Movielens Report

Samuel Pereira

31/05/2021

Introduction

In this report I will display how a data-set containing medical data of heart parameters was used to create a prediction system based on results of exams made on the patients, they contain a wide number of variables to analyze.

In order to better predict the presence of a heart disease the prediction system will feature more than one method of prediction and it will be evaluated based on it's accuracy.

Method

About the data

The data used for the analysis was downloaded from: https://www.kaggle.com/ronitf/heart-disease-uci via the kaggle plataform, it's a modified data-frame from the machine learning repository from of the University of California Irvine made by the following creators:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Data description and exploration

The data set is already in the tidy format but many of the columns as described by the authors use integers in place of characters in test the result in categorical data.

A user of the plataform also pointed that some modifications were not documented in the post here is the link of the comment: https://www.kaggle.com/ronitf/heart-disease-uci/discussion/105877.

target the target column is the variable being predicted by the model, It's a registry that shows if a patient has a heart disease or not with 0 = yes and 1 = no.

```
heart_data <- heart_data %>% mutate(target = as.factor(target))
levels(heart_data$target) <- c("yes","no")
summary(heart_data$target)
```

Age The first column is the age column, this section of the data contains the ages of the patients gathered for the study as a numeric variable.

```
summary(heart_data$ï..age)
```

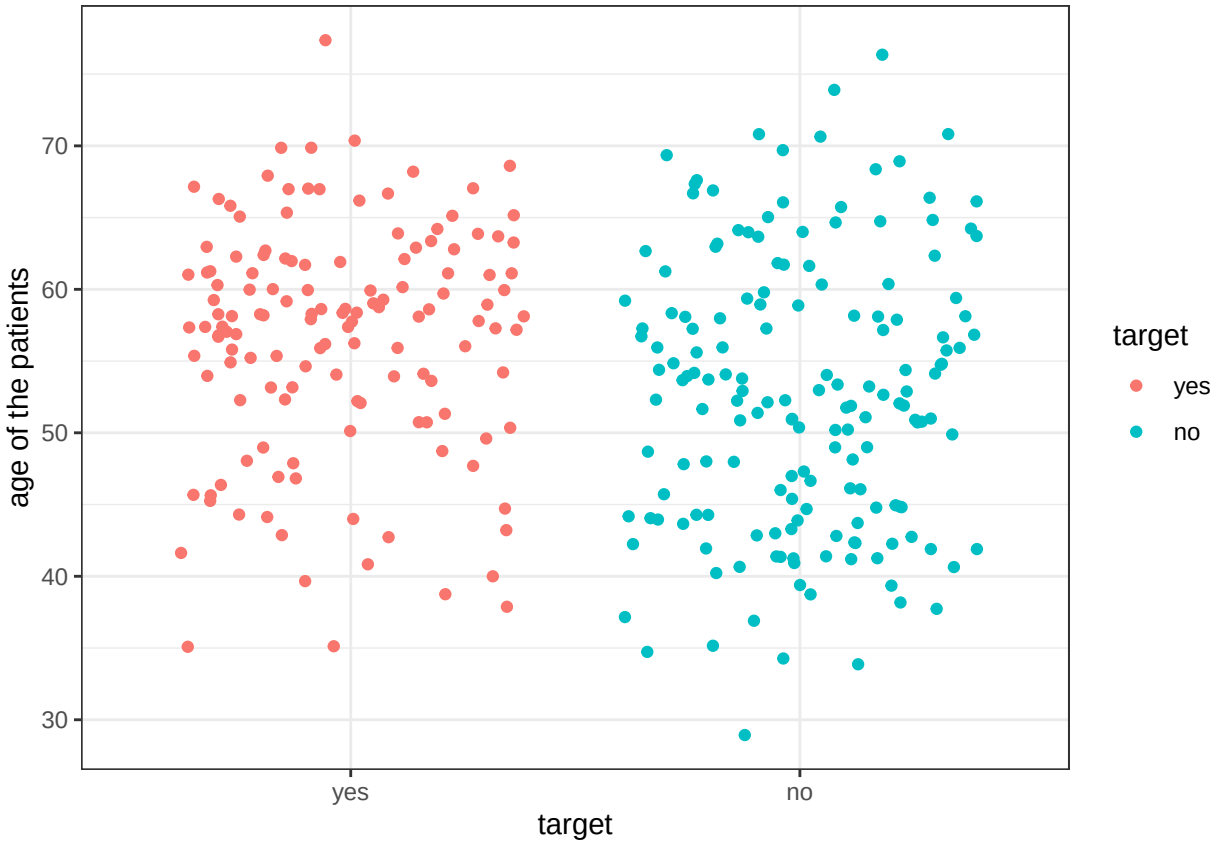
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   47.50   55.00   54.37   61.00   77.00
```

```
head(heart_data$ï..age)
```

```
## [1] 63 37 41 56 57 57
```

```
colnames(heart_data)[1] <- "age"

