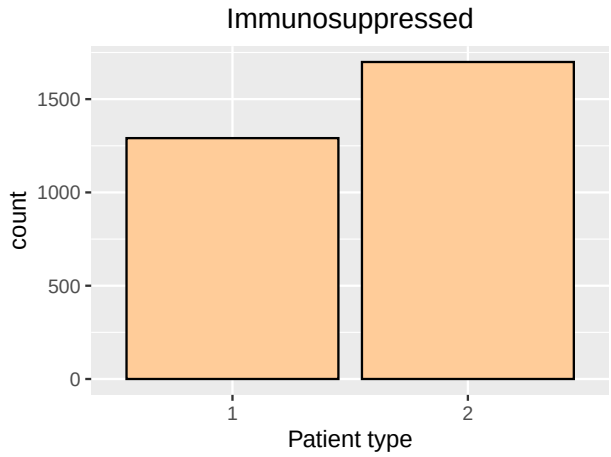
Parameter: Asthma

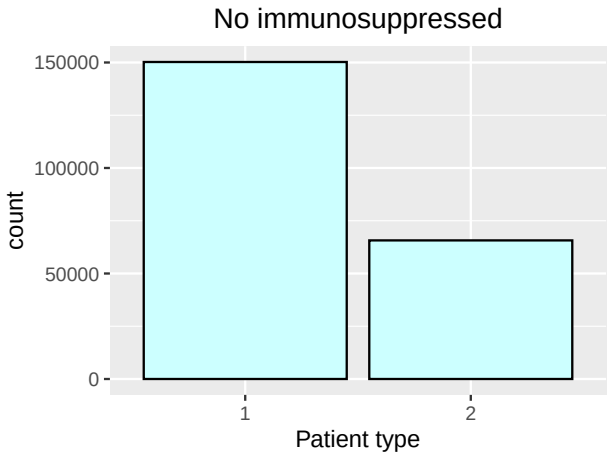




Both patients with and without asthma have more outpatients than inpatients.

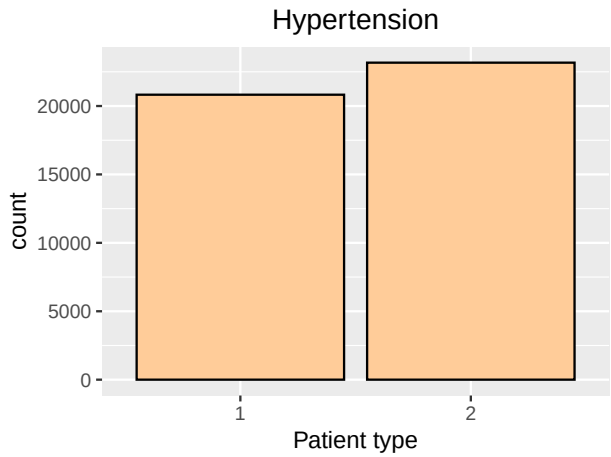
Parameter: Immunosuppression

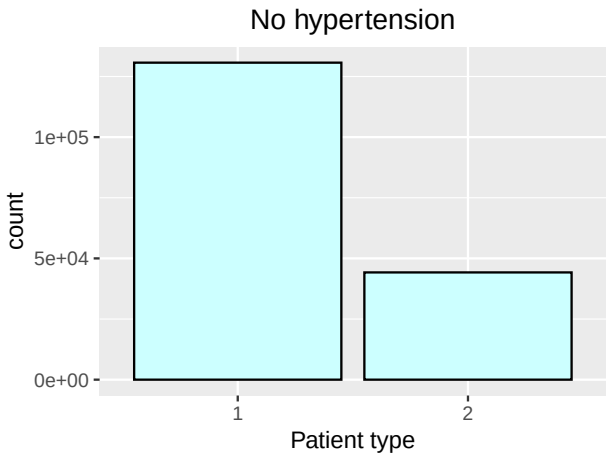




There are slightly more patients with immunosuppression that are inpatient than outpatient and there are more outpatients without immunosuppression than inpatients.

