Block III

Sara Dovalo and Javier Muñoz Flores

21/3/2022

Contents

Introduction

The content of this section is, mainly, a further description of the dataset used as well as the exposition of the problem to solve.

The dataset selected contains several patient records of different medical measurements in order to predict wether a person is more likely to suffer a heart disease, i.e. a failure . The data has been retrieved from the public *kaggle* repository and it is the product of the combination of five different datasets from different regions of EEUU. The final dataset contains in total 918 instances and 12 attributes, which 7 of them are categorical and the five remaining are numerical:

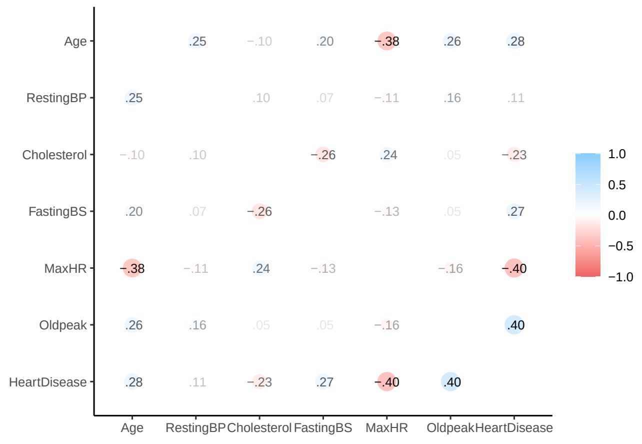
- `Age`(*quantitative*): age of the patient in years
- `Sex`(*qualitative*): sex of the patient [*M*: Male, *F*: Female]
- `ChestPainType`(*qualitative*): Angina type, i.e. chest pain, frequently caused when the heart muscle is not able to get enough oxygen-rich blood [*TA*: Typical Angina, *ATA*: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- `RestingBP`(*quantitative*): resting blood pressure [mm Hg]. A normal level is less than 180 mm Hg.
- `Cholesterol`(*quantitative*): serum cholesterol [mm/dl]
- `FastingBS`(*qualitative*): fasting blood sugar [1: if FastingBS $>$ 120 mg/dl, 0: otherwise]
- `RestingECG`(*qualitative*): resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $>$ 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- `MaxHR`(*quantitative*): maximum heart rate achieved [Numeric value between 60 and 202]
- `ExerciseAngina`: exercise-induced angina [Y: Yes, N: No]
- `Oldpeak`(*quantitative*): oldpeak = ST [Numeric value measured in depression]
- `ST_Slope`(*qualitative*): the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- `HeartDisease`(*qualitative*): class variable [1: heart disease, 0: Normal]

Clearly, it is a binary classification problem since the target variable has two levels.

Preprocessing

First of all, it is suitable to visualize the data and to identify if there are missing values which could add noise and disrupt the performance of the future models created.

We show the correlation between the numeric variables through a nice plot to carry out a first exploratory analysis.



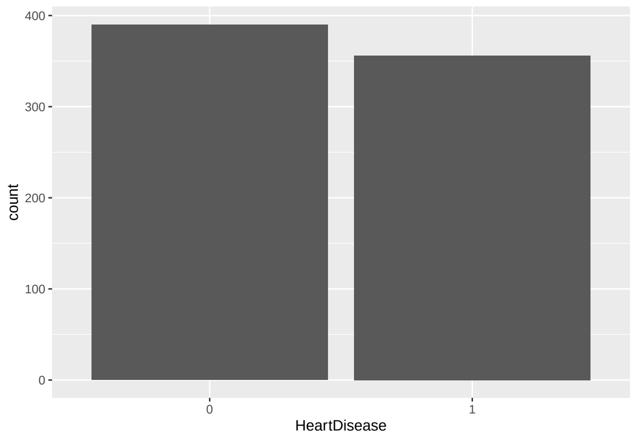
We notice that the variable `Cholesterol` contains 172 values equal to zero. They must be considered as missing values as a person cannot that such a low level of cholesterol in blood. We decide to eliminate those observations that have a zero in that variable.

```r
# Count number of zeros (missing values)
heart.data %>% filter(Cholesterol == 0) %>% summarize(count = n())
# Eliminate rows which contain 0 in variable Cholesterol
heart.data = heart.data %>%
filter(Cholesterol != 0)
# Number of missing values
heart.data %>% filter(Cholesterol == 0) %>% summarize(count = n())
```

Furthermore, since the dataset includes several categorical attributes, it is necessary to transform them into factor variables.

```r
# Categorical variables as factors
heart.data = heart.data %>%
        mutate_each_(funs(factor(.)),c(2,3,6,7,9,11,12))
str(heart.data)
```

To finish this section, we check if the classes of the response `HeartDisease` are balanced or not. We visualize it through a barplot to see it more clearly.



