



# BIX TECNOLOGIA

## DESAFIO CIENTISTA DE DADOS

RODRIGO VIEIRA



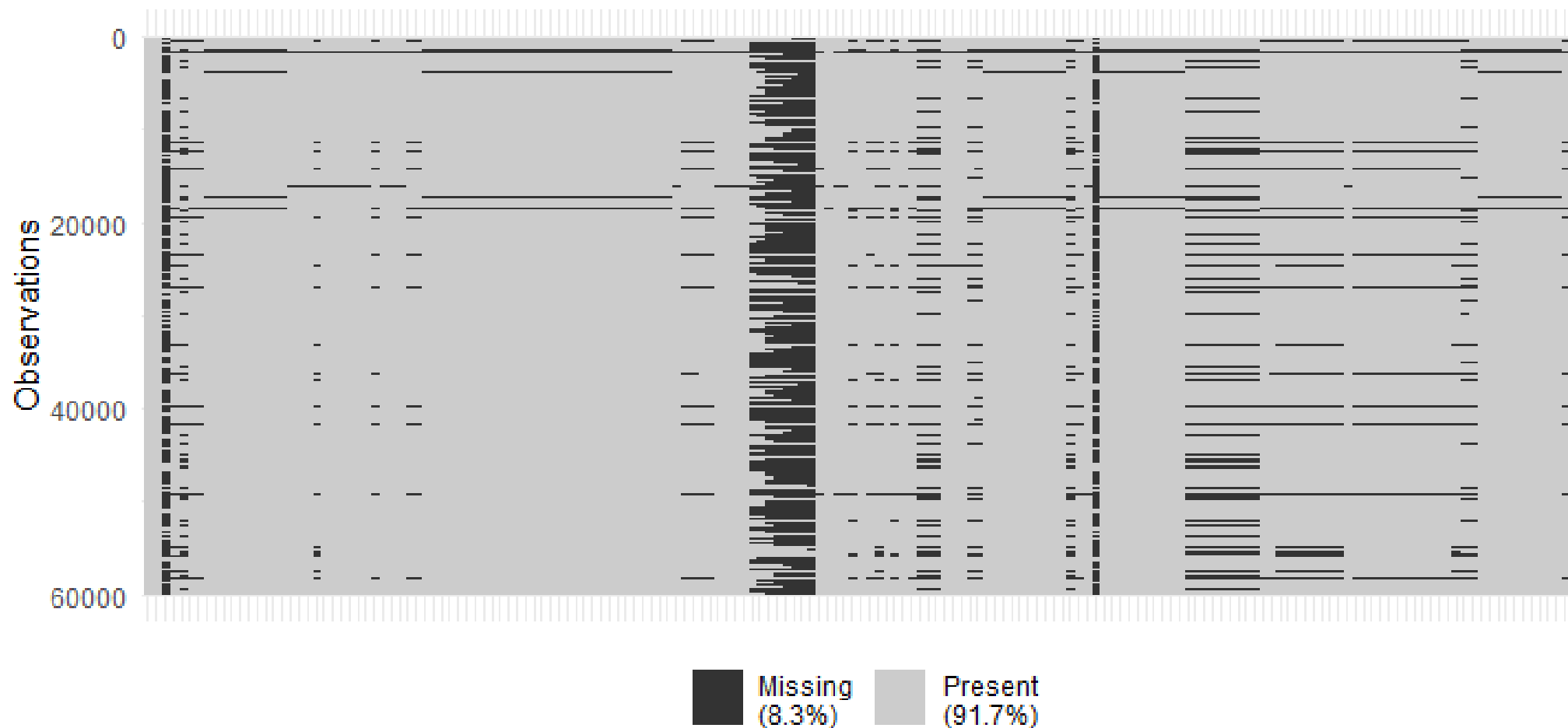
# TOPICOS DA APRESENTAÇÃO

- PRINCIPAIS AÇÕES DE LIMPEZA E TRANSFORMAÇÃO
- ANÁLISE EXPLORATÓRIA
- ALGORITMOS APLICADOS
- RESULTADOS ALCANÇADOS
- SUGESTÕES E RECOMENDAÇÕES



