Model Selection

Eduardo Yuki Yada

Imports

```
library(tidyverse)
library(yaml)
library(tidymodels)
library(usemodels)
library(vip)
library(bonsai)
library(lightgbm)
```

Minutes to run: 0

Loading data

```
load('../dataset/processed_data.RData')
load('../dataset/processed_dictionary.RData')

columns_list <- yaml.load_file("./auxiliar/columns_list.yaml")

outcome_column <- params$outcome_column
features_list <- params$features_list</pre>
```

Minutes to run: 0.002

Filtering eligible pacients

```
df = df %>%
  filter(disch_outcomes_t0 == 0)

df %>% dim

## [1] 15766 239

Minutes to run: 0.008
```

Eligible features

```
'suporte_hemod' # com proced_invasivos_qtde
eligible_features = eligible_columns %>%
  base::intersect(c(columns_list$categorical_columns, columns_list$numerical_columns)) %>%
  setdiff(c(exception_columns, correlated_columns))
if (is.null(features_list)) {
  features = eligible_features
} else {
  features = base::intersect(eligible_features, features_list)
}
gluedown::md_order(features, seq = TRUE, pad = TRUE)
## 01. sex
## 02. age
## 03. education_level
## 04. patient_state
## 05. underlying_heart_disease
## 06. heart disease
## 07. nyha_basal
## 08. prior_mi
## 09. heart_failure
## 10. af
## 11. cardiac_arrest
## 12. transplant
## 13. valvopathy
## 14. endocardites
## 15. diabetes
## 16. renal_failure
## 17. hemodialysis
## 18. copd
## 19. comorbidities_count
## 20. procedure_type_1
## 21. reop_type_1
## 22. procedure_type_new
## 23. cied_final_1
## 24. cied_final_group_1
## 25. admission_t0
## 26. admission_pre_t0_180d
## 27. icu_t0
## 28. dialysis_t0
## 29. disch_outcomes_t0
## 30. admission_t0_emergency
## 31. aco
## 32. antiarritmico
## 33. betabloqueador
## 34. ieca_bra
## 35. dva
## 36. digoxina
## 37. estatina
## 38. diuretico
## 39. vasodilatador
## 40. insuf cardiaca
## 41. espironolactona
## 42. bloq_calcio
## 43. antiplaquetario_ev
## 44. insulina
## 45. anticonvulsivante
## 46. psicofarmacos
```

```
## 47. antifungico
## 48. antiviral
## 49. antiretroviral
## 50. classe_meds_qtde
## 51. meds_cardiovasc_qtde
## 52. meds_antimicrobianos
## 53. vni
## 54. cec
## 55. transplante_cardiaco
## 56. outros_proced_cirurgicos
## 57. icp
## 58. intervencao_cv
## 59. angioplastia
## 60. cateterismo
## 61. eletrofisiologia
## 62. cateter_venoso_central
## 63. proced_invasivos_qtde
## 64. cve_desf
## 65. transfusao
## 66. interconsulta
## 67. equipe_multiprof
## 68. ecg
## 69. holter
## 70. teste_esforco
## 71. espiro_ergoespiro
## 72. tilt_teste
## 73. metodos_graficos_qtde
## 74. laboratorio
## 75. cultura
## 76. analises_clinicas_qtde
## 77. citologia
## 78. biopsia
## 79. histopatologia_qtde
## 80. angio_rm
## 81. angio_tc
## 82. cintilografia
## 83. ecocardiograma
## 84. endoscopia
## 85. flebografia
## 86. pet_ct
## 87. ultrassom
## 88. tomografia
## 89. radiografia
## 90. ressonancia
## 91. exames_imagem_qtde
## 92. bic
## 93. mpp
Minutes to run: 0
```

Train test split (70%/30%)

