Covid Pre-Condition

Rafael Pereira

5/1/2021

Introduction

This algorithm uses 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 the 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 nine models are used: linear discriminant analysis (LDA), generalized linear model (LGM), quadratic discriminant analysis (QDA), classification and regression tree (RPART),Boosted classification tree (BTREE), Conditional interference tree (CTREE) and three ensembles of the other six 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")
if(!require(ada)) install.packages("ada", repos = "http://cran.us.r-project.org")
if(!require(plyr)) install.packages("plyr", repos = "http://cran.us.r-project.org")
if(!require(xgboost)) install.packages("xgboost", repos = "http://cran.us.r-project.org")
if(!require(party)) install.packages("party", repos = "http://cran.us.r-project.org")

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

```
library(plyr)
library(xgboost)
library(party)
```

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$copd%in%c(97,98,99)|covid_dat$asthma%in%c(97,98,99)|
covid_dat$inmsupr%in%c(97,98,99)|covid_dat$hypertension%in%c(97,98,99)|
covid_dat$other_disease%in%c(97,98,99)|covid_dat$cardiovascular%in%c(97,98,99)|
covid_dat$obesity%in%c(97,98,99)|covid_dat$renal_chronic%in%c(97,98,99)|
covid_dat$tobacco%in%c(97,98,99))
covid_dat<-covid_dat[-ind,]
```

The number of patients in the data set is:

```
## [1] 218902
```

The number of outpatients in the data set is:

```
## [1] 151518
```

The number of inpatients in the data set is:

```
## [1] 67384
```

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

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

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

```r
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 number of patients in the training set is:

```
## [1] 175121
```

The number of outpatients in the training set is:

```
## [1] 121078
```

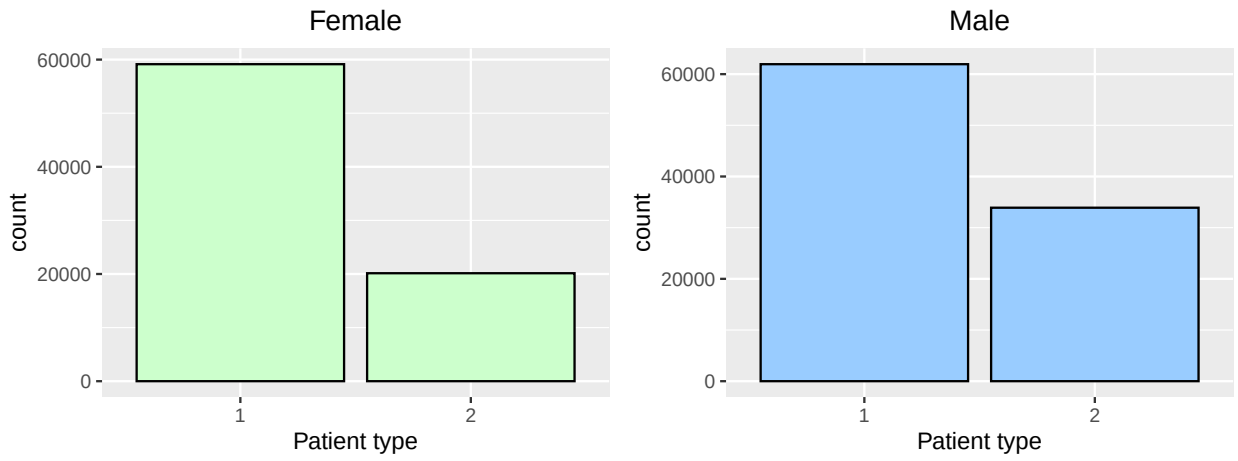
The number of inpatients in the training set is:

```
## [1] 54043
```