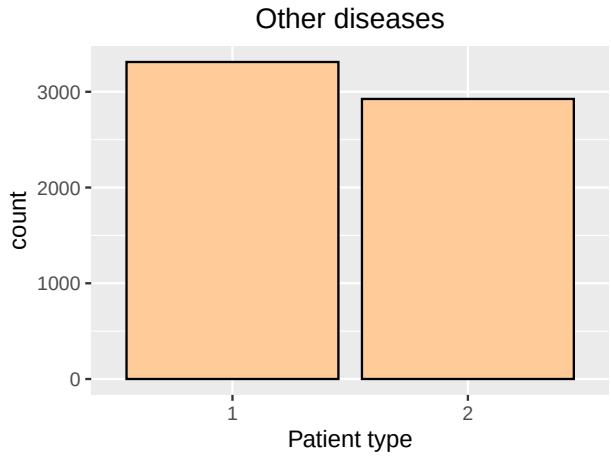
Parameter: Hypertension

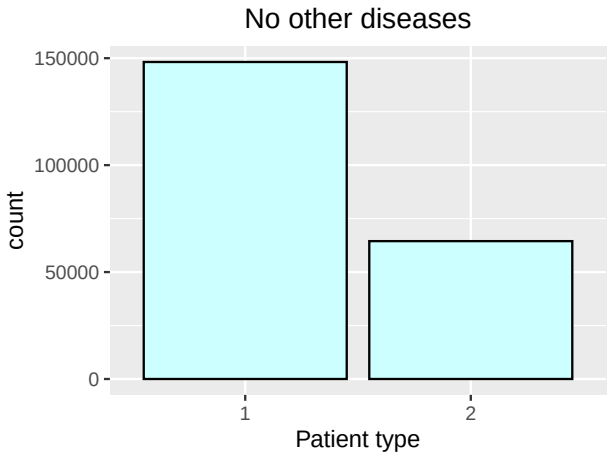




There are slightly more patients with hypertension that are inpatients than outpatients and more patients without hypertension that are outpatients than inpatients.

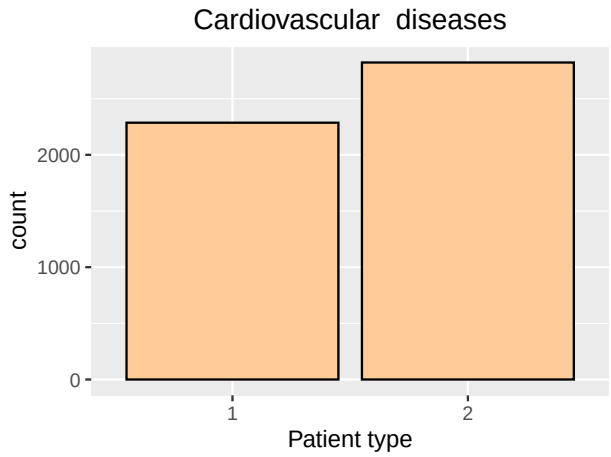
Parameter: Other diseases

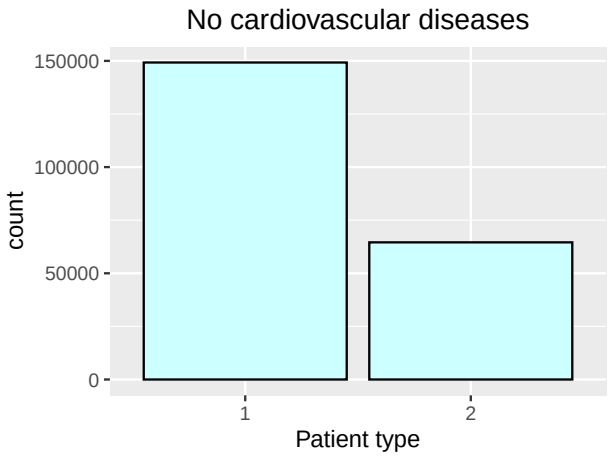




There are slightly more outpatients with other diseases than inpatients and there are more outpatients without other diseases than inpatients.