Minutes to run: 0.005 Minutes to run: 0

Global parameters

Minutes to run: 0

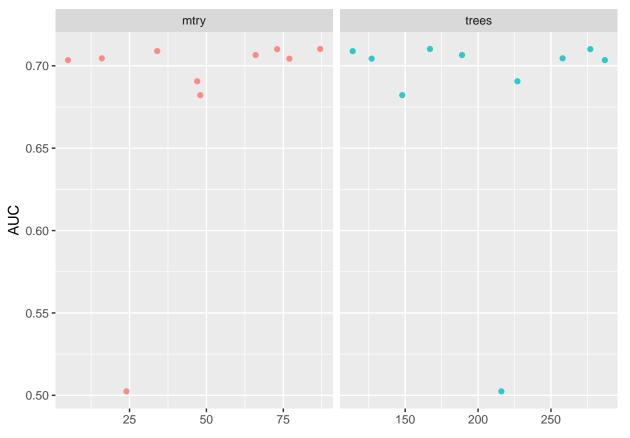
Functions

```
validation = function(model_fit, new_data, plot=TRUE) {
  library(pROC)
 library(caret)
 test_predictions_prob <-</pre>
    predict(model_fit, new_data = new_data, type = "prob") %>%
    rename_at(vars(starts_with(".pred_")), ~ str_remove(., ".pred_")) %>%
    .$`1`
 pROC_obj <- roc(</pre>
    new_data[[outcome_column]],
    test_predictions_prob,
    direction = "<",
    levels = c(0, 1),
    smoothed = TRUE,
    ci = TRUE,
    ci.alpha = 0.9,
    stratified = FALSE,
    plot = plot,
    auc.polygon = TRUE,
   max.auc.polygon = TRUE,
    grid = TRUE,
    print.auc = TRUE,
    show.thres = TRUE
  )
  test_predictions_class <-</pre>
    predict(model_fit, new_data = new_data, type = "class") %>%
    rename_at(vars(starts_with(".pred_")), ~ str_remove(., ".pred_")) %>%
    .$class
  conf_matrix <- table(test_predictions_class, new_data[[outcome_column]])</pre>
  if (plot) {
    sens.ci <- ci.se(pROC_obj)</pre>
    plot(sens.ci, type = "shape", col = "lightblue")
    plot(sens.ci, type = "bars")
    confusionMatrix(conf_matrix) %>% print
 }
```

```
return(pROC_obj)
}
```

Boosted Tree (XGBoost)

```
xgboost_recipe <-</pre>
  recipe(formula = sprintf("%s ~ .", outcome_column) %% as.formula, data = df_train) %>%
  step_novel(all_nominal_predictors()) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_other(all_nominal_predictors(), threshold = 0.05, other=".merged") %>%
  step_dummy(all_nominal_predictors(), one_hot = TRUE) %>%
  step_zv(all_predictors())
xgboost_spec <- boost_tree(</pre>
 mtry = tune(),
 trees = tune(),
 min_n = tune(),
 tree_depth = tune(),
 learn_rate = tune(),
  loss_reduction = tune()
) %>%
  set_engine("xgboost") %>%
  set_mode("classification")
xgboost_grid <- grid_latin_hypercube(</pre>
  finalize(mtry(), df_train),
  dials::trees(range = c(100L, 300L)),
 min_n(),
 tree_depth(),
 learn_rate(),
 loss_reduction(),
  size = grid_size
xgboost_workflow <-</pre>
  workflow() %>%
  add_recipe(xgboost_recipe) %>%
  add_model(xgboost_spec)
xgboost_tune <-
  xgboost_workflow %>%
  tune_grid(resamples = df_folds,
            grid = xgboost_grid)
xgboost_tune %>%
  show_best("roc_auc")
## # A tibble: 5 x 12
     mtry trees min_n tree_depth learn_rate loss_r~1 .metric .esti~2 mean
                                                                                   n std_err .config
##
     <int> <int> <int>
                            <int>
                                         <dbl>
                                                  <dbl> <chr>
                                                                <chr>
                                                                         <dbl> <int>
                                                                                       <dbl> <chr>
## 1
        87 167
                    27
                                       4.88e-2 6.27e- 6 roc auc binary 0.710
                                                                                   4 0.00927 Prepro~
## 2
        73 277
                                7
                                       4.98e-3 2.69e- 2 roc_auc binary 0.710
                    36
                                                                                   4 0.00693 Prepro~
## 3
        34
             114
                    25
                                8
                                       4.21e-4 2.18e-10 roc_auc binary 0.709
                                                                                   4 0.00691 Prepro~
        66
             189
                    31
                                9
## 4
                                       1.57e-4 1.03e- 7 roc_auc binary 0.707
                                                                                   4 0.00792 Prepro~
             258
                               14
                                       4.26e-8 5.95e+ 0 roc_auc binary 0.705
                                                                                   4 0.00767 Prepro~
## # ... with abbreviated variable names 1: loss_reduction, 2: .estimator
best_xgboost <- xgboost_tune %>%
 select_best("roc_auc")
```