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

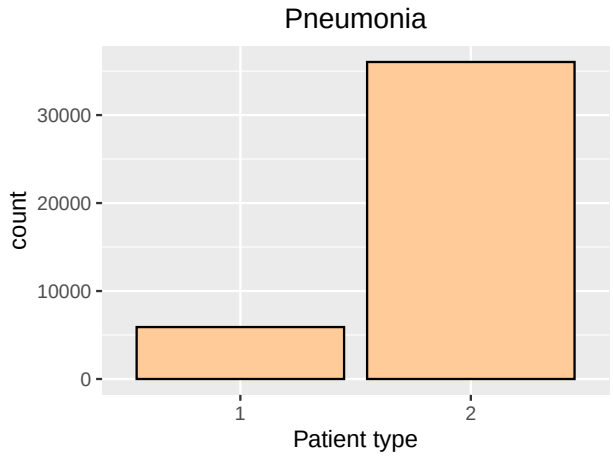
```
## [1] 0.6913962
```

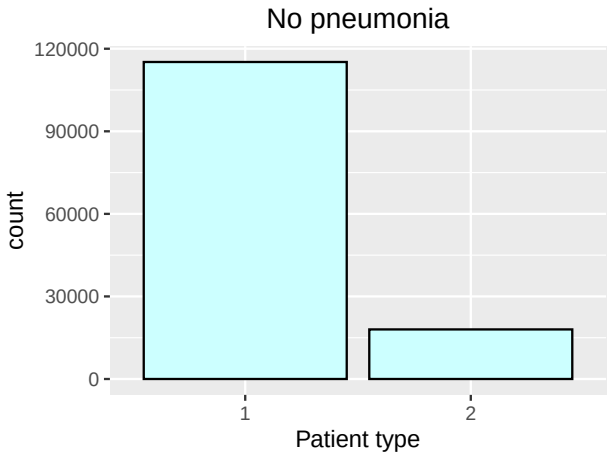
Doing a visual inspection of the training set: First parameter to analyze: Sex



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

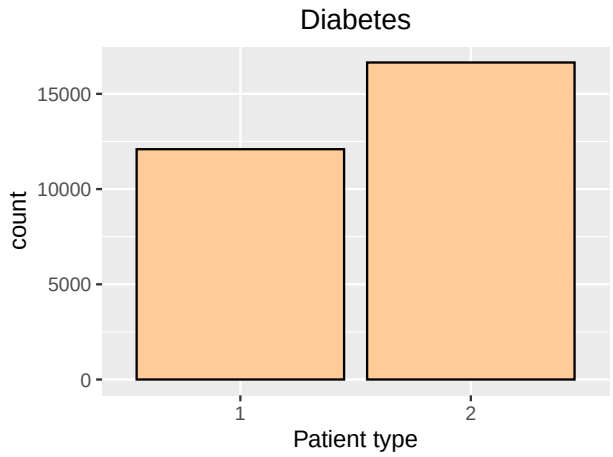
Parameter: Pneumonia

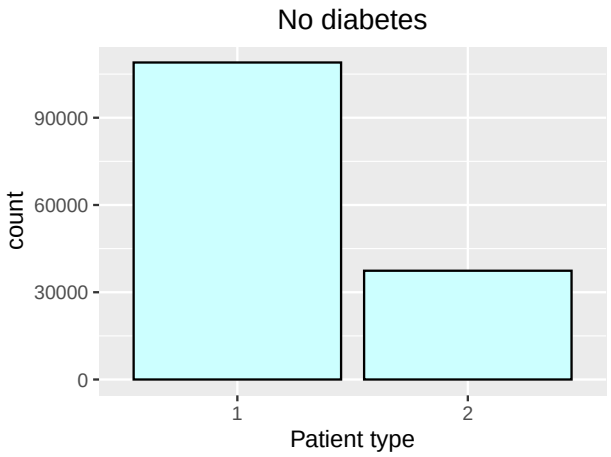




There are more inpatients with pneumonia than outpatients with pneumonia and there are more outpatients without pneumonia than inpatients with pneumonia.