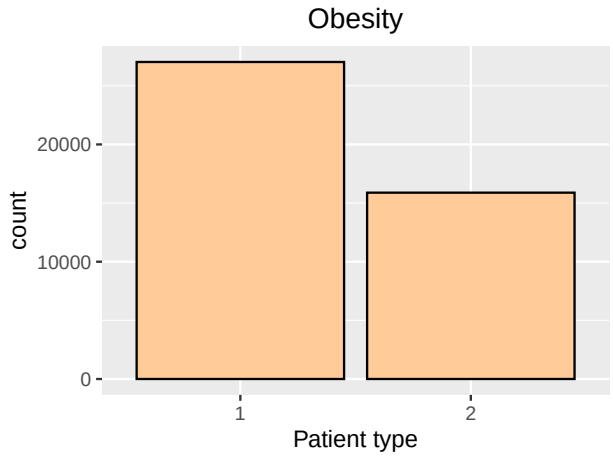
Parameter: Cardiovascular diseases

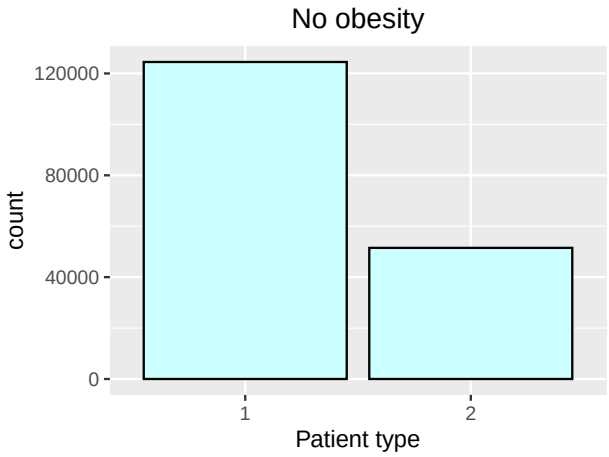




There are slightly more inpatients with other cardiovascular diseases than outpatients and there are more outpatients without cardiovascular diseases than inpatients.

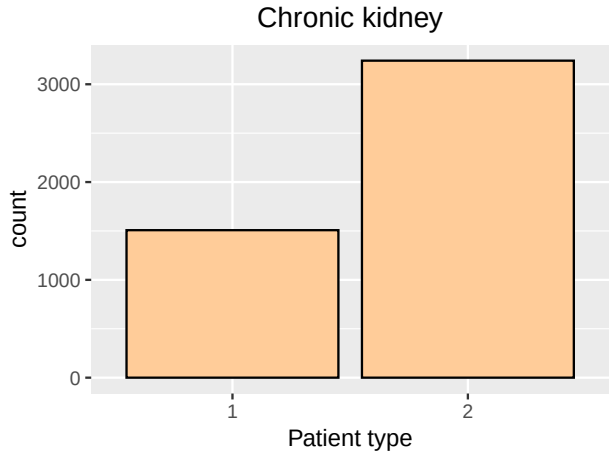
Parameter: Obesity

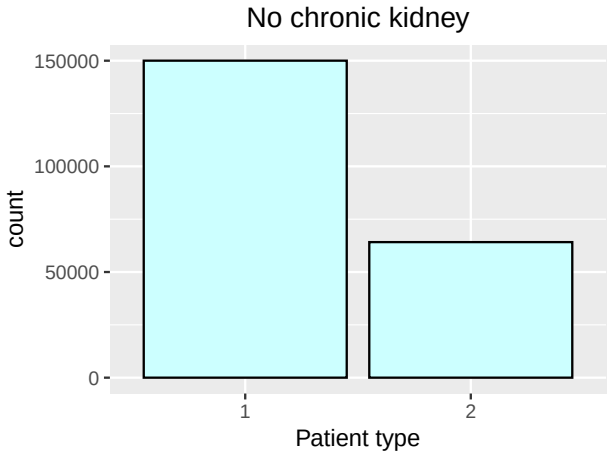




There are slightly more outpatients with obesity than inpatients and there are more outpatients without obesity than inpatients.