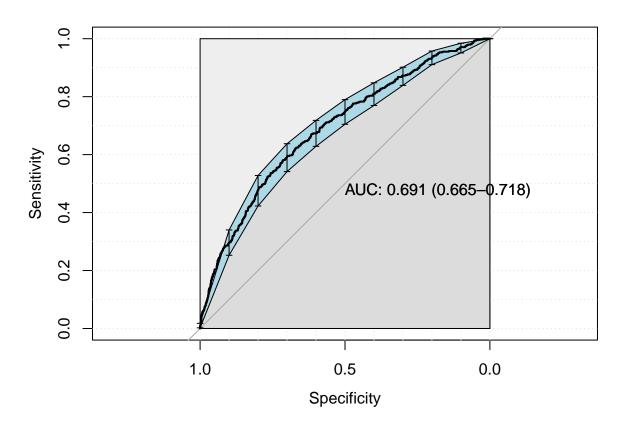


```
final_xgboost_workflow <-
    xgboost_workflow %>%
    finalize_workflow(best_xgboost)

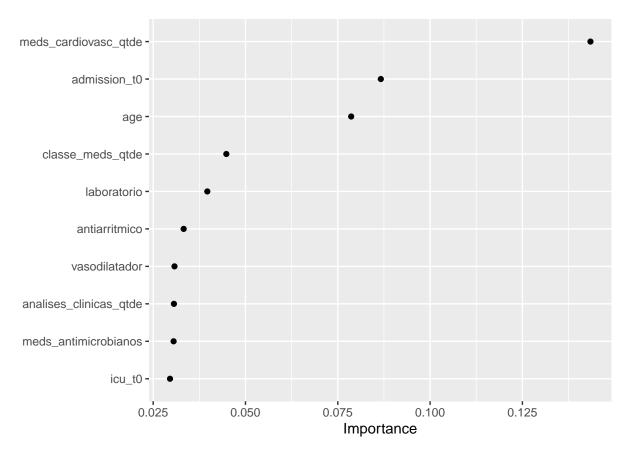
last_xgboost_fit <-
    final_xgboost_workflow %>%
    last_fit(df_split)

final_xgboost_fit <- extract_workflow(last_xgboost_fit)

xgboost_auc <- validation(final_xgboost_fit, df_test)</pre>
```



```
Confusion Matrix and Statistics
##
##
##
   test_predictions_class
                                   1
                        0 4265
##
                                 434
##
                        1
                             18
                                  13
##
                  Accuracy : 0.9044
##
                    95% CI: (0.8957, 0.9127)
##
       No Information Rate: 0.9055
##
       P-Value [Acc > NIR] : 0.6101
##
##
##
                     Kappa: 0.0427
##
##
    Mcnemar's Test P-Value : <2e-16
##
##
               Sensitivity: 0.99580
##
               Specificity: 0.02908
##
            Pos Pred Value: 0.90764
##
            Neg Pred Value: 0.41935
                Prevalence: 0.90550
##
##
            Detection Rate: 0.90169
##
      Detection Prevalence: 0.99345
##
         Balanced Accuracy: 0.51244
##
##
          'Positive' Class : 0
##
final_xgboost_fit %>%
  fit(data = df_train) %>%
  extract_fit_parsnip() %>%
  vip(geom = "point")
```



```
xgboost_parameters <- xgboost_tune %>%
    show_best("roc_auc", n=1) %>%
    select(trees, mtry, min_n, tree_depth, learn_rate, loss_reduction) %>%
    as.list

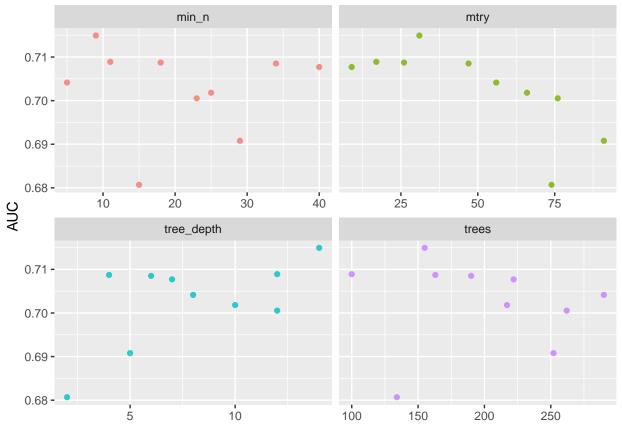
saveRDS(
    xgboost_parameters,
    file = sprintf(
        "../EDA/auxiliar/hyperparameters/model_selection/xgboost_parameters_%s.rds",
        outcome_column
    )
)
```