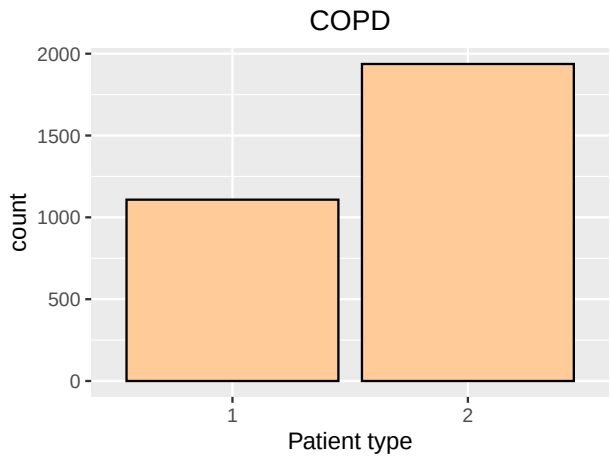
Parameter: Diabetes

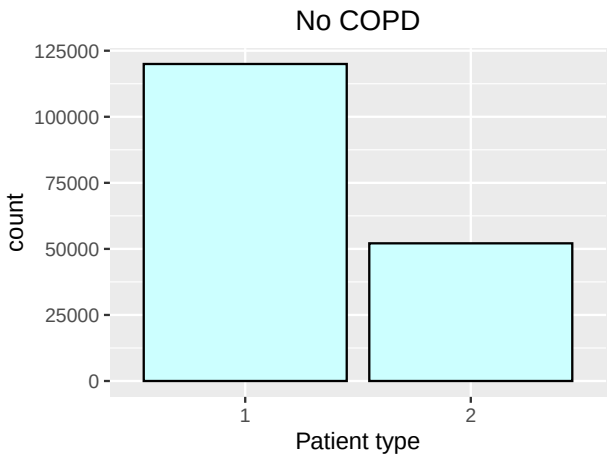




There are slightly more inpatients with diabetes than outpatients with diabetes and there are more outpatients without diabetes than inpatients without diabetes.

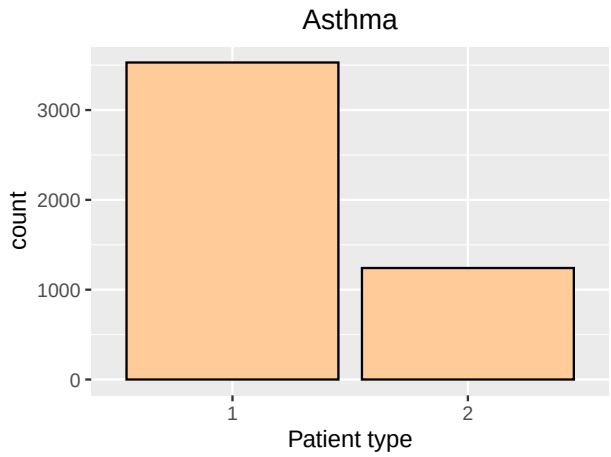
Parameter: Chronic obstructive pulmonary disease

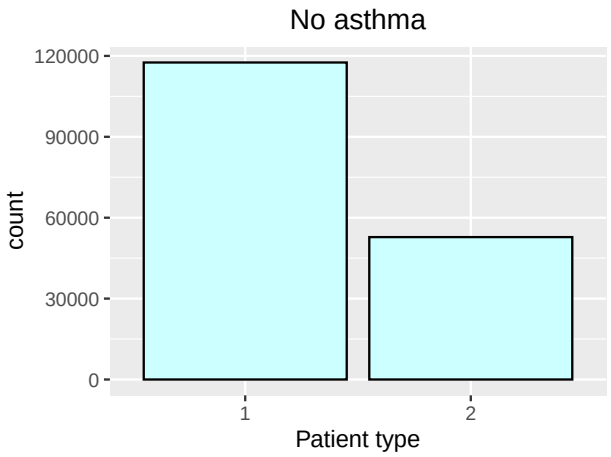




There are more inpatients with COPD than outpatients with COPD and there are more outpatients without COPD than inpatients without COPD.