exp_01 <- heart_data %>% ggplot(aes(x = target, y = age,  col = target))+
  geom_jitter()+
  labs(x = "target", y = "age of the patients")+
  theme_bw()
exp_01
```

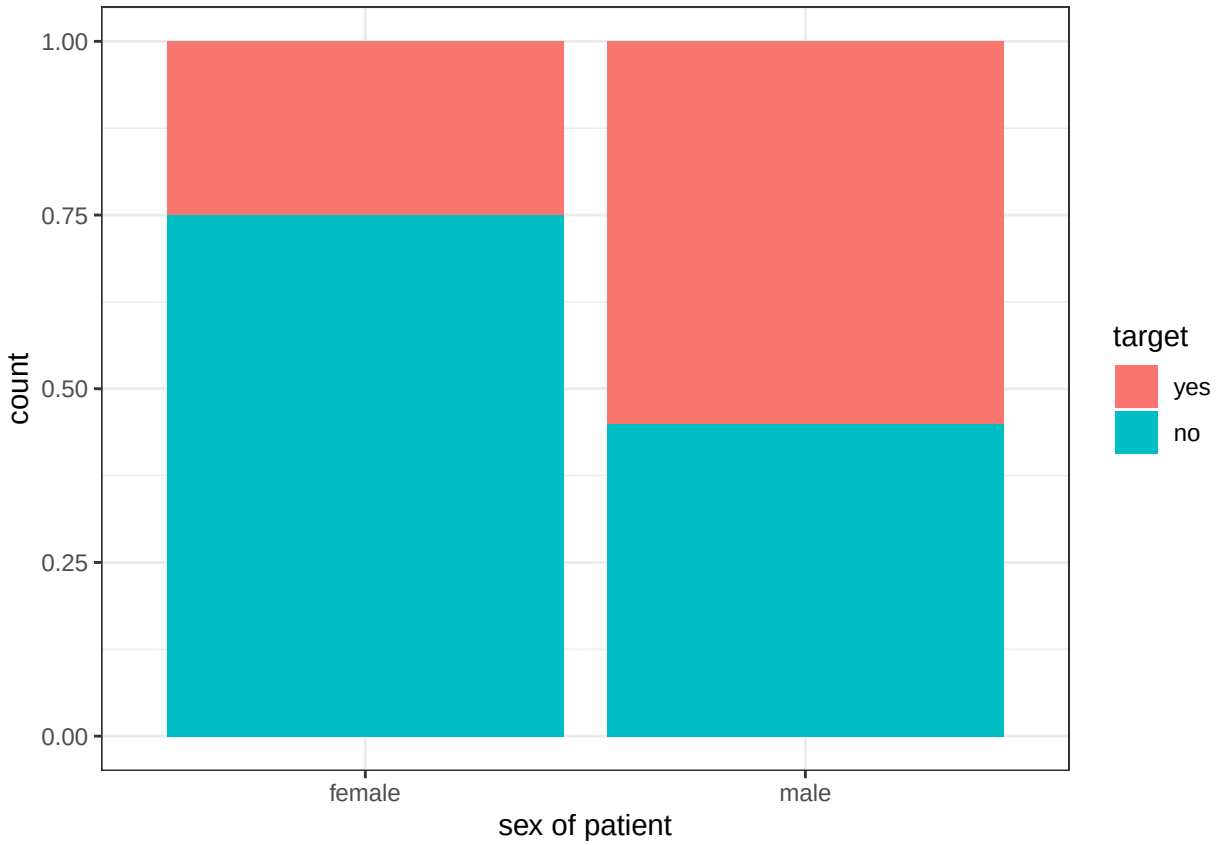


Sex The second column contains the sex of the patients divided in 0 = female and 1 = male, this column was converted from integer to a factor in order to facilitate analysis and modeling of the data in question.

```
heart_data <- heart_data %>% mutate(sex = as.factor(sex))
levels(heart_data$sex) <- c("female","male")
summary(heart_data$sex)
```

```
## female   male
##     96    207
```

```
exp_02 <- heart_data %>% ggplot(aes(x = sex, fill = target))+
  geom_bar(stat = "count",position = "fill")+
  labs(x = "sex of patient")+
  theme_bw()
exp_02
```



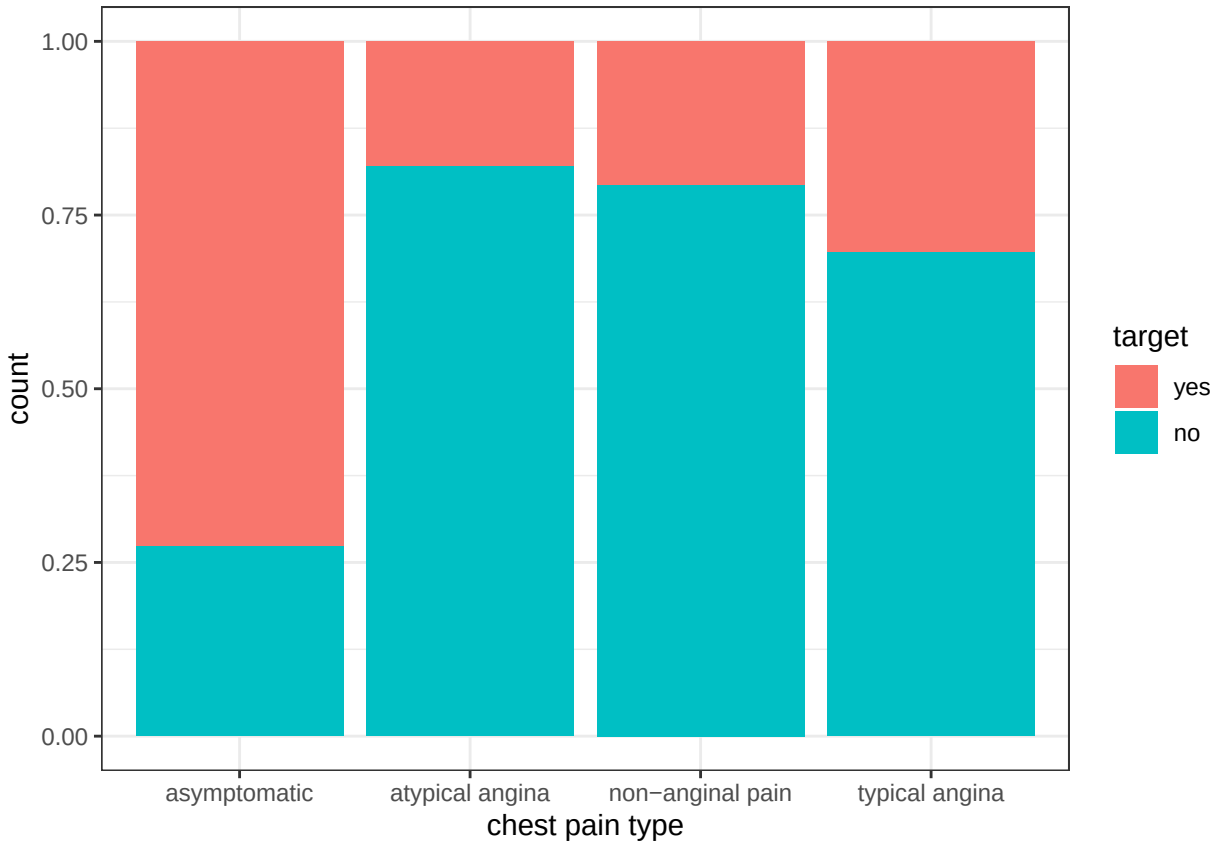
Chest pain (cp) The cp column contains the response of the patients when asked if they were felling chest pain and which kind of pain was perceived. This column was converted from an integer (0 = asymptomatic, 1 = atypical angina, 2 = non-anginal pain ,3 = typical angina) to a factor to facilitate it's study.