Boosted Tree (LightGBM)

```
lightgbm_recipe <-
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula, data = df_train) %>%
  step_novel(all_nominal_predictors()) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_other(all_nominal_predictors(), threshold = 0.05, other=".merged") %>%
  step_impute_mean(all_numeric_predictors()) %>%
  step_zv(all_predictors())

lightgbm_spec <- boost_tree(
  mtry = tune(),
  trees = tune(),
  min_n = tune(),
  tree_depth = tune(),
  learn_rate = tune(),
  loss_reduction = tune(),
  sample_size = 1</pre>
```

```
) %>%
  set_engine("lightgbm") %>%
  set mode("classification")
lightgbm_grid <- grid_latin_hypercube(</pre>
  finalize(mtry(), df_train),
  dials::trees(range = c(100L, 300L)),
  min_n(),
  tree_depth(),
  learn_rate(),
  loss_reduction(),
  size = grid_size
lightgbm_workflow <-</pre>
  workflow() %>%
  add_recipe(lightgbm_recipe) %>%
  add_model(lightgbm_spec)
lightgbm tune <-
  lightgbm_workflow %>%
  tune_grid(resamples = df_folds,
            grid = lightgbm_grid)
lightgbm_tune %>%
  show_best("roc_auc")
## # A tibble: 5 x 12
##
     mtry trees min_n tree_depth learn_rate loss_re~1 .metric .esti~2 mean
                                                                                  n std_err .config
##
     <int> <int> <int>
                            <int>
                                       <dbl>
                                                 <dbl> <chr>
                                                                <chr>
                                                                        <dbl> <int>
                                                                                      <dbl> <chr>
## 1
                                                                                  4 0.00583 Prepro~
        31
             155
                    9
                               14
                                    6.42e- 3 1.75e- 3 roc_auc binary 0.715
             100
                               12
## 2
        17
                    11
                                   1.02e-10 1.41e- 1 roc_auc binary 0.709
                                                                                  4 0.00669 Prepro~
## 3
        26 163
                    18
                               4
                                    1.45e- 9 1.11e+ 1 roc_auc binary 0.709
                                                                                  4 0.00709 Prepro~
## 4
        47
             190
                    34
                                6
                                    9.70e- 9 1.51e- 6 roc_auc binary 0.709
                                                                                  4 0.00701 Prepro~
## 5
         9
             222
                    40
                                7
                                    1.19e- 7 2.34e-10 roc_auc binary 0.708
                                                                                  4 0.00560 Prepro~
## # ... with abbreviated variable names 1: loss_reduction, 2: .estimator
best_lightgbm <- lightgbm_tune %>%
  select_best("roc_auc")
lightgbm_tune %>%
  collect_metrics() %>%
  filter(.metric == "roc_auc") %>%
  select(mean, mtry:tree_depth) %>%
  pivot_longer(mtry:tree_depth,
               values_to = "value",
               names_to = "parameter"
  ) %>%
  ggplot(aes(value, mean, color = parameter)) +
  geom_point(alpha = 0.8, show.legend = FALSE) +
  facet_wrap(~parameter, scales = "free_x") +
  labs(x = NULL, y = "AUC")
```