The classes are balanced, it will be not needed to over/undersampling the sample.

Modelling with `H2O` package

We follow the same procedure that we did in classroom for using `h2o.automl()`, i.e. fitting, benchmarking, predicting and explaining, in that order.

Fitting

Before starting with the selection of the models, the *local h2o cluster* is initialized in order to leverage the parallelization in the virtual machine which the package provides and to use h2o functions.

The first step is to convert the data into a `h2o` object and then, to identify the names of the predictors as well as the name of the response.

Then, it is important to mention that there will not be splitting into train and test, since we will use cross-validation metrics on the leaderboard model. Thus, we have to specify the number of folds (in our case will be `nfolds = 8`) needed to the cross-validation process. However, we do not have to indicate the predictors names, instead only the `dataframe` and the response variable. In addition, we do not exclude any algorithm in `h2o.automl()` method, but we limit the time for fitting the models (`max_runtime_secs = 30` and `max_runtime_secs_per_model = 5`).

Benchmarking

In this section, we have to explore the leaderboard of models in order to carry out a first comparison. We visualize the errors associated to each of the models.

```
# Leaderboard
lead_h <- model_h@leaderboard
names(lead_h)[5] <- "mpce" # Rename mean_per_class_error to shorten output
print(lead_h[2, -6], n = nrow(lead_h)) # Exclude final column to fit the table in one page
```

```
## [1] model_id auc      logloss  aucpr    mpce     mse
## <0 rows> (or 0-length row.names)
##
## [0 row x 6 columns]
```

The leader has been a *StackedEnsemble* model with the lowest errors.

Prediction

We select the leader of the models and predict a few rows in order to show the probabilities associated with each class, since it is a binary prediction problem.

```
h2o.predict(object = model_h@leader, newdata = data.h[1:2, ])
```

Explanation

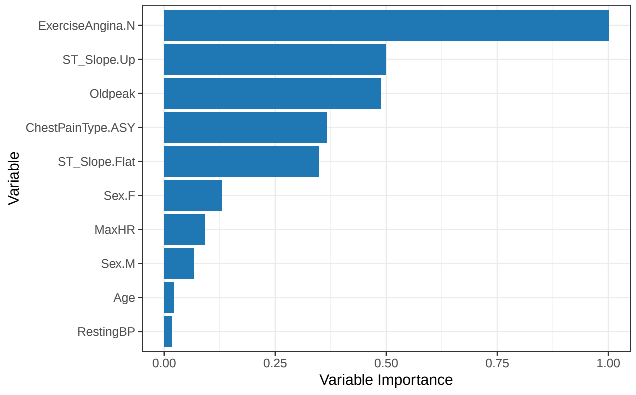
Because of the short report we must to hand in, it has not be possible to do a large interpretation of the plot results. However, it is suitable to identify what are the variables most important for the model and the correlation of them with the response. Thus, we use SHAP plot and variable importance plot to carry out a short explanation.

```
exp <- h2o.explain(object = model_h, newdata = data.h)
exp
```

```
##
##
## Confusion Matrix
## ================
##
## > Confusion matrix shows a predicted class vs an actual class.
##
##
##
```

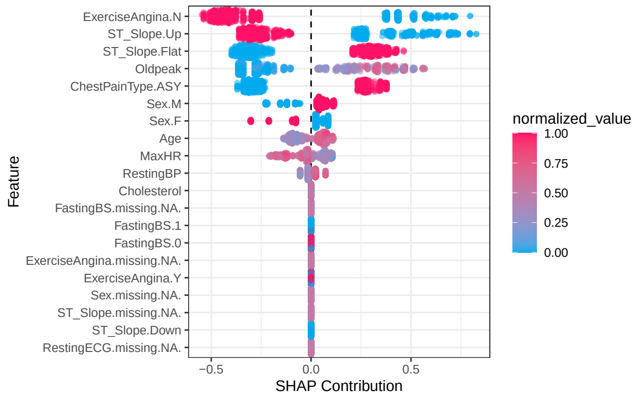
```
## XGBoost_1_AutoML_21_20220324_232221
## -----------------------------------
##
## |  | 0 | 1 | Error | Rate
## |:---:|:---:|:---:|:---:|:---:|
## | **0** |338 | 52 | 0.133333333333333 |  =52/390 |
## | **1** |41 | 315 | 0.115168539325843 |  =41/356 |
## | **Totals** |379 | 367 | 0.124664879356568 |  =93/746 |
##
##
## Variable Importance
## ===================
##
## > The variable importance plot shows the relative importance of the most important variables in the 
```