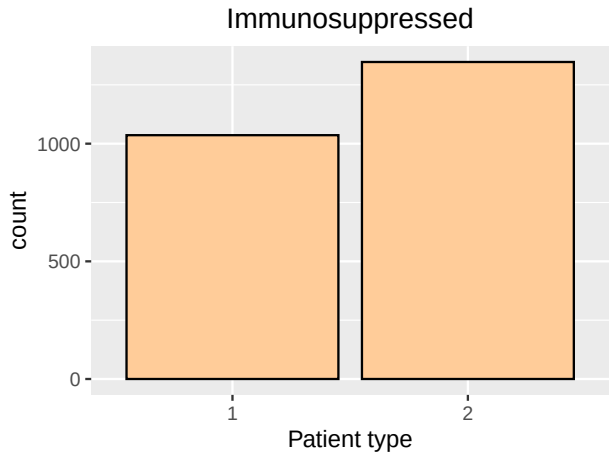
Parameter: Asthma

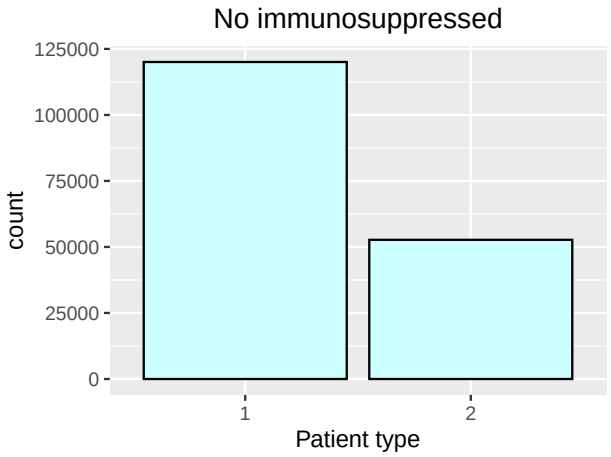




Both patients with and without asthma have more outpatients than inpatients.

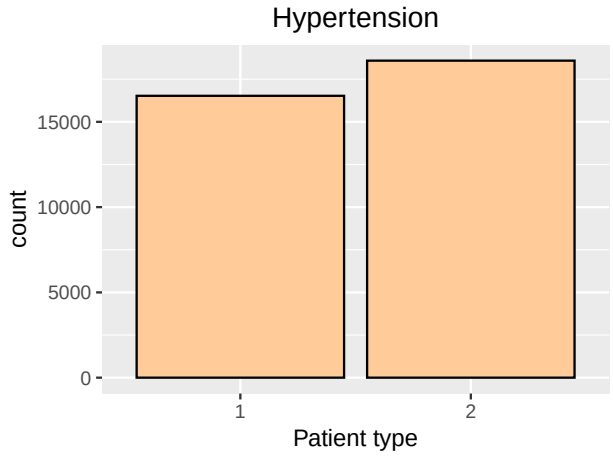
Parameter: Immunosuppression





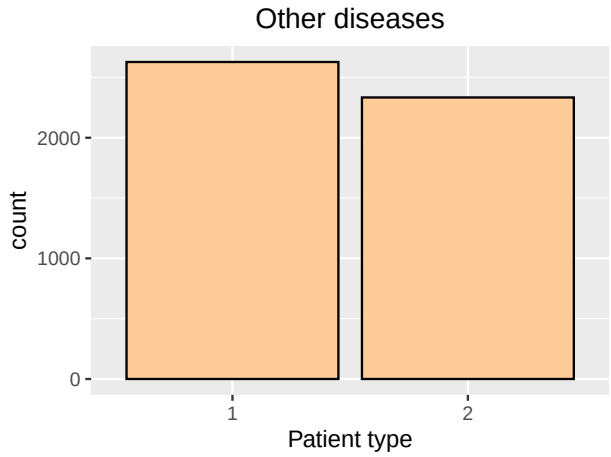
There are slightly more inpatients with immunosuppression than outpatient with immunosuppression and there are more outpatients without immunosuppression than inpatients without immunosuppression.