```
heart_data <- heart_data %>% mutate(cp = as.factor(cp))
levels(heart_data$cp) <- c("asymptomatic","atypical angina","non-anginal pain","typical angina")

summary(heart_data$cp)
```

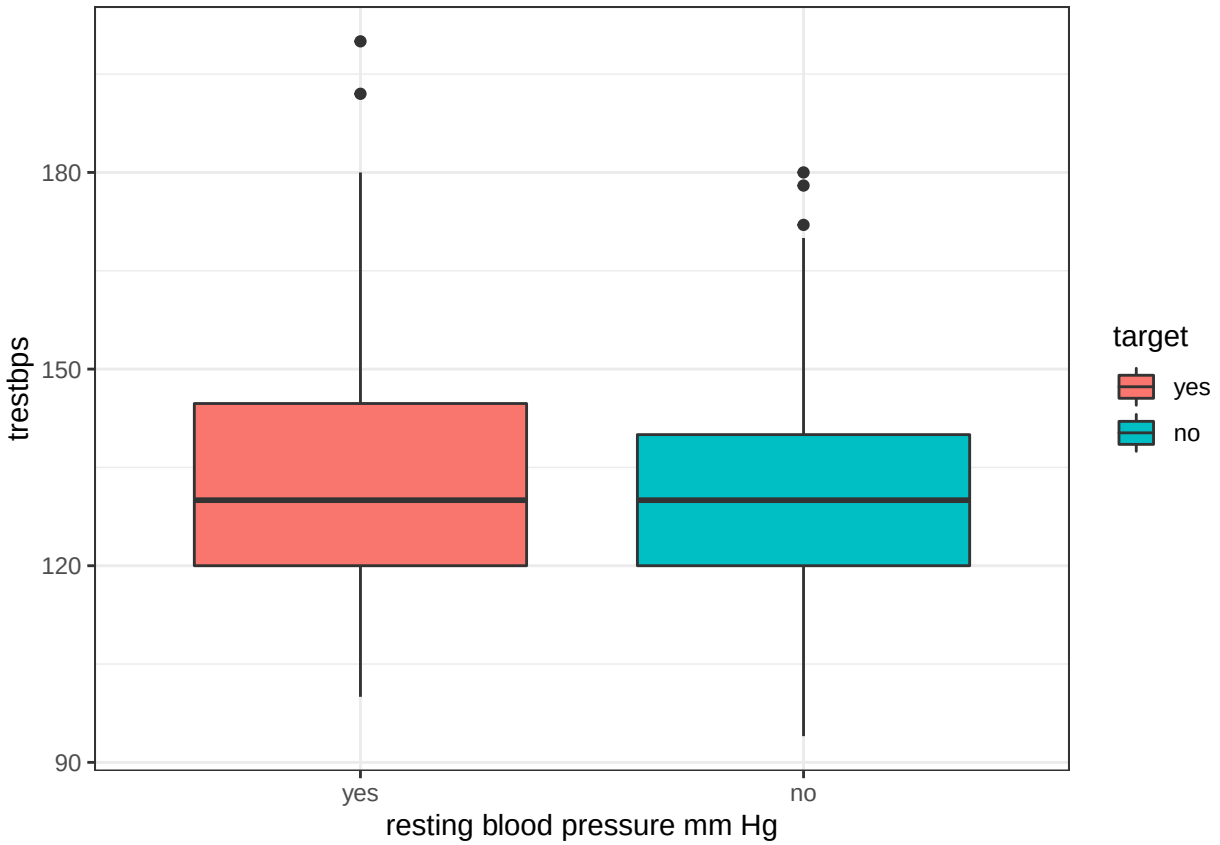
```
##     asymptomatic  atypical angina non-anginal pain   typical angina
##              143               50               87               23
```