Variable importance
for "XGBoost_1_AutoML_21_20220324_232221"



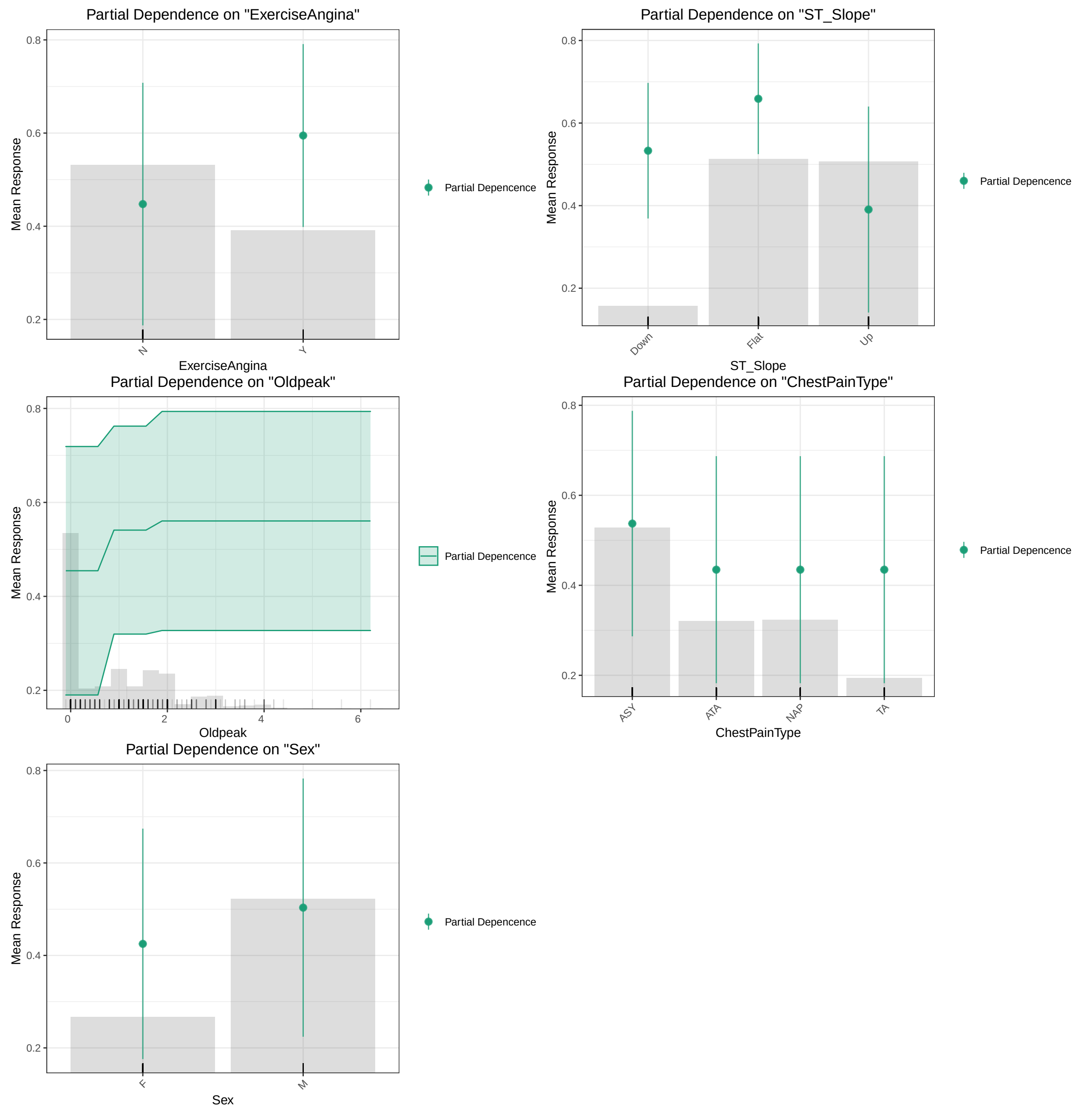
```
##
##
## SHAP Summary
## ============
##
## > SHAP summary plot shows the contribution of the features for each instance (row of data). The sum 
```