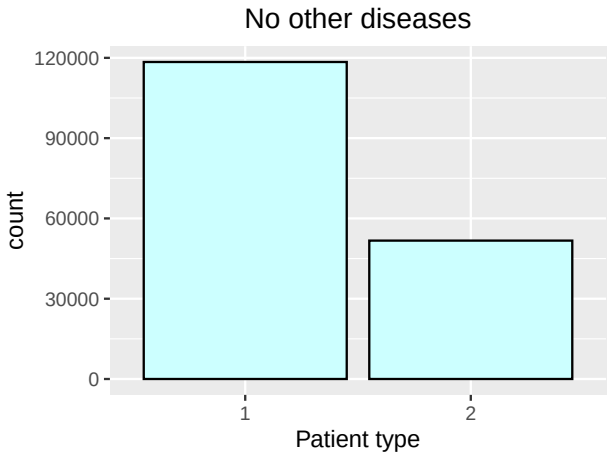
Parameter: Hypertension



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

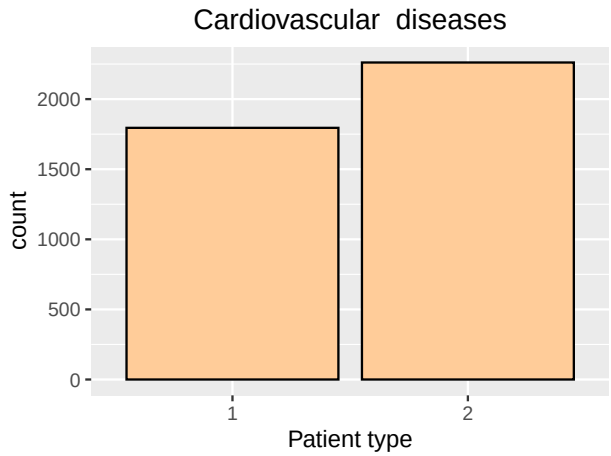
Parameter: Other diseases

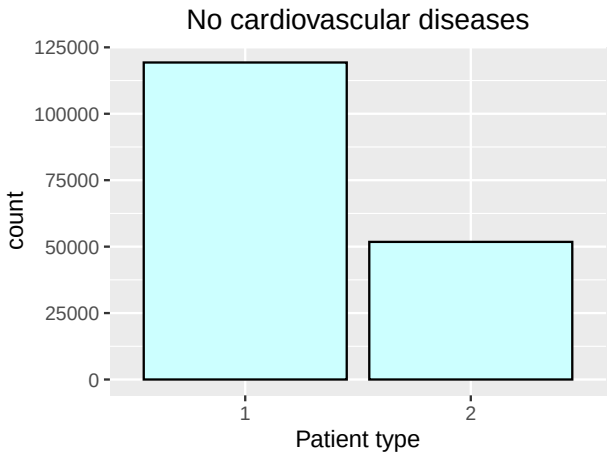




There are slightly more outpatients with other diseases than inpatients and there are more outpatients without other diseases than inpatients without other diseases.

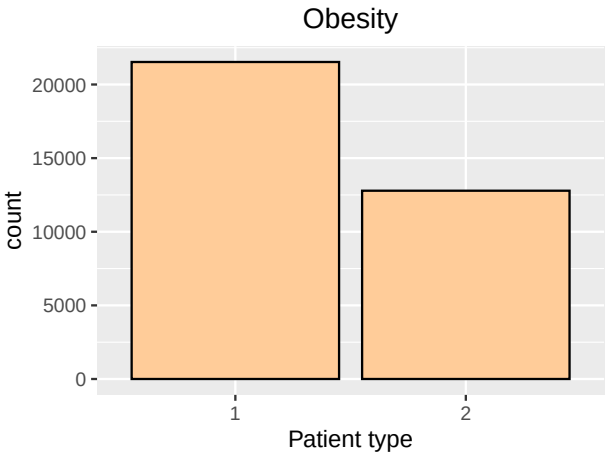
Parameter: Cardiovascular diseases

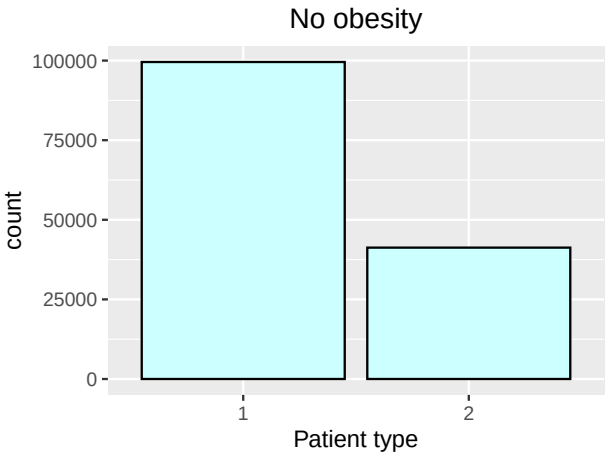




There are slightly more inpatients with cardiovascular diseases than outpatients and there are more outpatients without cardiovascular diseases than inpatients without cardiovascular diseases.