# AÇÕES DE LIMPEZA E TRANSFORMAÇÃO

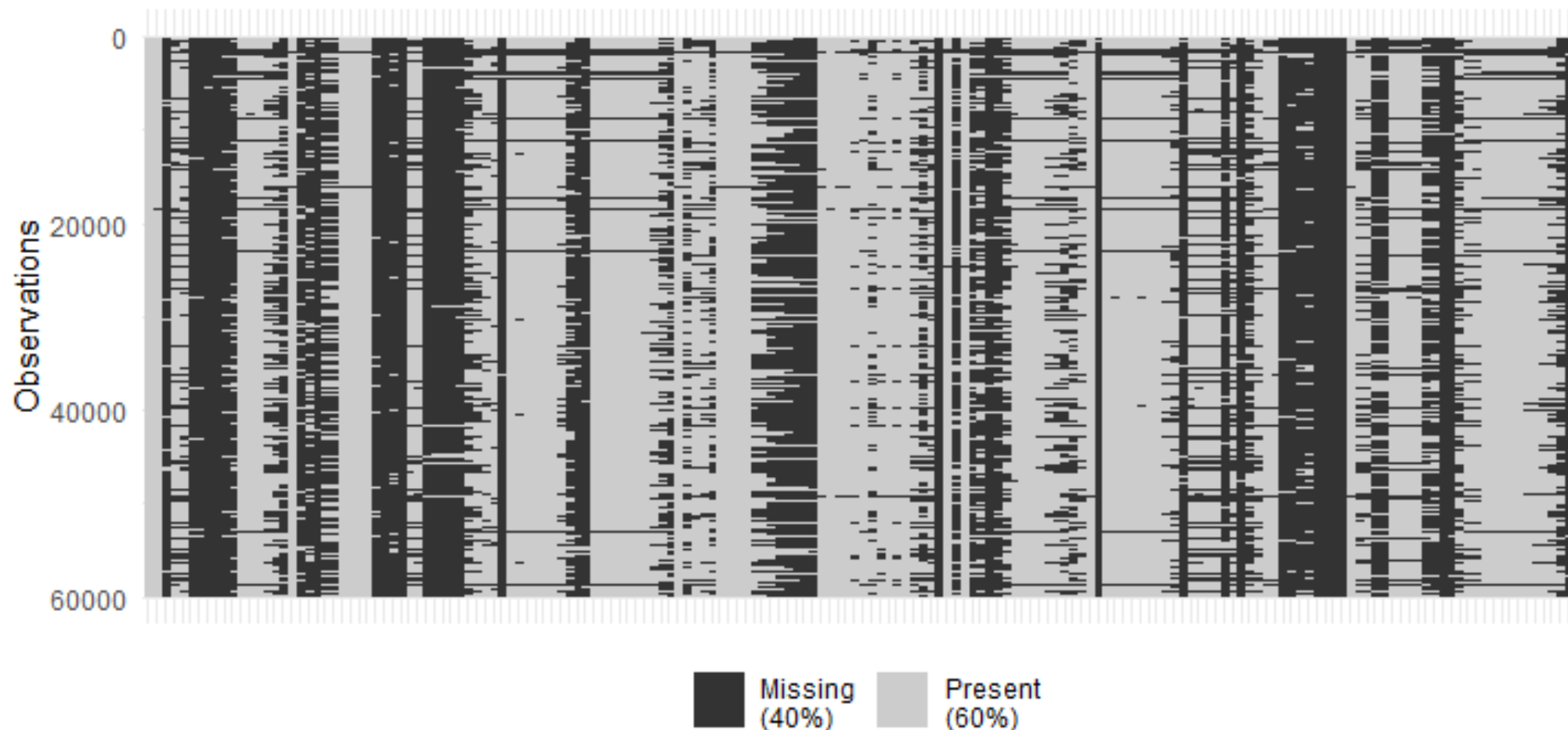
- ESTAVA ASSIM.