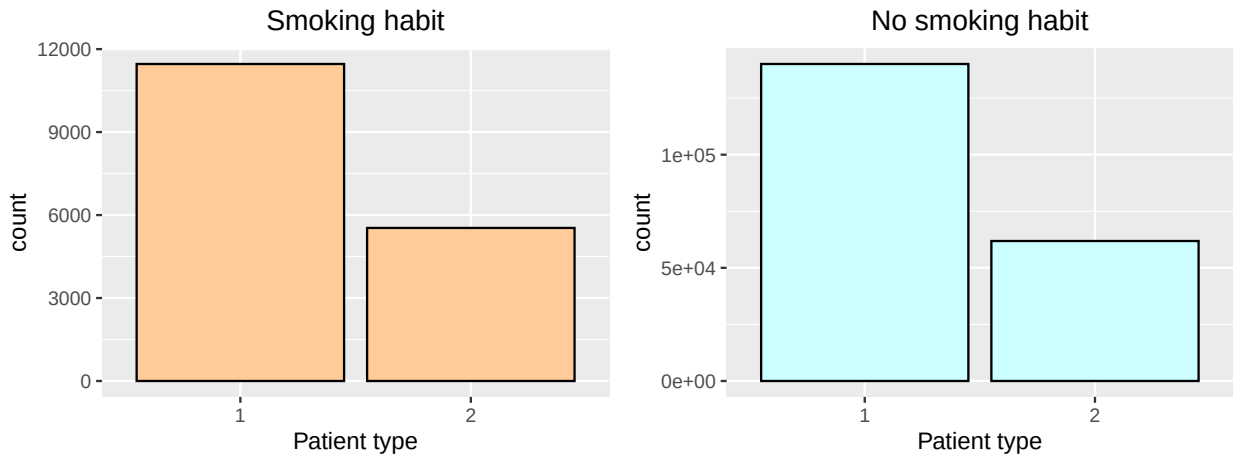
Parameter:Chronic kidney disease





There are more inpatients with chronic kidney disease than outpatients and more outpatients without chronic kidney disease than inpatients.