Summary Plot
for "XGBoost_1_AutoML_21_20220324_232221"



```
##
```

```
##
## Partial Dependence Plots
## ========================
##
## > Partial dependence plot (PDP) gives a graphical depiction of the marginal effect of a variable on
```



```
h2o.shutdown(prompt = FALSE)
```

Clearly, the most important variable is `ST_Slope`, an important feature which can be seen in an electrocardiogram. As expected, these unbiased features related with the functionality of the heart have an crucial role in the prediction.

SHAP plot, for instance, reveals that women are more correlated with the response than men, a point that is not so clear in a first view.

Modelling with `tidymodels` package

`Tidymodels`is an interface that unifies hundreds of functions from different packages, facilitating all stages of pre-processing, training, optimization and validation of predictive models.

This package offers so many possibilities that they can hardly be shown with a single example.

The first step is to split the database into two subsets, the one used for training and the one used for validation. This is done with the `initial_split()` command of the `rsample` package. The training dataset, the one used for training the model and the one used to validate the model metrics can then be separated to help the selection, from a set of models applied on the same data, that one which performs best. It could be separated into two subsets (training and test), but it is a better practice to perform a cross-validation, so this one will be applied.

```r
# Partition on training and test
heart_split <- initial_split(heart.data, prop = .8)
data_train <- training(heart_split)
data_validate <- testing(heart_split)
```

For simplicity we only consider 7 of the 11 explanatory variables.

```r
variables <- c("Age", "Sex", "RestingBP", "Cholesterol", "FastingBS",
               "RestingECG", "MaxHR", "HeartDisease")
data_train <- data_train %>% dplyr::select(all_of(variables))
data_validate <- data_validate %>% dplyr::select(all_of(variables))
```

Logistic model

Build a model

As the variable of interest `HeartDisease` is a binary variable, a logistic regression model is chosen, which will be our basis. An object shall be created that stores the formula for later use:

```r
glm_fit <- glm(HeartDisease ~ Age + Sex + RestingBP + Cholesterol + FastingBS +
                RestingECG + MaxHR , data = data_train, family = binomial)
glm_fit_formula <- as.formula(HeartDisease ~ Age + Sex + RestingBP + Cholesterol +
                FastingBS + RestingECG + MaxHR)
```

```r
tidy(glm_fit)
```

```
## # A tibble: 9 x 5
##   term             estimate std.error statistic       p.value
##   <chr>               <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)      -1.59      1.34       -1.19  0.233
## 2 Age               0.0330    0.0118      2.79  0.00520
## 3 SexM              1.50      0.246       6.09  0.00000000110
## 4 RestingBP         0.0126    0.00580     2.18  0.0293
## 5 Cholesterol       0.00420   0.00174     2.41  0.0158
## 6 FastingBS1        0.300     0.271       1.11  0.268
## 7 RestingECGNormal -0.536     0.242      -2.21  0.0269
## 8 RestingECGST     -0.120     0.317      -0.378 0.705
## 9 MaxHR            -0.0271    0.00448    -6.04  0.00000000153
```

With the `broom::tidy()` object that returns information about the components of the model we can clearly see that men are more likely to have heart disease. Furthermore, the positive sign of the coefficients associated with the variables `Age`, `RestingBP` and `Cholesterol` means that they are positively correlated positively with our response variable, `HeartDisease`, that is, a person with an advanced age and a high cholesterol level will be more likely to have a heart disease, as expected.

However, the p-value associated with the variables `FastingBS1` and `RestingECGST` is quite large which means that perhaps these variables are not statistically significant in predicting whether a person will have heart disease or not.