# AÇÕES DE LIMPEZA E TRANSFORMAÇÃO

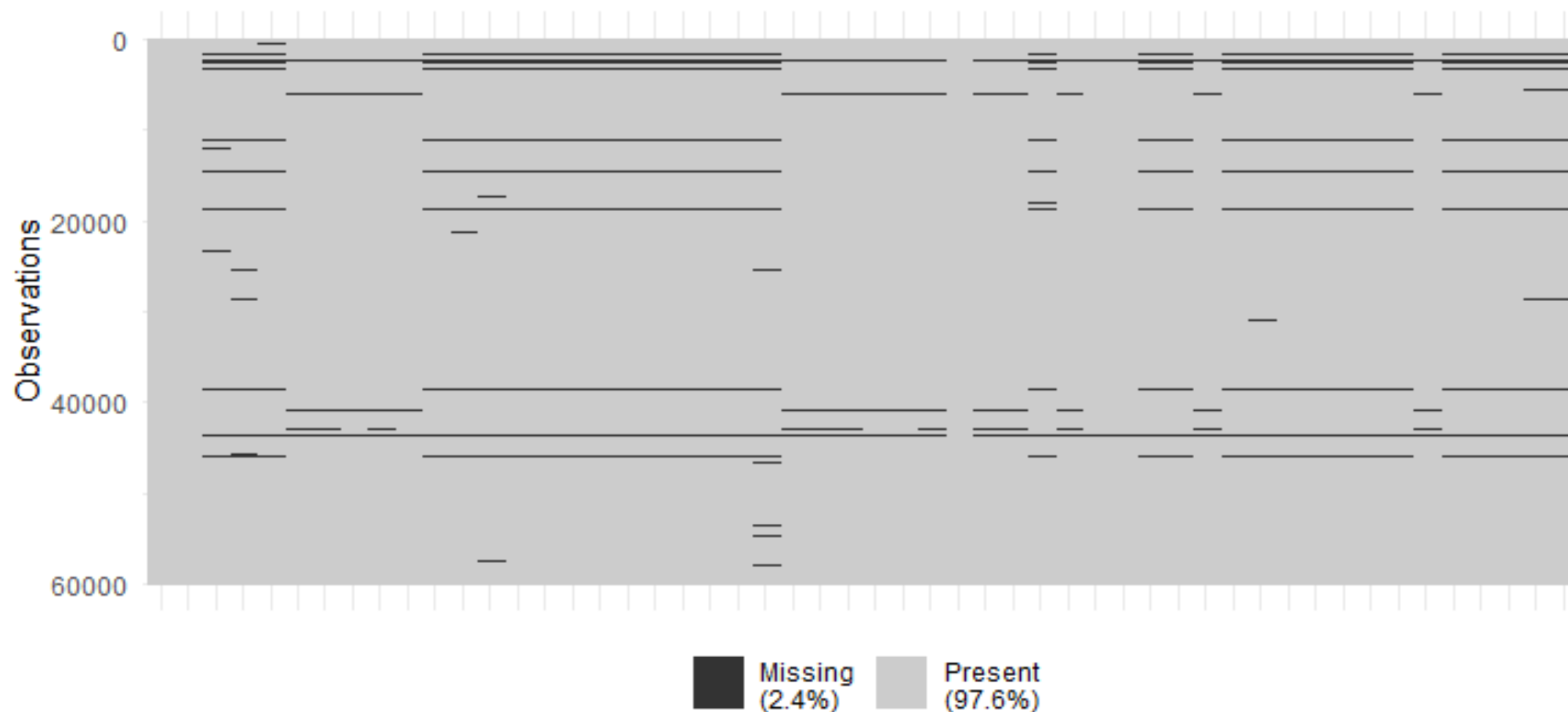
- PRECISOU FICAR ASSIM.





# AÇÕES DE LIMPEZA E TRANSFORMAÇÃO

- PARA ENTÃO FICAR ASSIM.