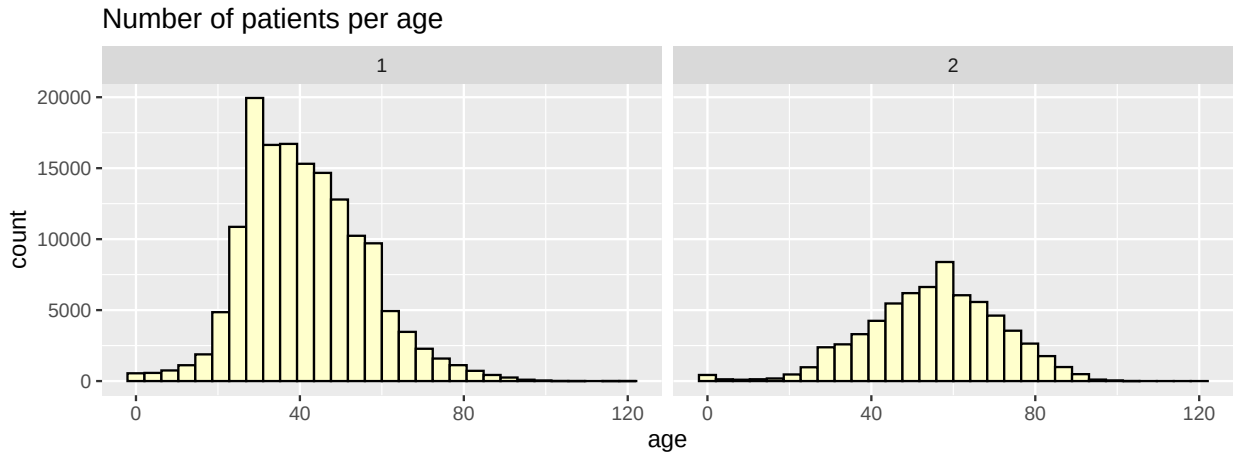
Parameter: Smoking habit



\end{figure}

With or without smoking habits the number of outpatients is bigger than the number of inpatients.

Parameter: Age



The proportion of inpatients change with age, it grows with the age.

Splitting the data in a training and test sets. The test set will be 20% of the original data set.

```
set.seed(1, sample.kind="Rounding")
test_index <- createDataPartition(y = covid_dat$patient_type, times = 1, p = 0.2, list = FALSE)
train_set<-covid_dat[-test_index,]
test_set<-covid_dat[test_index,]
ind<-which(test_set$patient_type==1)#patients of the test set that are outpatients
```

The first model will be linear discriminatory analysis.

```
train_lda<-train(as.factor(patient_type)~as.factor(sex)+as.factor(pneumonia)+age+
as.factor(diabetes)+as.factor(copd)+as.factor(asthma)+as.factor(inmsupr)+
as.factor(hypertension)+as.factor(other_disease)+as.factor(cardiovascular)+
as.factor(obesity)+as.factor(renal_chronic)+as.factor(tobacco),
```

```
data =train_set,method="lda")
lda_pred<-predict(train_lda,test_set)
cm<-confusionMatrix(table(as.numeric(lda_pred), test_set$patient_type))
lda_ac<-cm$overall[["Accuracy"]]
lda_out<-mean(lda_pred[ind]==1)#Accuracy on outpatients
lda_in<-mean(lda_pred[-ind]==2)#Accuracy on inpatients
```

The second model will be a generalized linear model.

```
train_glm<-train(as.factor(patient_type)~as.factor(sex)+as.factor(pneumonia)+age+
as.factor(diabetes)+as.factor(copd)+as.factor(asthma)+as.factor(inmsupr)+
as.factor(hypertension)+as.factor(other_disease)+as.factor(cardiovascular)+
as.factor(obesity)+as.factor(renal_chronic)+as.factor(tobacco),
method="glm",data =train_set)
glm_pred<-predict(train_glm,test_set)
cm<-confusionMatrix(table(as.numeric(glm_pred), test_set$patient_type))
glm_ac<-cm$overall[["Accuracy"]]
glm_out<-mean(glm_pred[ind]==1)#Accuracy on outpatients
glm_in<-mean(glm_pred[-ind]==2)#Accuracy on inpatients
```

The third model will be a quadratic discriminatory analysis.

```
train_qda<-train(as.factor(patient_type)~as.factor(sex)+as.factor(pneumonia)+age+
as.factor(diabetes)+as.factor(copd)+as.factor(asthma)+as.factor(inmsupr)+
as.factor(hypertension)+as.factor(other_disease)+as.factor(cardiovascular)+
as.factor(obesity)+as.factor(renal_chronic)+as.factor(tobacco),
data =train_set,method="qda")
qda_pred<-predict(train_qda,test_set)
cm<-confusionMatrix(table(as.numeric(qda_pred), test_set$patient_type))
qda_ac<-cm$overall[["Accuracy"]]
qda_out<-mean(qda_pred[ind]==1)#Accuracy on outpatients
qda_in<-mean(qda_pred[-ind]==2)#Accuracy on inpatients
```