```
exp_03 <- heart_data %>% ggplot(aes(x = cp, fill = target))+
  geom_bar(stat = "count",position = "fill")+
  labs(x = "chest pain type")+
  theme_bw()
exp_03
```



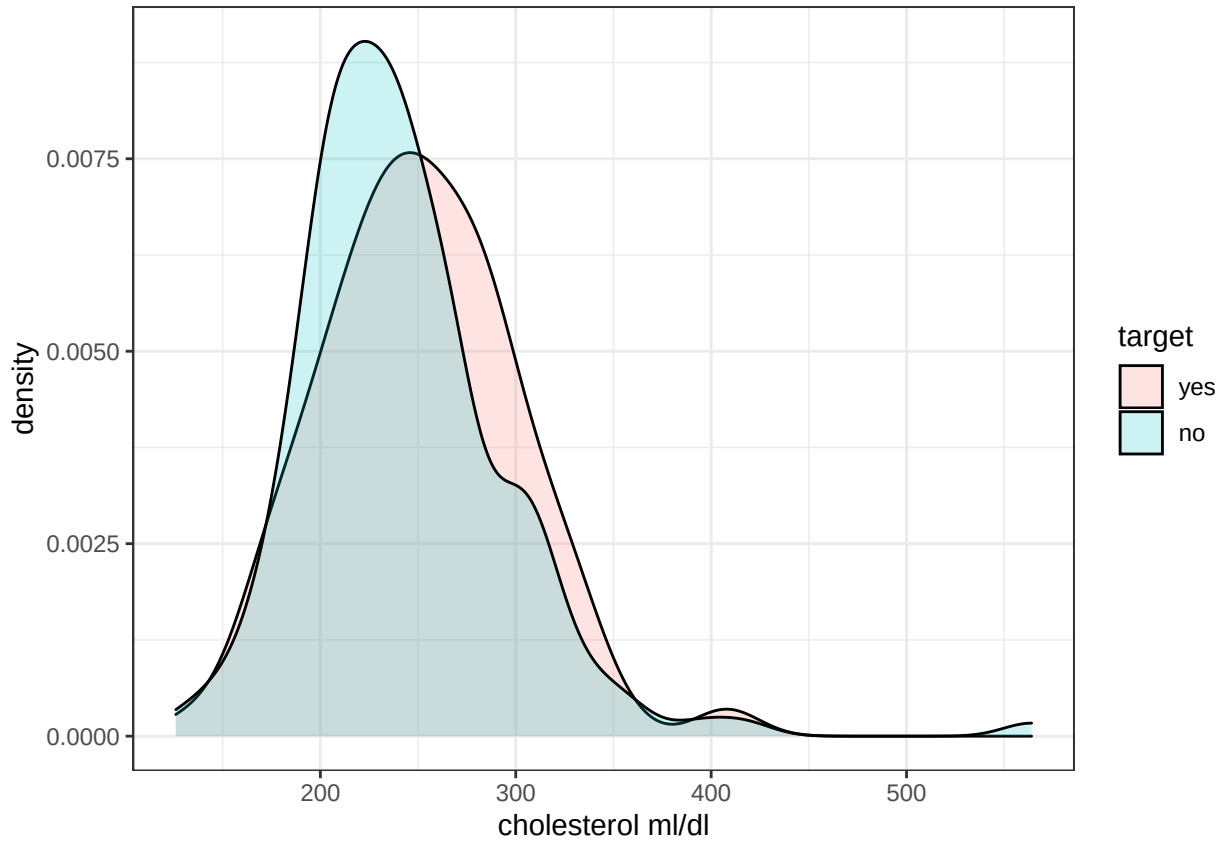
Resting blood pressure (trestbps) The trestbps column contains the resting blood pressure of the patients in millimeters of mercury on admission to the hospital.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    94.0   120.0   130.0   131.6   140.0   200.0
```



Cholesterol (chol) The chol column contains the cholesterol concentration of the patients in mg/dl.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   126.0   211.0   240.0   246.3   274.5   564.0
```



Fasting blood sugar (fbs) This column contains the result of blood sugar test after fasting and separated patients that had sugar levels above 120 mg/dl, the class was changed to factor given the nature of the test and the values were changed to True = 1 or False = 0

```
heart_data <- heart_data %>% mutate(fbs = as.factor(fbs))
levels(heart_data$fbs) <- c("False","True")
summary(heart_data$fbs)
```

```
## False  True
##   258    45
```

```
exp_06 <- heart_data %>% ggplot(aes(x = fbs, fill = target))+
  geom_bar(position = "fill")+
  labs(x = "fasting blood sugar > 120 mg/dl")+
  theme_bw()
```

Resting electrocardiographic results (restecg) The restecg column contains the results of a resting electrocardiographic exam divided into Normal = 0, ST abnormality = 1 , 2 hyperventricular.abnormality = 2. This column was converted into a factor by the following code:

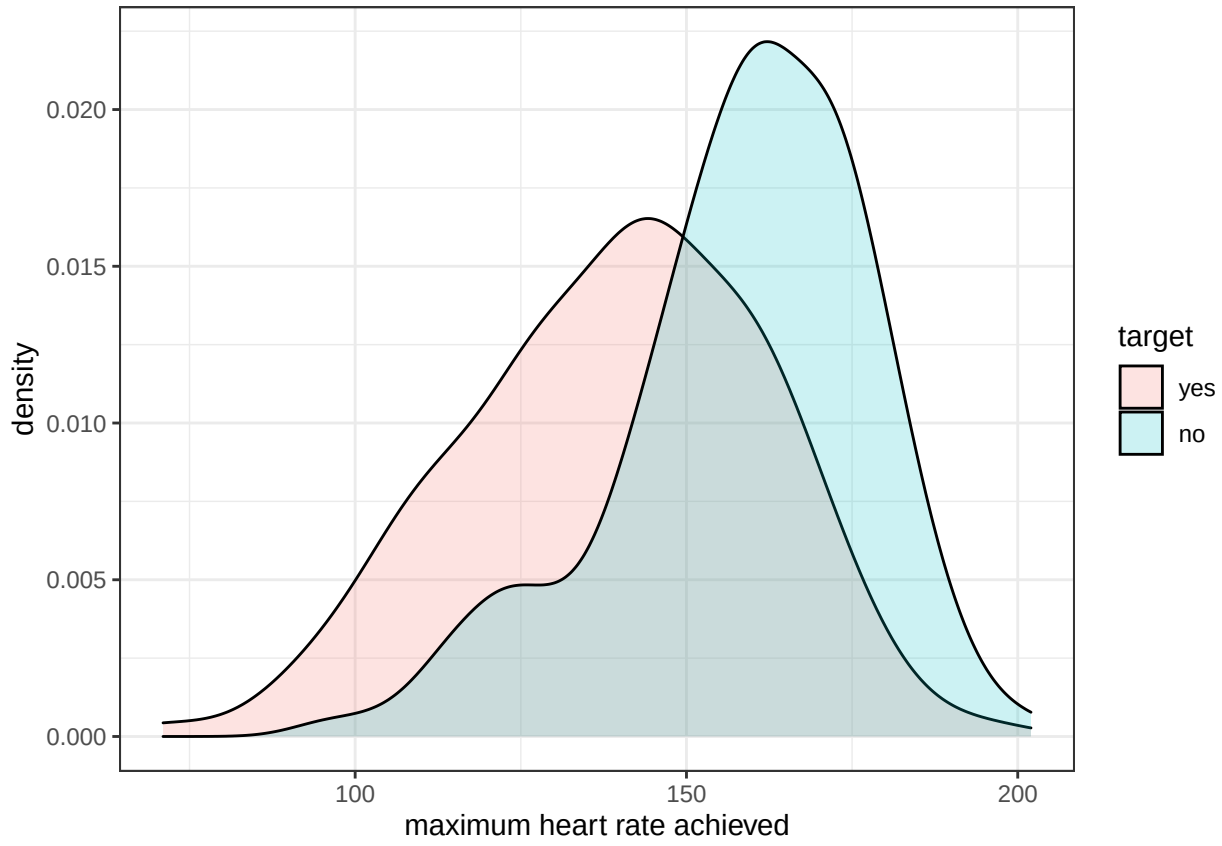
```
heart_data <- heart_data %>% mutate(restecg = as.factor(restecg))
levels(heart_data$restecg) <- c("normal","ST","Hyper")
summary(heart_data$restecg)
```