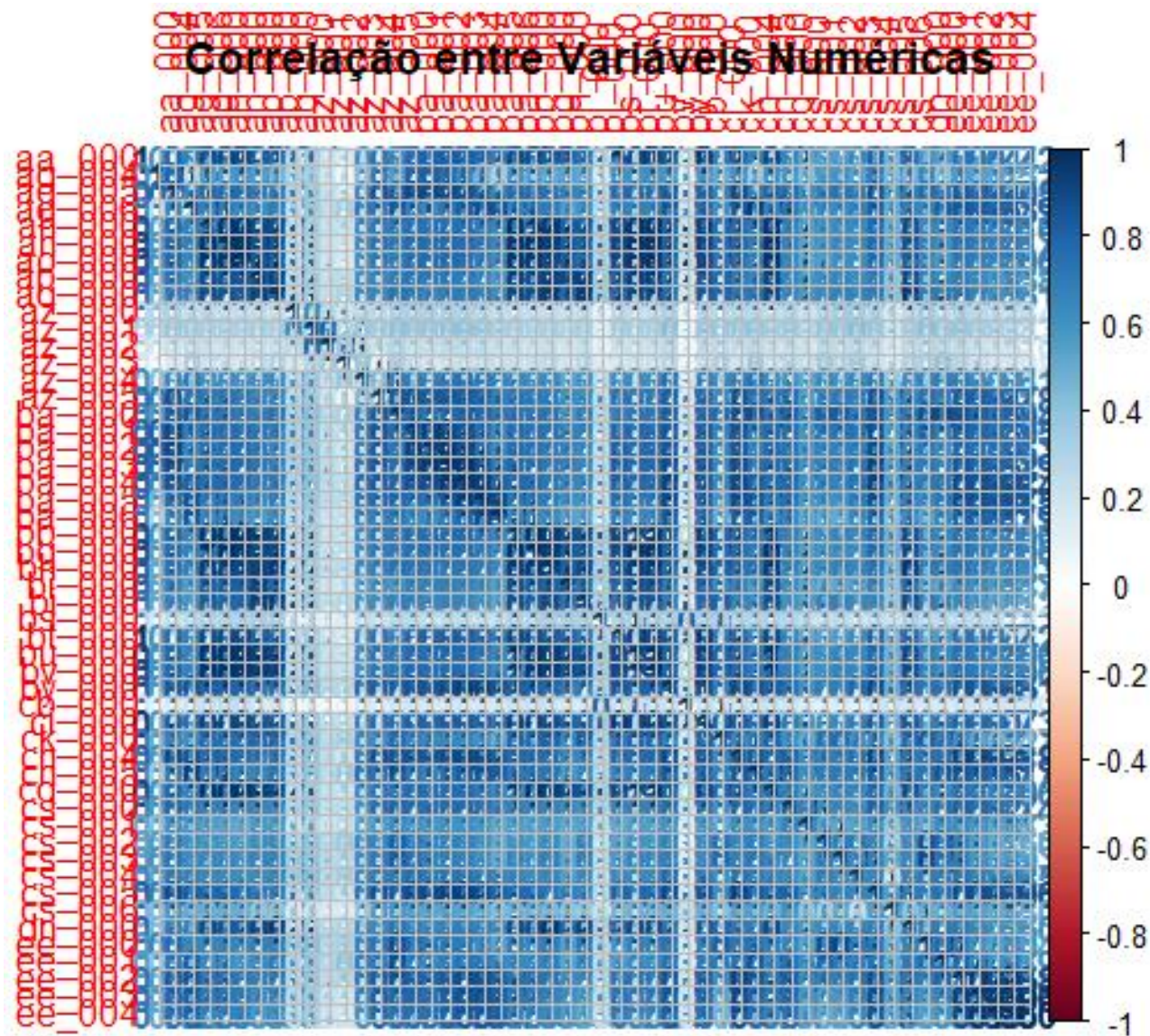
# AÇÕES DE LIMPEZA E TRANSFORMAÇÃO

- SUBSTITUIR VALORES “0” POR “NA”.
- CONTABILIZAR O TOTAL DE VALORES “NA” POR COLUNA.
- EXCLUIR COLUNAS COM MAIS DE 5% DE VALORES “NA”.
- DATA SETE RESULTANTE COM 52 COLUNAS, 2,4% DE VALORES “NA”.
- IMPUTAR A MEDIANA PARA OS DEMAIS DADOS “NA” RESTANTES.
- PADRONIZAR OS VALORES COM A FUNÇÃO “SCALE”.