The fourth model will be a classification tree.

```
train_rpart<-train(as.factor(patient_type)~as.factor(sex)+as.factor(pneumonia)+age+
as.factor(diabetes)+as.factor(copd)+as.factor(asthma)+as.factor(inmsupr)+
as.factor(hypertension)+as.factor(other_disease)+as.factor(cardiovascular)+
as.factor(obesity)+as.factor(renal_chronic)+as.factor(tobacco),
data = train_set,method="rpart")
rpart_pred<-predict(train_rpart,test_set)
cm<-confusionMatrix(table(as.numeric(rpart_pred), test_set$patient_type))
rpart_ac<-cm$overall[["Accuracy"]]
rpart_out<-mean(rpart_pred[ind]==1)#Accuracy on outpatients
rpart_in<-mean(rpart_pred[-ind]==2)#Accuracy on inpatients
```

The fifth model will be an ensemble of the other four models.If the majority of the models predict inpatient, the ensemble will predict inpatient. If the majority of models predict outpatient it will predict outpatient. If there is a tie, the ensemble will predict an outpatient, because there are more outpatients.

```
ensemble<-data.frame(LDA=as.numeric(lda_pred),
                     QDA=as.numeric(qda_pred),
                     GLM=as.numeric(glm_pred),
                     RPART=as.numeric(rpart_pred))
ensemble_pred<-ifelse(rowMeans(ensemble)<=6/4,1,2)
cm<-confusionMatrix(table(ensemble_pred, test_set$patient_type))
ensemble_ac<-cm$overall[["Accuracy"]]
ensemble_out<-mean(ensemble_pred[ind]==1)#Accuracy on outpatients
ensemble_in<-mean(ensemble_pred[-ind]==2)#Accuracy on inpatients
```

The sixth model will be an ensemble as the fifth model, but using the information acquired by analyzing the plots,From the plots could be seen that there are considerable more inpatients with pneumonia,chronic kidney disease, COPD and diabetes than outpatients with the same preconditions. In case of tie, the ensemble will consider these preconditions to predict, in case the patient has it, it will predict inpatient:

```
#same preconditions
ensemble_pred2<-ifelse(rowMeans(ensemble)<=(1*2+2*2)/4,1,2)
ind2<-which(rowMeans(ensemble)==6/4)#for ties in the ensemble
ensemble_pred2[ind2]<-ifelse(test_set$pneumonia[ind2]==1,2,
                             ifelse(test_set$renal_chronic[ind2]==1,2,
                                    ifelse(test_set$copd[ind2]==1,2,
                                           ifelse(test_set$diabetes[ind2]==1,2,1))))
cm<-confusionMatrix(table(ensemble_pred2, test_set$patient_type))
ensemble2_ac<-cm$overall[["Accuracy"]]#Accuracy 0.8631
ensemble2_out<-mean(ensemble_pred2[ind]==1)#Accuracy on outpatients
ensemble2_in<-mean(ensemble_pred2[-ind]==2)#Accuracy on inpatients
```

The seventh model is similar to Ensemble 2, but using all the pre-conditions that have more inpatients than outpatients (pneumonia,chronic kidney disease, COPD, diabetes, immunosuppression, hypertension and cardiovascular disease), it will predict inpatient in case that is a tie in the ensemble and the patient has at least one of the pre-conditions:

```
ensemble_pred3<-ifelse(rowMeans(ensemble)<=6/4,1,2)
ind2<-which(rowMeans(ensemble)==6/4)#for ties in the ensemble
ensemble_pred3[ind2]<-ifelse(test_set$pneumonia[ind2]==1,2,
                      ifelse(test_set$renal_chronic[ind2]==1,2,
                      ifelse(test_set$copd[ind2]==1,2,
                      ifelse(test_set$diabetes[ind2]==1,2,
                      ifelse(test_set$inmsupr[ind2]==1,2,
                      ifelse(test_set$hypertension[ind2]==1,2,
                      ifelse(test_set$cardiovascular[ind2]==1,2,1)))))))
cm<-confusionMatrix(table(ensemble_pred3, test_set$patient_type))
ensemble3_ac<-cm$overall[["Accuracy"]]#Accuracy 0.8
ensemble3_out<-mean(ensemble_pred3[ind]==1)#Accuracy on outpatients
ensemble3_in<-mean(ensemble_pred3[-ind]==2)#Accuracy on inpatients
```