```
## normal     ST  Hyper
##    147    152      4
```

```
exp_07 <- heart_data %>% ggplot(aes(x = restecg, fill = target))+
  geom_bar(position = "fill")+
  labs(x = "fasting blood sugar > 120 mg/dl")+
  theme_bw()
```

Maximum heart rate achieved (thalach) This column have the maximum heart rate achived by the patients in their hospital time up to the data collection.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    71.0   133.5   153.0   149.6   166.0   202.0
```

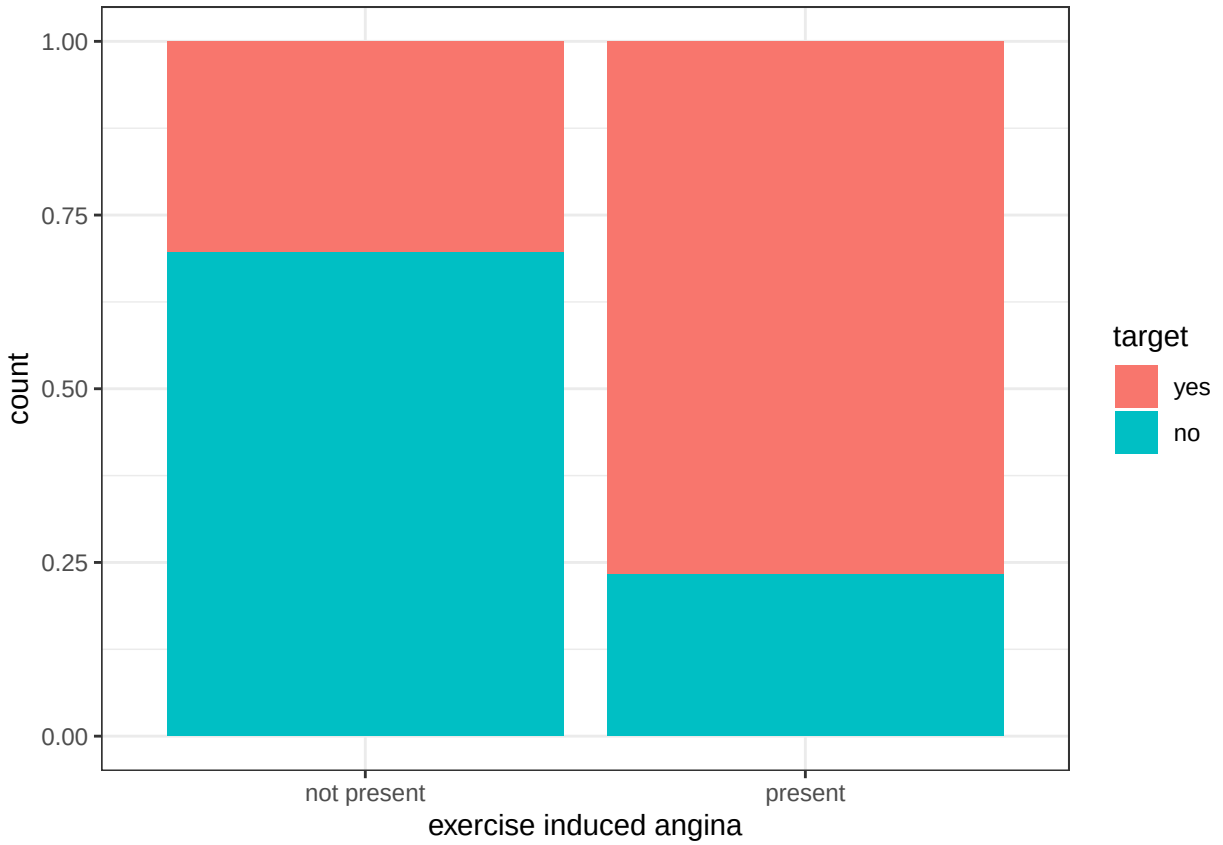


Exercise induced angima (exang) This column is a lists if a patient have angima induced by physical exercise

```
heart_data <- heart_data %>% mutate(exang = as.factor(exang))
levels(heart_data$exang) <- c("not present","present")
summary(heart_data$thalach)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    71.0   133.5   153.0   149.6   166.0   202.0
```

```
exp_09 <- heart_data %>% ggplot(aes(x = exang, fill = target))+
  geom_bar(position = "fill")+
  labs(x="exercise induced angina")+
  theme_bw()
exp_09
```