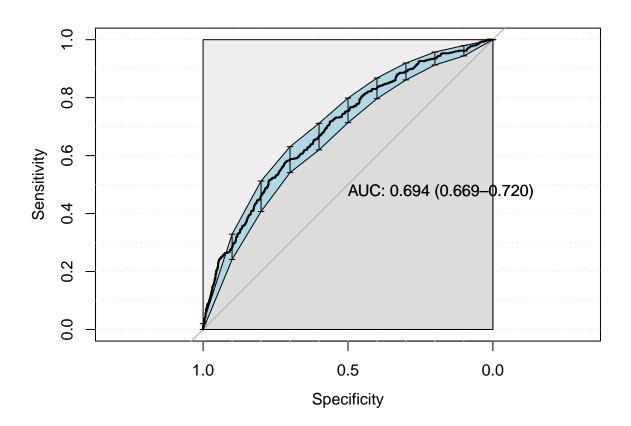


```
final_lightgbm_workflow <-
    lightgbm_workflow %>%
    finalize_workflow(best_lightgbm)

last_lightgbm_fit <-
    final_lightgbm_workflow %>%
    last_fit(df_split)

final_lightgbm_fit <- extract_workflow(last_lightgbm_fit)

lightgbm_auc <- validation(final_lightgbm_fit, df_test)</pre>
```



Confusion Matrix and Statistics

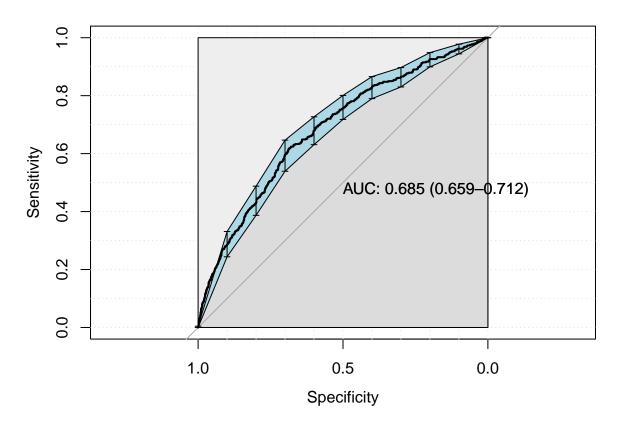
##

```
##
   test_predictions_class
                        0 4282 445
##
##
                        1
                                   2
##
                  Accuracy : 0.9057
##
                    95% CI: (0.897, 0.9139)
##
       No Information Rate: 0.9055
##
       P-Value [Acc > NIR] : 0.4928
##
##
##
                     Kappa : 0.0076
##
##
    Mcnemar's Test P-Value : <2e-16
##
##
               Sensitivity: 0.999767
##
               Specificity: 0.004474
##
            Pos Pred Value : 0.905860
##
            Neg Pred Value: 0.666667
                Prevalence: 0.905497
##
##
            Detection Rate: 0.905285
##
      Detection Prevalence: 0.999366
##
         Balanced Accuracy : 0.502120
##
##
          'Positive' Class : 0
##
lightgbm_parameters <- lightgbm_tune %>%
  show_best("roc_auc", n=1) %>%
  select(trees, mtry, min_n, tree_depth, learn_rate, loss_reduction) %>%
  as.list
saveRDS(
```

```
lightgbm_parameters,
file = sprintf(
   "../EDA/auxiliar/hyperparameters/model_selection/lightgbm_parameters_%s.rds",
   outcome_column
)
)
```