Preprocess our data with recipes

The `resample` package will be used to evaluate the model. Ten cross-validations, each of 8 *folds*, will be used, therefore 80 samples will be obtained to evaluate the accuracy of the model.

```r
set.seed(1234)
folds_with_repeats <- rsample::vfold_cv(data = data_train, v = 8, repeats = 10)
print(folds_with_repeats)
```

```
## #  8-fold cross-validation repeated 10 times
## # A tibble: 80 x 3
##    splits            id       id2
##    <list>            <chr>    <chr>
##  1 <split [521/75]> Repeat01 Fold1
##  2 <split [521/75]> Repeat01 Fold2
##  3 <split [521/75]> Repeat01 Fold3
##  4 <split [521/75]> Repeat01 Fold4
##  5 <split [522/74]> Repeat01 Fold5
##  6 <split [522/74]> Repeat01 Fold6
##  7 <split [522/74]> Repeat01 Fold7
##  8 <split [522/74]> Repeat01 Fold8
##  9 <split [521/75]> Repeat02 Fold1
## 10 <split [521/75]> Repeat02 Fold2
## # ... with 70 more rows
```

The command `rsample::vfold_cv()` generates the 80 subsets of equal size, varying the seed of the subdivision. The result generates an object with three variables: the repetition (`id`), the fold (`id2`) and splits. `splits` is a variable that stores a list of size ($v$repeats*)*3, where v is the number of folds and *repeat* the number of repetitions. Each element of the list is an object of the tibble class with several variables:

- `data`: matrix of dimension n*p, where n is the number of records and p the number of variables.
- `in_id`: logical index of the position of the records in the data that belong to the records being analysed. This set of records is called analisys, and the one left out is called assessment.
- `id`: stores the id of the fold and of the repetition to which the data corresponds.

Prediction

Next, to assess the accuracy of the model for each of the samples generated with the cross-validation we will create a function that will mainly do the following:

1. A logistic regression and model fit for the *analysis* data.
2. Prediction on the *assessment* data using one of the most prominent packages of the `tidymodels` package, `broom`.
3. Checking whether the prediction was performed correctly.

```
res_leftout <- function(samplecv, model) {
  # Fit the model
  glm_model <- glm(model, data = analysis(samplecv), family = binomial)

  # Identify the dataset left out
  holdout <- assessment(samplecv)

  # Perfoms prediction on the data set hold out using augment()
  res <- broom::augment(glm_model, newdata = holdout)

  # lvls will be the levels of the factor with the predictions
  # the prediction (res$.fitted) is transformed into a nominal variable
  # depending on whether it is greater or less than zero
  # If greater than zero, then No (no heart attack), otherwise Yes.
  lvls <- levels(holdout$HeartDisease)
  predictions <- factor(ifelse(res$.fitted > 0, lvls[2], lvls[1]), levels = lvls)

  # We check if the prediction is correct
  res$correct <- predictions == holdout$HeartDisease

  # Return the dataset with the additional columns
  res
}
```

For example, if we implement this procedure for the first of the 80 samples generated we obtain:

```
firstsample <- res_leftout(folds_with_repeats$splits[[1]], glm_fit_formula)
dim(firstsample)
```

```
## [1] 75 11
```

```
dim(rsample::assessment(folds_with_repeats$splits[[1]]))
```

```
## [1] 75  8
```

Therefore, the name of the columns added for the first of the samples is:

```
print(firstsample[1:7, setdiff("correct", names("HeartDisease"))])
```

```
## # A tibble: 7 x 1
##   correct
##   <lgl>
## 1 TRUE
## 2 TRUE
## 3 FALSE
## 4 TRUE
## 5 TRUE
## 6 FALSE
## 7 FALSE
```

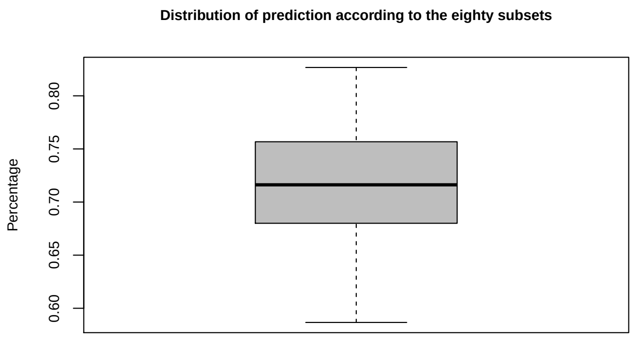
As we can see, half of the predictions in the first sample are incorrect. For this particular model, `.fitted` refers to the linear predictor of the *log-odds*.