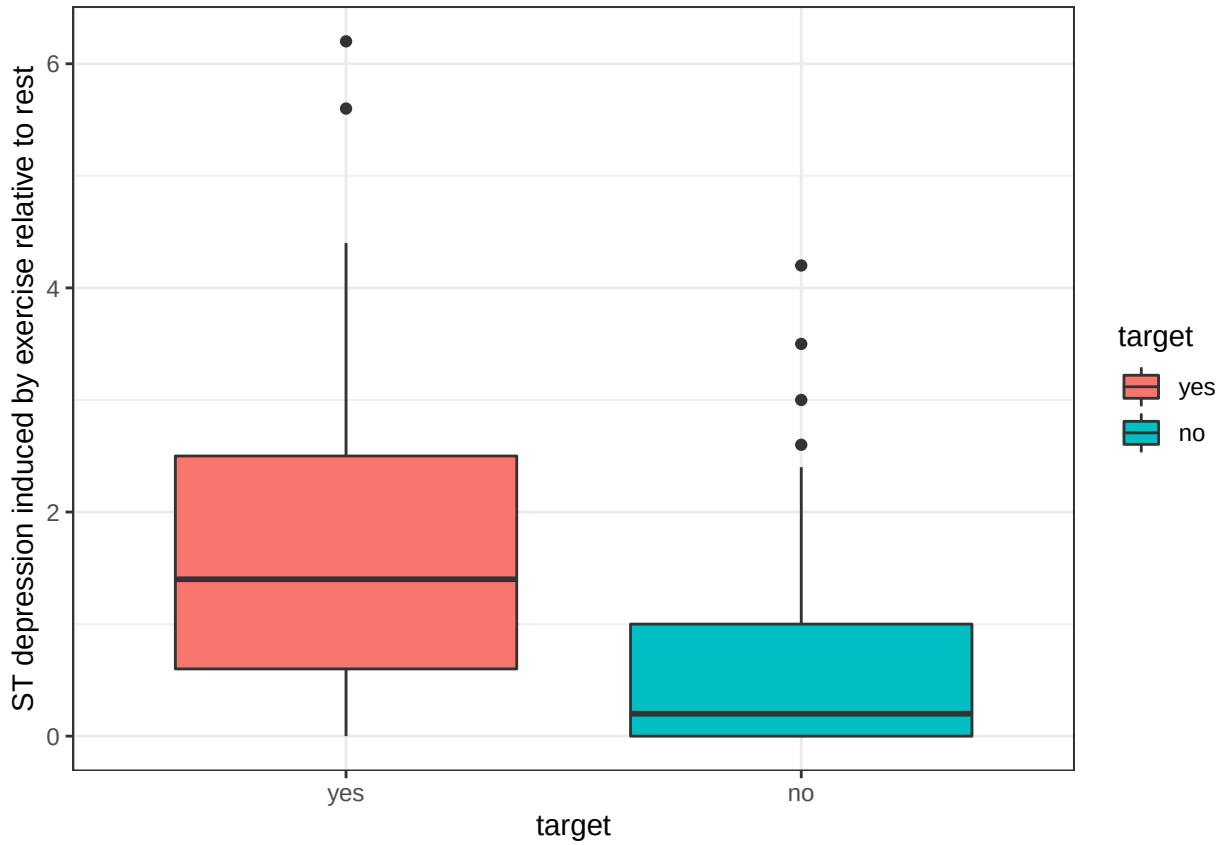
Stress depression induced by exercise relative to rest (oldpeak) This column shows the stress depression induced by physical exercise, measured and registered by the hospital.

```
summary(heart_data$oldpeak)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.80    1.04    1.60    6.20
```

```
exp_10 <- heart_data %>% ggplot(aes(x = target, y = oldpeak, fill = target))+
  geom_boxplot()+
  labs(x = "target", y = "ST depression induced by exercise relative to rest")+
  theme_bw()

exp_10
```



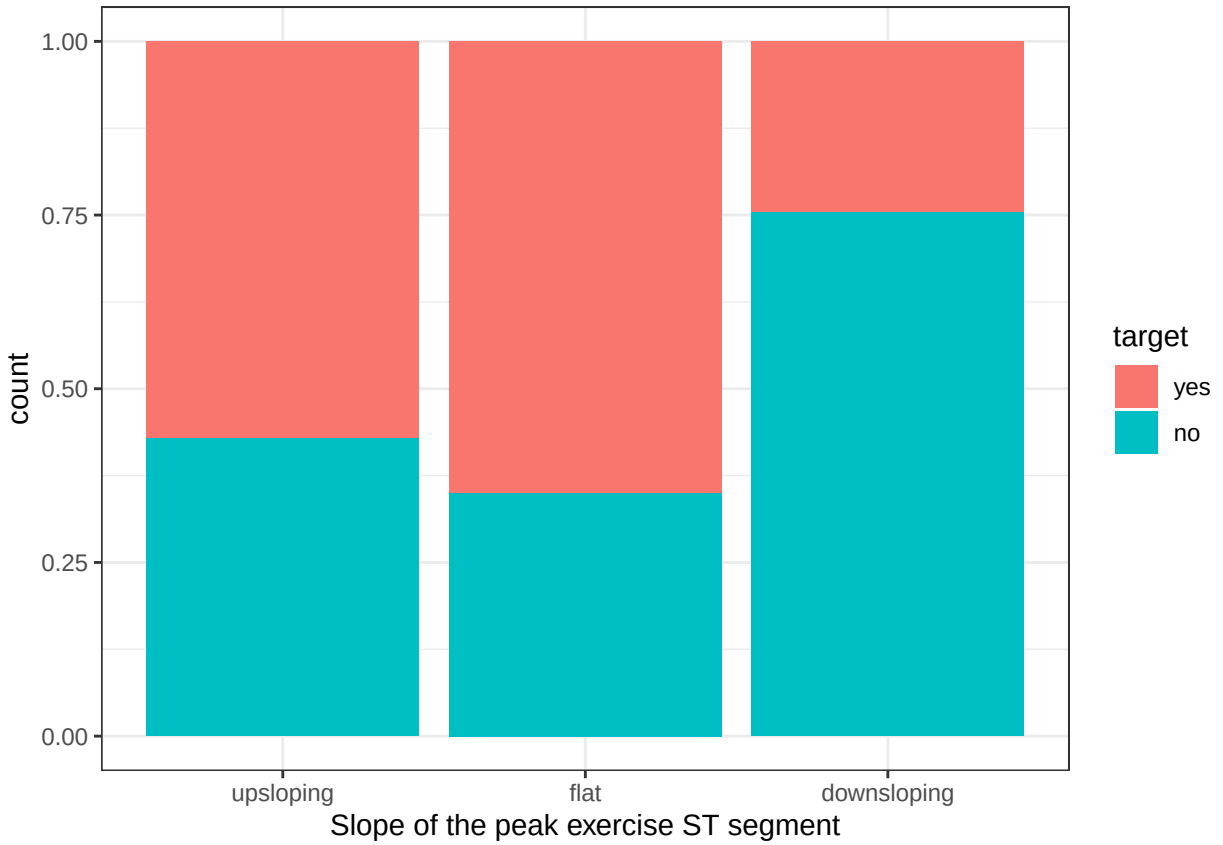
Slope The slope column contains the slope if the heart rate during the peak of the stress section of the exercise examination. This column was converted into a factor (0 = downsloping, 1 = flat, 2 = up sloping) using the following code:

```
heart_data <- heart_data %>% mutate(slope = as.factor(slope))
levels(heart_data$slope) <- c("upsloping","flat","downsloping")
summary(heart_data$slope)
```

```
##   upsloping        flat downsloping
##          21         140         142
```

```
exp_11 <- heart_data %>% ggplot(aes(x = slope, fill = target))+
  geom_bar(position = "fill")+
  labs(x = "Slope of the peak exercise ST segment")+
  theme_bw()

exp_11
```