GLM

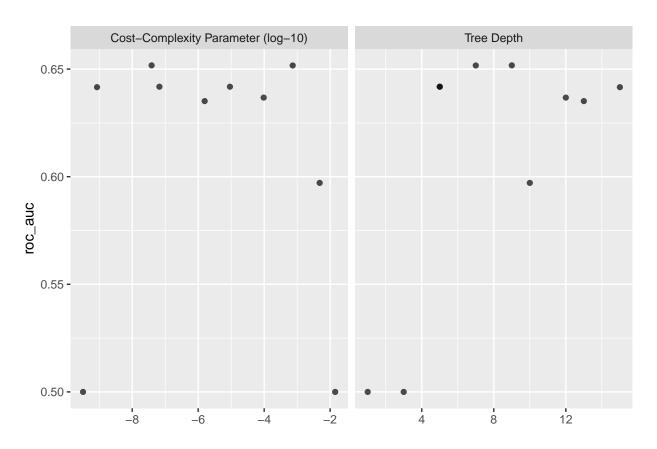
```
glmnet_recipe <-</pre>
  recipe(formula = sprintf("%s ~ .", outcome_column) %% as.formula, data = df_train) %>%
  step_novel(all_nominal_predictors()) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_other(all_nominal_predictors(), threshold = 0.05, other=".merged") %>%
  step_dummy(all_nominal_predictors(), one_hot = TRUE) %>%
  step_zv(all_predictors()) %>%
  step_normalize(all_numeric_predictors())
glmnet_spec <-</pre>
  logistic_reg(penalty = 0) %>%
  set_mode("classification") %>%
  set_engine("glmnet")
glmnet_workflow <-</pre>
  workflow() %>%
  add_recipe(glmnet_recipe) %>%
  add_model(glmnet_spec)
glm_fit <- glmnet_workflow %>%
  fit(df_train)
glm_auc = validation(glm_fit, df_test)
```



```
Confusion Matrix and Statistics
##
##
##
  test_predictions_class
                                   1
##
                        0 4273
                                 440
                            10
##
                        1
                                   7
##
                  Accuracy : 0.9049
##
##
                    95% CI: (0.8961, 0.9131)
       No Information Rate: 0.9055
##
       P-Value [Acc > NIR] : 0.5716
##
##
##
                     Kappa : 0.0234
##
    Mcnemar's Test P-Value : <2e-16
##
##
##
               Sensitivity: 0.99767
##
               Specificity: 0.01566
##
            Pos Pred Value: 0.90664
##
            Neg Pred Value: 0.41176
##
                Prevalence: 0.90550
##
            Detection Rate: 0.90338
##
      Detection Prevalence: 0.99641
##
         Balanced Accuracy : 0.50666
##
##
          'Positive' Class : 0
##
```

Decision Tree

```
tree_recipe <-
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula, data = df_train) %>%
  step_novel(all_nominal_predictors()) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_other(all_nominal_predictors(), threshold = 0.05, other=".merged") %%
  step_dummy(all_nominal_predictors(), one_hot = TRUE) %>%
  step_zv(all_predictors())
tree_spec <-
  decision_tree(cost_complexity = tune(),
                tree_depth = tune()) %>%
  set_mode("classification") %>%
  set_engine("rpart")
tree_grid <- grid_latin_hypercube(cost_complexity(),</pre>
                                   tree_depth(),
                                   size = grid_size)
tree_workflow <-</pre>
 workflow() %>%
 add_recipe(tree_recipe) %>%
  add_model(tree_spec)
tree_tune <-
 tree_workflow %>%
 tune_grid(resamples = df_folds,
            grid = tree_grid)
tree_tune %>%
  collect_metrics()
autoplot(tree_tune, metric = "roc_auc")
```



```
tree_tune %>%
    show_best("roc_auc")

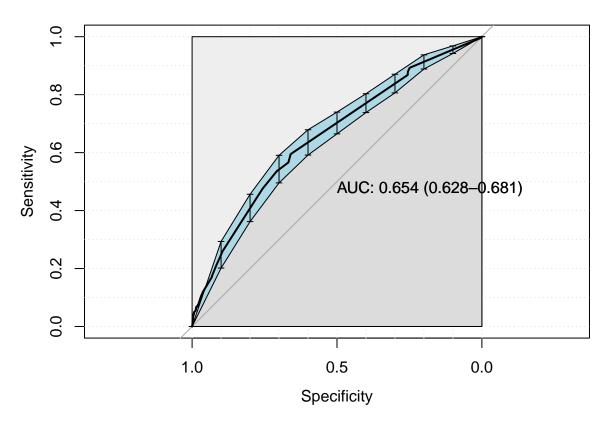
best_tree <- tree_tune %>%
    select_best("roc_auc")

final_tree_workflow <-
    tree_workflow %>%
    finalize_workflow(best_tree)

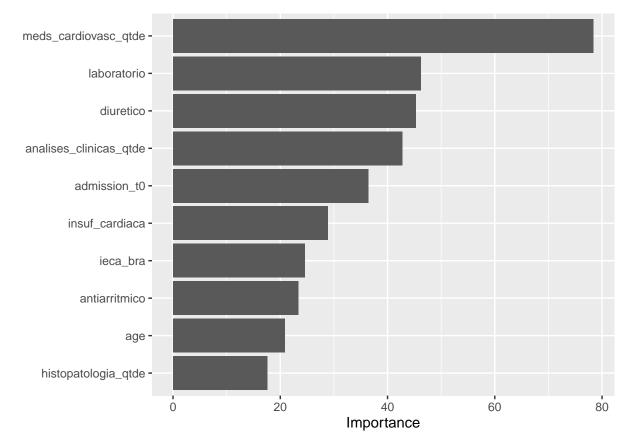
last_tree_fit <-
    final_tree_workflow %>%
    last_fit(df_split)

final_tree_fit <- extract_workflow(last_tree_fit)

tree_auc = validation(final_tree_fit, df_test)</pre>
```