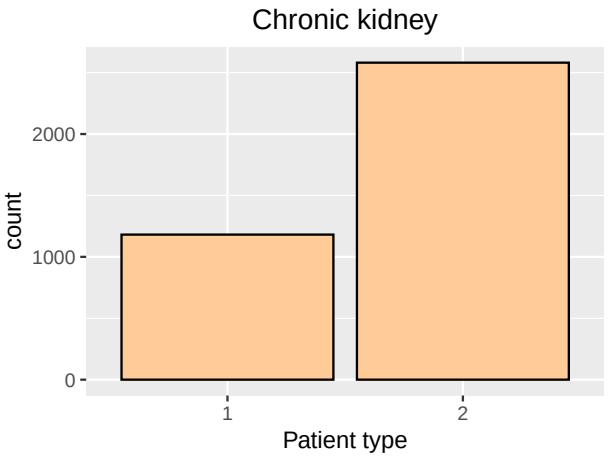
Parameter: Obesity

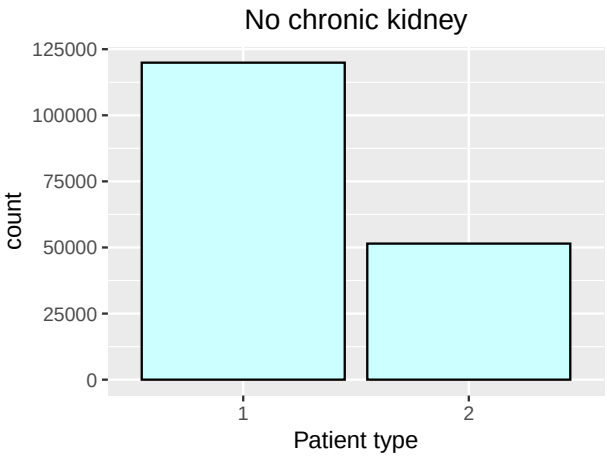




There are slightly more outpatients with obesity than inpatients and there are more outpatients without obesity than inpatients without obesity.