To do the above with the rest of the 79 samples obtained, we will apply the `purrr::map()` function, which applies the requested on each list:

```
folds_with_repeats$results <- purrr::map(folds_with_repeats$splits, res_leftout,
                                glm_fit_formula)
```

We can now calculate the metric for all *assessment* datasets. The percentage of heart diseases with a correct prediction is calculated:

```
folds_with_repeats$results <- map_dbl(folds_with_repeats$results, function(x) {mean(x$correct)})
boxplot(folds_with_repeats$results,
        main = 'Distribution of prediction according to the eighty subsets',
        col = 'grey', ylab = 'Percentage', cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.8)
```



The accuracy to be exceeded from this baseline is 0.73.

Classification: Implementation of k-means algorithm

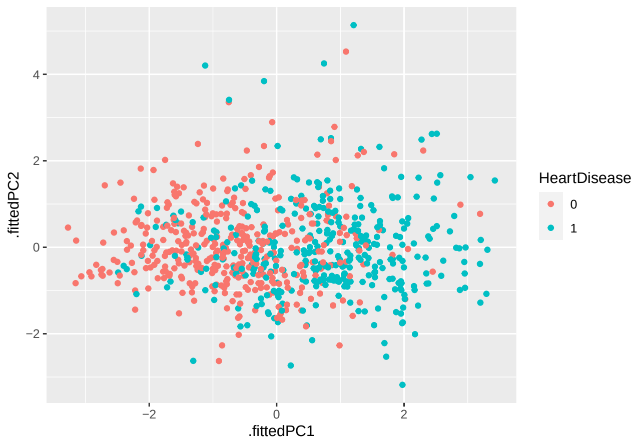
We attempt to implement a new method in order to achieve a further analysis of the data. Before that, we apply PCA and select the principal components which explain most of the variability of the dataset. To apply both techniques, it is compulsory to select the numeric attributes and highly recommendable to scale them.

```
# PCA components
pca_comp <- heart.data %>%
select(where(is.numeric)) %>%
prcomp(scale. = TRUE)
tidy(pca_comp)
```

```
## # A tibble: 3,730 x 3
##      row    PC   value
##    <int> <dbl>   <dbl>
##  1     1     1 -1.56
##  2     1     2  1.40
##  3     1     3 -0.654
##  4     1     4  0.0201
##  5     1     5 -0.347
##  6     2     1 -0.0289
##  7     2     2 -0.255
##  8     2     3 -2.00
##  9     2     4  0.162
## 10     2     5 -0.343
## # ... with 3,720 more rows
```

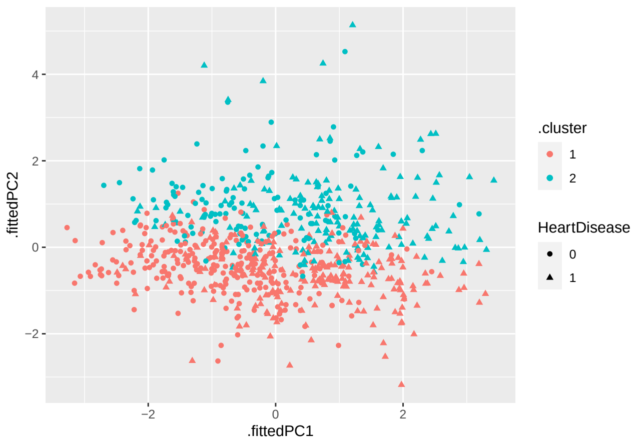
First, we plot the first two PCA components and visualize if the classes form differentiated clusters. By using the `broom::augment()` function, we can complement the results as the PCA technique as the *k-means* algorithm. It is very useful in order to plot the categorization *k-means* in terms of the first two PCA components.

```
aug_pca <- augment(pca_comp, data = heart.data)
ggplot(data = aug_pca, aes(x = .fittedPC1, y = .fittedPC2)) +
geom_point(aes(col = HeartDisease))
```



We cannot distinguish two groups in the data. We apply *k-means* method and again visualize through the first two PCA components the division in clusters which has carried out the algorithm.

```
k_m <- heart.data %>%
select(where(is.numeric)) %>%
kmeans(centers = 2)
aug_k_m <- augment(k_m, data = heart.data)
aug_k_m <- cbind(aug_k_m, aug_pca[".fittedPC1"], aug_pca[".fittedPC2"])
ggplot(data = aug_k_m, mapping = aes(x = .fittedPC1, y = .fittedPC2)) +
geom_point(aes(col = .cluster, pch = HeartDisease))
```



It seems that now there is less overlapping, if we look at the cluster *k-means* division. It suggests that cluster *k-means* division may distinguish two different population in a better way than the classes do it.