Finally a accuracy (one for general, other for outpatient and other for inpatient) data.frame is created:

```
accuracy<-data.frame(row.names = c("LDA","GLM","QDA","RPART","ENSEMBLE","ENSEMBLE2","ENSEMBLE3"),
Accuracy=c(lda_ac,glm_ac,qda_ac,rpart_ac,ensemble_ac,ensemble2_ac,ensemble3_ac))

acc_out<-data.frame(row.names = c("LDA","GLM","QDA","RPART","ENSEMBLE","ENSEMBLE2","ENSEMBLE3"),
```

```
Accuracy=c(lda_out,glm_out,qda_out,rpart_out,ensemble_out,ensemble2_out,ensemble3_out))

acc_in<-data.frame(row.names = c("LDA","GLM","QDA","RPART","ENSEMBLE","ENSEMBLE2","ENSEMBLE3"),
Accuracy=c(lda_in,glm_in,qda_in,rpart_in,ensemble_in,ensemble2_in,ensemble3_in))
```

Results

The general accuracy of the models are:

```
##            Accuracy
## LDA       0.8618579
## GLM       0.8621777
## QDA       0.8423289
## RPART     0.8637308
## ENSEMBLE  0.8626345
## ENSEMBLE2 0.8631141
## ENSEMBLE3 0.8632283
```

The three ensembles have better performances than LDA,GLM and QDA models but less than the RPART model's performance. All the models have an accuracy around of 86%, only the QDA method is inferior to that with an accuracy of 0.842. One of the motives that the QDA model has a considerably low accuracy in comparison to the other models is that there are 13 predictors and QDA doesn't perform well with many predictors.

The accuracy of the models considering only the outpatients are:

```
##            Accuracy
## LDA       0.9493758
## GLM       0.9441853
## QDA       0.8981275
## RPART     0.9445795
## ENSEMBLE  0.9476018
## ENSEMBLE2 0.9442510
## ENSEMBLE3 0.9439225
```

The accuracy of the models of only the outpatients of all the models are very good, all except QDA are greater than 94%. The LDA model has the best accuracy of the outpatients.

The accuracy of the models considering only the inpatients are:

```
##            Accuracy
## LDA       0.6621693
## GLM       0.6750618
## QDA       0.7150139
## RPART     0.6792594
## ENSEMBLE  0.6687655
## ENSEMBLE2 0.6779852
## ENSEMBLE3 0.6791095
```

Those are all around 67% of accuracy, except for QDA model that outperforms the others models with an accuracy of 71.5%

The model with the lowest general accuracy and lowest accuracy for outpatients is the best to predict inpatients (QDA).

The performance of the third ensemble is better than the other two in general accuracy and accuracy of inpatients, but has the worst accuracy of outpatients of the ensembles.The performance of the second ensemble is better than the first one in all the cases (general accuracy, accuracy of outpatients and accuracy of inpatients).

Conclusion

The analysis of the plots are based on the sample, for confirmation of those conclusions more studies are needed.

The models could help to prioritize the medical attention of patients that are more susceptible to need a hospitalization based on the pre-conditions.

None of the models uses pregnancy, dates, intubation or UCI to make predictions, using these or more parameters (others pre-conditions as diet,physical activity, consumption of alcohol,etc.) could help to improve the performance of the models.

More models (knn, random forest, etc.) could be used to try to improve the accuracy of the algorithm.