# ANÁLISE EXPLORATÓRIA

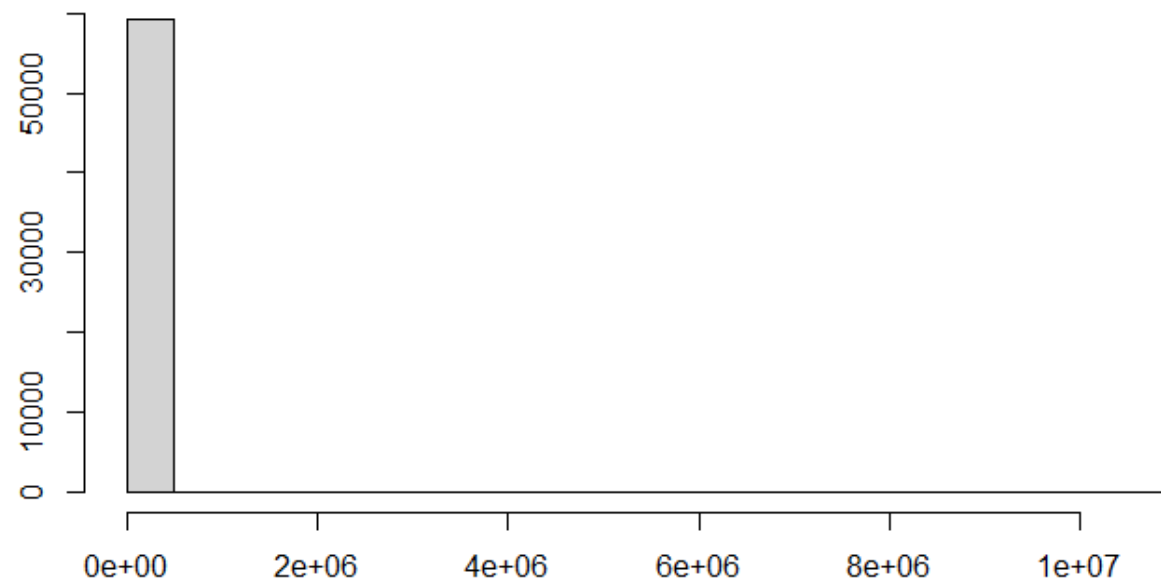
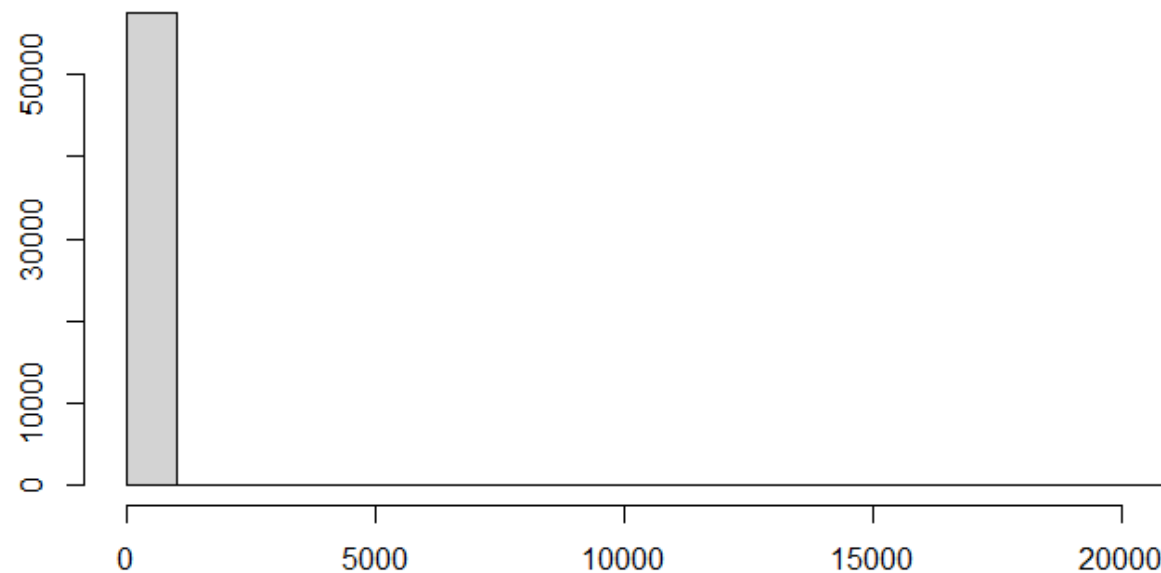
- As variáveis não apresentam correlação negativa;
- Os grupos az, bs e co apresentam FRACA correlação positiva;
- As demais apresentam FORTE correlação positiva;