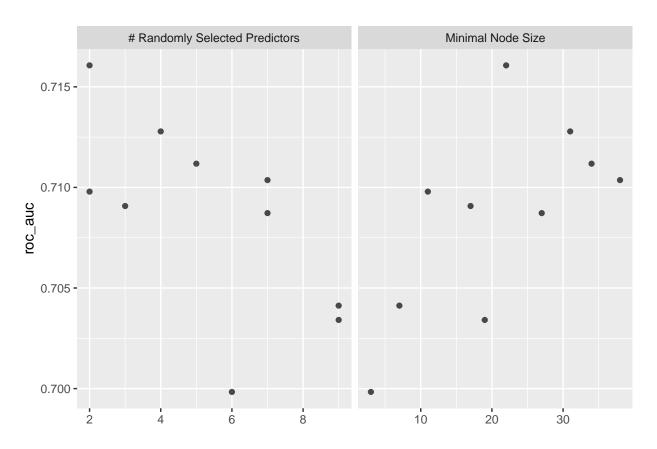


```
if (tree_auc$auc > 0.55){
  final_tree_fit %>%
    extract_fit_parsnip() %>%
    vip()
}
```



Random Forest

```
rf_recipe <-
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula, data = df_train) %>%
  step_novel(all_nominal_predictors()) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_other(all_nominal_predictors(), threshold = 0.05, other=".merged") %%
  step_dummy(all_nominal_predictors(), one_hot = TRUE) %>%
  step_zv(all_predictors()) %>%
  step_impute_mean(all_numeric_predictors())
rf_spec <-
  rand_forest(mtry = tune(),
              trees = 100,
              min_n = tune()) %>%
  set_mode("classification") %>%
  set_engine("ranger")
rf_grid <- grid_latin_hypercube(mtry(range = c(1, 10)),</pre>
                                min_n(),
                                 size = grid_size)
rf_workflow <-
  workflow() %>%
  add_recipe(rf_recipe) %>%
  add_model(rf_spec)
rf_tune <-
  rf_workflow %>%
  tune_grid(resamples = df_folds,
            grid = rf_grid)
rf_tune %>%
  collect_metrics()
autoplot(rf_tune, metric = "roc_auc")
```



```
rf_tune %>%
    show_best("roc_auc")

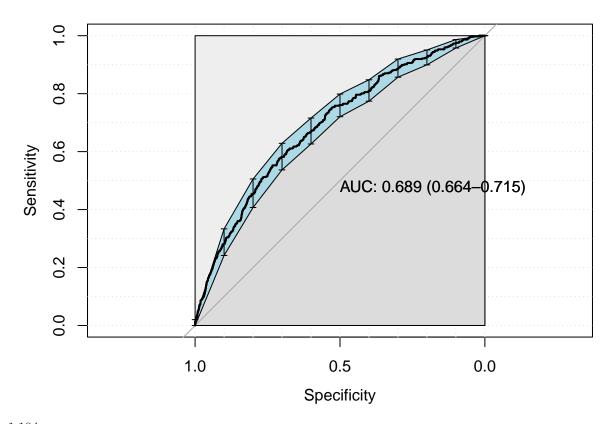
best_rf <- rf_tune %>%
    select_best("roc_auc")

final_rf_workflow <-
    rf_workflow %>%
    finalize_workflow(best_rf)

last_rf_fit <-
    final_rf_workflow %>%
    last_fit(df_split)

final_rf_fit <- extract_workflow(last_rf_fit)

rf_auc = validation(final_rf_fit, df_test)</pre>
```