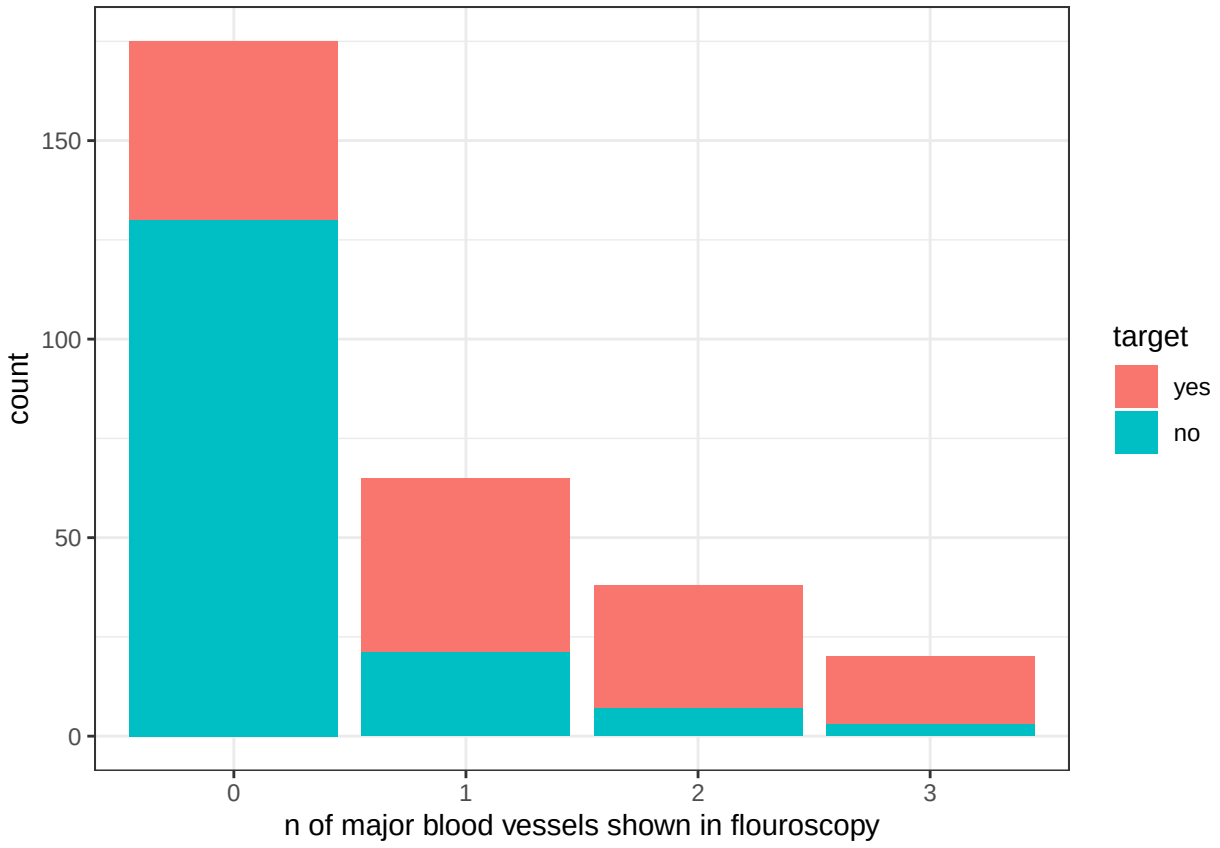


Major blood vessels colored (ca) This column have the number of major blood vessels colored by a flourosopy examination ranging from 0 to 3, in order better study this variable it was converted as is into a factor for exploratory analysis.