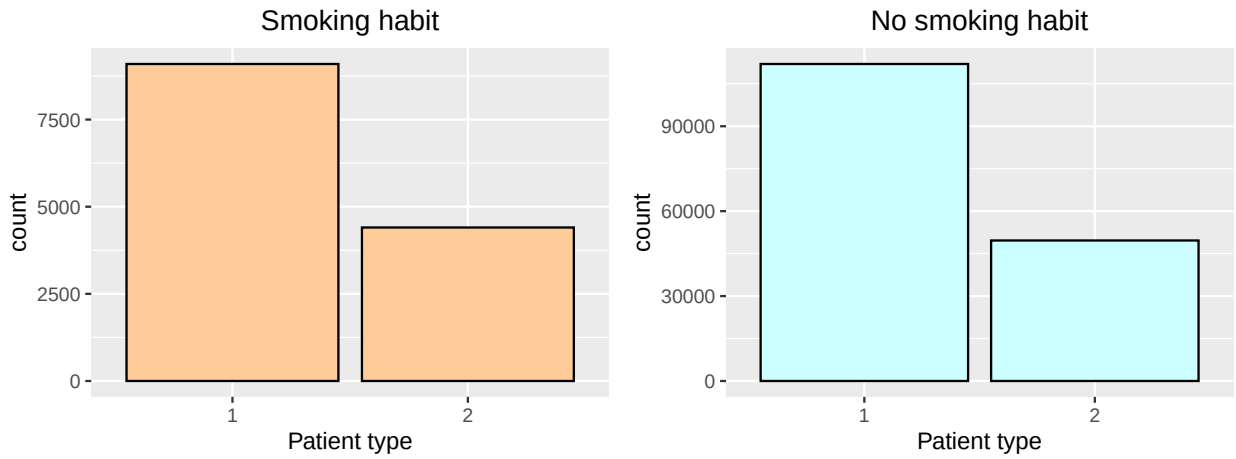
Parameter:Chronic kidney disease





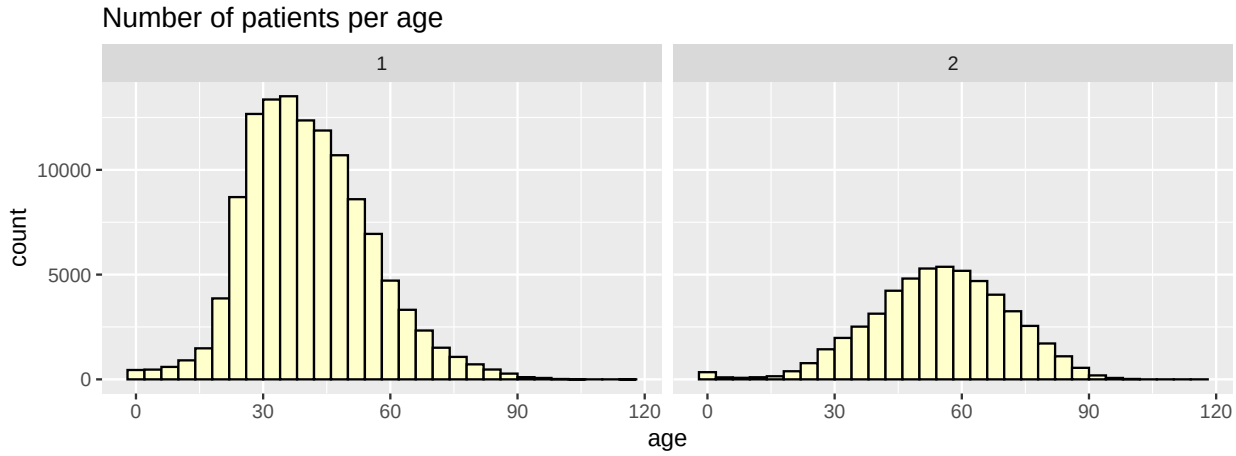
There are more inpatients with chronic kidney disease than outpatients and more outpatients without chronic kidney disease than inpatients without chronic kidney disease.

Parameter: Smoking habit



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.

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,family = "binomial")
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 a Boosted Classification Tree. It uses a gradient descent algorithm which can optimize any differentiable loss function. An ensemble of trees are built one by one and individual trees are summed sequentially. The next tree tries to recover the loss (difference between actual and predicted values).

```
train_ada<-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="xgbTree", trControl = trainControl("cv", number = 5), data =train_set)
train_ada$bestTune
```

```
##    nrounds max_depth eta gamma colsample_bytree min_child_weight subsample
## 50     100         3 0.3     0              0.8                1      0.75
```

```
ada_pred<-predict(train_ada,test_set)
cm<-confusionMatrix(table(as.numeric(ada_pred), test_set$patient_type))
ada_ac<-cm$overall[["Accuracy"]] #Accuracy 0.8621
ada_out<-mean(ada_pred[ind]==1)#Accuracy on outpatients
ada_in<-mean(ada_pred[-ind]==2)#Accuracy on inpatients
```

The sixth model will be a Conditional Inference Tree. This model (ctree), according to its authors avoids the following variable selection bias of rpart: they tend to select variables that have many possible splits or many missing values. Unlike the rpart model, ctree uses a significance test procedure in order to select variables instead of selecting the variable that maximizes an information measure.

```
train_cit<-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="ctree", data =train_set)
cit_pred<-predict(train_cit,test_set)
cm<-confusionMatrix(table(as.numeric(cit_pred), test_set$patient_type))
cit_ac<-cm$overall[["Accuracy"]] #Accuracy 0.8627
cit_out<-mean(cit_pred[ind]==1)#Accuracy on outpatients
cit_in<-mean(cit_pred[-ind]==2)#Accuracy on inpatients
```

The seventh model will be an ensemble of the other six models.If the majority of the models predict an inpatient, the ensemble will predict an inpatient. If the majority of models predict an outpatient it will predict an outpatient. If there is a tie, the ensemble will predict an outpatient, because there are more outpatients in the data set.

```
ensemble<-data.frame(LDA=as.numeric(lda_pred),
                     QDA=as.numeric(qda_pred),
                     GLM=as.numeric(glm_pred),
                     RPART=as.numeric(rpart_pred),
                     ADA=as.numeric(ada_pred),
                     CIT=as.numeric(cit_pred))
ensemble_pred<-ifelse(rowMeans(ensemble)<=9/6,1,2)#There are more outpatients
#than inpatients, so in case of tie, predict outpatient
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 eighth model will be an ensemble similar to the seventh model, but using the information acquired by analyzing the plots. From the plots, it's visible that there is a higher amount of inpatients with pneumonia,chronic kidney disease, COPD and diabetes than outpatients with the same pre-conditions. In case of a tie, the ensemble will consider these pre-conditions to predict the outcome, in case the patient has one of them, it will predict an inpatient, otherwise it will predict an outpatient:

```
ensemble_pred2<-ifelse(rowMeans(ensemble)<=9/6,1,2)
ind2<-which(rowMeans(ensemble)==9/6)#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
ensemble2_out<-mean(ensemble_pred2[ind]==1)#Accuracy on outpatients
ensemble2_in<-mean(ensemble_pred2[-ind]==2)#Accuracy on inpatients
```

The ninth 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). In case of a tie, the ensemble will consider these pre-conditions to predict the outcome, in case the patient has one of them, it will predict an inpatient, otherwise it will predict an outpatient:

```
ensemble_pred3<-ifelse(rowMeans(ensemble)<=9/6,1,2)
ind2<-which(rowMeans(ensemble)==9/6)#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
ensemble3_out<-mean(ensemble_pred3[ind]==1)#Accuracy on outpatients
ensemble3_in<-mean(ensemble_pred3[-ind]==2)#Accuracy on inpatients
```

Finally three accuracy data frames are created(one for overall, one for outpatient and other for inpatient):

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

#Creating a data frame with the accuracy of the outpatients
acc_out<-data.frame(row.names =
c("LDA","GLM","QDA","RPART","BTREE","CTREE","ENSEMBLE","ENSEMBLE2","ENSEMBLE3"),
Accuracy=c(lda_out,glm_out,qda_out,rpart_out,ada_out,cit_out,ensemble_out,
ensemble2_out,ensemble3_out))

#Creating a data frame with the accuracy of  the inpatients
acc_in<-data.frame(row.names =
c("LDA","GLM","QDA","RPART","BTREE","CTREE","ENSEMBLE","ENSEMBLE2","ENSEMBLE3"),
Accuracy=c(lda_in,glm_in,qda_in,rpart_in,ada_in,cit_in,ensemble_in,ensemble2_in,
ensemble3_in))
```

Results

The overall accuracy of the models are:

```
##            Accuracy
## LDA       0.8618579
## GLM       0.8621777
## QDA       0.8423289
```

```
## RPART     0.8637308
## BTREE     0.8635710
## CTREE     0.8627487
## ENSEMBLE  0.8633425
## ENSEMBLE2 0.8633882
## ENSEMBLE3 0.8632969
```

Each of the three ensembles have a better overall accuracy than the LDA, GLM, QDA and CTREE models but less than the BTREE and RPART models. The best model for overall accuracy is the RPART. All the models have an overall accuracy around of 86%, except the QDA model, which has an inferior overall accuracy of 84.2%.

The accuracy of the models considering only the outpatients with covid are:

```
##            Accuracy
## LDA       0.9493758
## GLM       0.9441853
## QDA       0.8981275
## RPART     0.9445795
## BTREE     0.9424113
## CTREE     0.9455979
## ENSEMBLE  0.9468791
## ENSEMBLE2 0.9446124
## ENSEMBLE3 0.9443495
```

The accuracy for predicting outpatients of all the models are greater than 94%, except the QDA with 89.8%. The LDA model has the best accuracy for predicting outpatients.

The accuracy of the models considering only the inpatients are:

```
##            Accuracy
## LDA       0.6621693
## GLM       0.6750618
## QDA       0.7150139
## RPART     0.6792594
## BTREE     0.6836819
## CTREE     0.6737126
## ENSEMBLE  0.6727382
## ENSEMBLE2 0.6780601
## ENSEMBLE3 0.6783599
```

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

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

The accuracy of the third ensemble is better than the other two for predicting inpatients, but has the worst overall accuracy and accuracy of outpatients of the ensembles.The second ensemble is better than the first one in overall accuracy and accuracy of inpatients. The first ensemble is the best in accuracy for predicting outpatients.

Conclusion

The analysis of the plots that are used for the second and third ensembles are based on the sample, for confirmation of those conclusions more studies are needed.

The models could help prioritize the medical attention for patients that are more in need of hospitalization based on their pre-conditions.

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

More models (knn, random forest, etc.) could be used to improve the accuracy of the algorithm, but the results show that different models have similar inpatient, outpatient and overall accuracy. We can conclude that the algorithm's performance depends more on the pre-conditions (predictors) than the models used.