1.104

KNN

```
# knn_recipe <-
    recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula, data = df_train) %>%
    step_novel(all_nominal_predictors()) %>%
   step_unknown(all_nominal_predictors()) %>%
    step_other(all_nominal_predictors(), threshold = 0.05, other=".merged") %>%
   step_dummy(all_nominal_predictors(), one_hot = TRUE) %>%
    step_zv(all_predictors()) %>%
    step_impute_mean(all_numeric_predictors())
# knn_spec <-
#
   nearest_neighbor(neighbors = tune(),
#
                     weight_func = tune(),
#
                     dist_power = tune()) %>%
   set_mode("classification") %>%
#
#
    set_engine("kknn")
# knn_grid <- grid_latin_hypercube(neighbors(),</pre>
                                   weight_func(),
#
                                    dist_power(),
                                    size = 5)
 knn_workflow <-
   workflow() %>%
    add_recipe(knn_recipe) %>%
    add_model(knn_spec)
# knn_tune <-
   knn_workflow %>%
    tune_grid(resamples = df_folds,
```

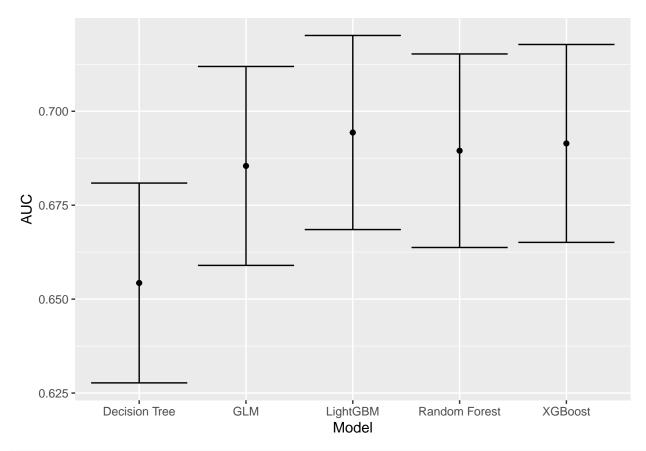
```
qrid = knn_qrid
#
# knn tune %>%
#
    collect_metrics()
# autoplot(knn_tune, metric = "roc_auc")
# knn_tune %>%
#
    show_best("roc_auc")
# best_knn <- knn_tune %>%
    select_best("roc_auc")
# final_knn_workflow <-</pre>
    knn_workflow %>%
    finalize_workflow(best_knn)
# last_knn_fit <-</pre>
#
   final_knn_workflow %>%
   last_fit(df_split)
# final_knn_fit <- extract_workflow(last_knn_fit)</pre>
# knn_auc = validation(final_knn_fit, df_test)
```

SVM

```
# svm_recipe <-
              recipe(formula = sprintf("%s ~ .", outcome\_column) \%>\% as.formula, data = df\_train) %>% as.formul
              step_novel(all_nominal_predictors()) %>%
              step_unknown(all_nominal_predictors()) %>%
              step_other(all_nominal_predictors(), threshold = 0.05, other=".merged") %>%
              step_dummy(all_nominal_predictors(), one_hot = TRUE) %>%
              step_zv(all_predictors()) %>%
#
              step_impute_mean(all_numeric_predictors())
# svm_spec <-
#
              svm_rbf(cost = tune(), rbf_sigma = tune()) %>%
              set_mode("classification") %>%
              set_engine("kernlab")
# svm_grid <- grid_latin_hypercube(cost(),</pre>
#
                                                                                                                                 rbf_sigma(),
#
                                                                                                                                 size = grid\_size)
#
# svm_workflow <-</pre>
              workflow() %>%
              add_recipe(svm_recipe) %>%
              add_model(svm_spec)
# svm_tune <-
#
              sum_workflow %>%
              tune_grid(resamples = df_folds,
#
                                                    grid = 5)
# svm_tune %>%
              collect_metrics()
```

```
# autoplot(svm_tune, metric = "roc_auc")
#
# svm_tune %>%
# show_best("roc_auc")
#
# best_svm <- svm_tune %>%
# select_best("roc_auc")
#
# final_svm_workflow <-
# svm_workflow %>%
# finalize_workflow(best_svm)
#
# last_svm_fit <-
# final_svm_workflow %>%
# last_stm_fit <-
# final_svm_workflow %>%
# last_fit(af_split)
#
# final_svm_fit <- extract_workflow(last_svm_fit)
#
# svm_auc = validation(final_svm_fit, df_test)</pre>
```

Models Comparison



saveRDS(df_auc, sprintf("../EDA/auxiliar/performance/%s_auc_result.RData", outcome_column))