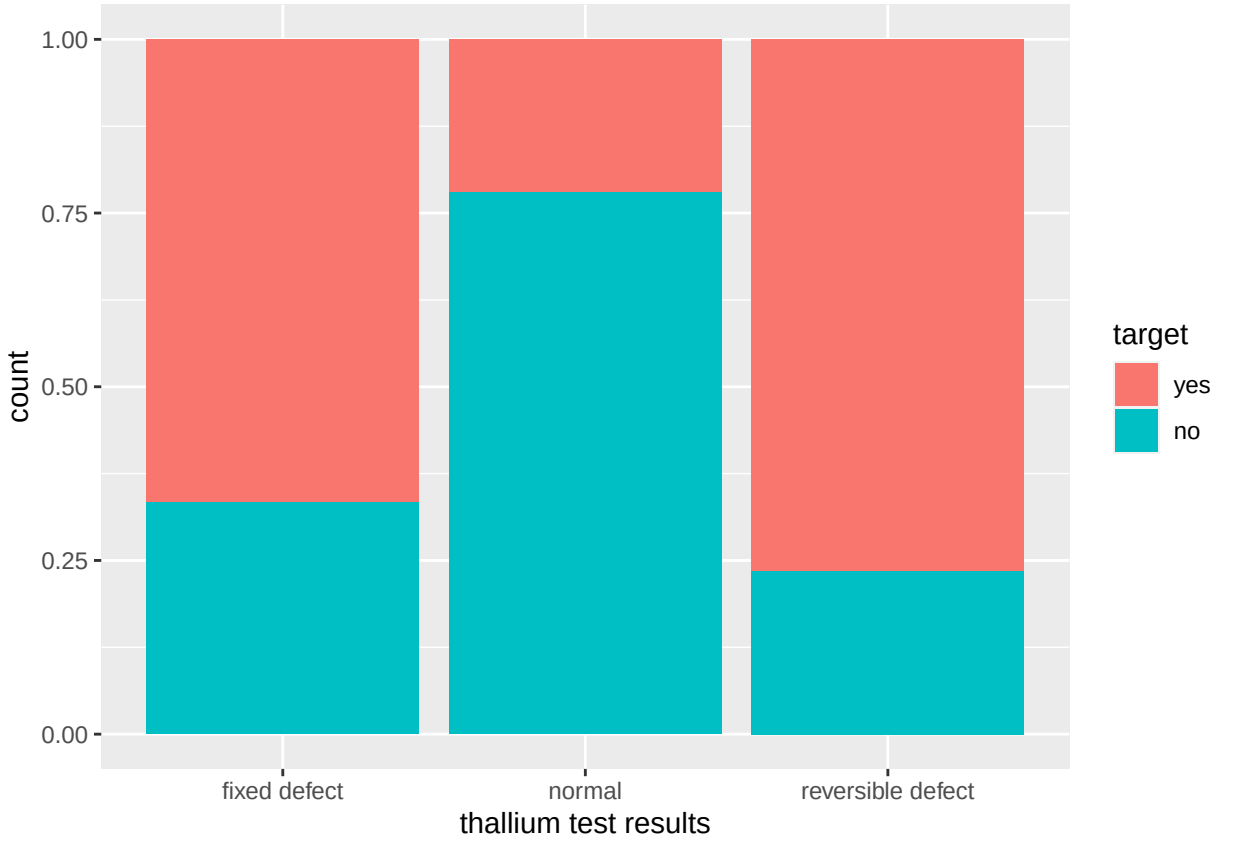
```
heart_data <- heart_data %>% filter(ca != 4)
# the 4 in the columns are NAs

exp_12 <- heart_data %>% ggplot(aes(x = as.factor(ca), fill = target))+
  geom_bar()+
  labs(x = "n of major blood vessels shown in flouroscopy")+
  theme_bw()

exp_12
```



Thallium test result (thal) The thal column contains the result of a thallium imaging exam made in order to detect heart defects, the results were divided in 1 = fixed defect, 2 = normal and 3 = reversible defect.



Visual analysis According to the graphical representations of the variables there is a clear indication that the variables are related to the presence of a heart disease and could be used to predict if the patients have them.

Data sub-setting The first step taken to was to create a validation set, this set comprised of 10% of the data and will be used to validate the models created in this project.

```
#creating validation set

val_index <- createDataPartition(y = heart_data$target, times = 1, p = 0.1, list = FALSE)
validation_set <- heart_data[val_index,]

rm(val_index)
```

The second step was to divide the data into train and test sets in order to develop the machine learning algorithm.

```
test_index <- createDataPartition(y = heart_data$target, times = 1, p = 0.2, list = FALSE)

test_set <- heart_data[test_index,]
train_set <- heart_data[-test_index,]
rm(test_index)
```

Modeling Initially the prediction system was made using logistic regression in order to see there was a strong correlation between the variables and the possible diagnosis.

```
#  logistic regression model

glm_fit <- train(x = train_set[,1:13],y = train_set$target,method = "glm")
```

```
## Warning in (function (kind = NULL, normal.kind = NULL, sample.kind = NULL) :
## non-uniform 'Rounding' sampler used
```

```
## Warning: predictions failed for Resample21: parameter=none Error in model.frame.default(Terms, newda
##   factor restecg has new levels Hyper
```

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo, :
## There were missing values in resampled performance measures.
```

```
glm_acc <- mean(test_set$target == predict(object = glm_fit, newdata = test_set[,1:13],
                                           test_set$target, type = "raw"))
```

The logistic regression showed good results but since it doesn't work much well with higher number of dimensions the next model tested was a decision tree model using the rpart library.

```
# decision tree

rpart_fit <- train(x = train_set[,1:13],y = train_set$target, method = "rpart",
                   tuneGrid= data.frame(cp = seq(0,0.04,len = 20)))

rpart_acc <- mean(test_set$target == predict(object = rpart_fit, newdata = test_set[,1:13],
                                             test_set$target, type = "raw"))
```

In order to enhance our results as random forest model was used with the randomForest library, this will also allow to check the influence of the variables in the calculations.

```
forest_fit <- train(x = train_set[,1:13],y = train_set$target, method = "rf")

forest_acc <- mean(test_set$target == predict(object = forest_fit, newdata = test_set[,1:13],
                                              test_set$target, type = "raw"))
```

The final model tested was an ensemble of the past models aiming to create a more robust prediction algorithm,providing not only a more precise but also less prone to overfitting.

```
# Ensemble

models<- c("glm","rpart","rf")

#training the models

fits <- lapply(models, function(model){
  train(x = train_set[,1:13],y = train_set$target, method = model)
})


#making predictions for each fitted model

names(fits) <- models

p_models <- sapply(fits, function(model){
  predict(object = model, newdata = test_set[,1:13],test_set$target,type="raw")
})
```

```r
probs <- rowMeans(p_models == "yes")
prediction <- ifelse(probs > 0.5,"yes","no")
ens_acc <- mean(prediction == test_set$target)
```

```
## [1] 0.9
```

Results

A quick overview of the graphics show that there is a great correlation between the predictors and the predicted variable, which indicates that is viable to make predictions through machine learning.

The accuracy of the models on the test set are as follow: logistic regression , decision tree , random forest , ensemble . The results of the models in the validation sets were as follow: ensemble , logistic regression , decision tree , random forest .

Conclusion

The project project has shown that it is possible to reliably predict is a patient have or not a heart disease using the given variables, considering that the ensemble and logistic regression generated a 0.9 overall accuracy.

In this project methods such as qda and lda were considered but ultimately failed to provide satisfactory results showing the challenge to generate a prediction given the high number of dimensions.