# ANÁLISE EXPLORATÓRIA

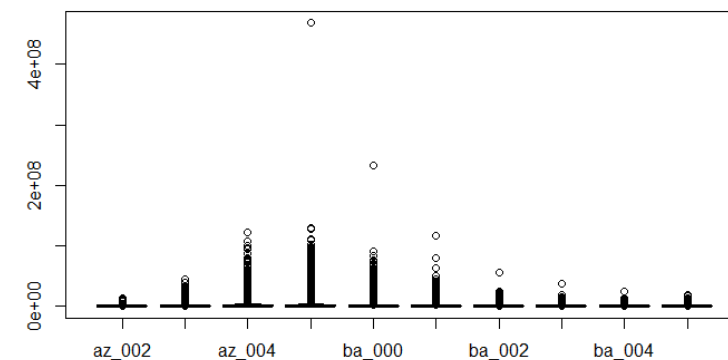
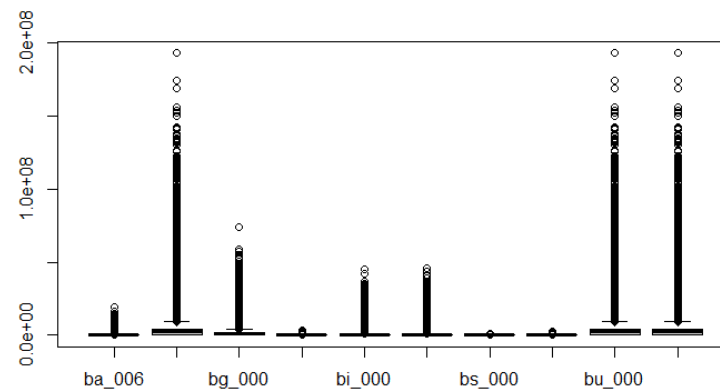
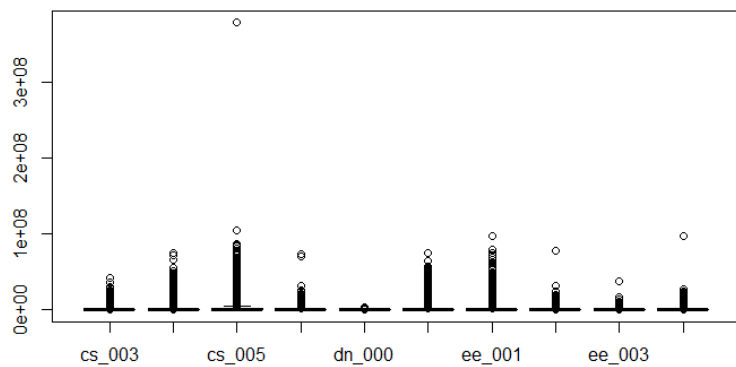
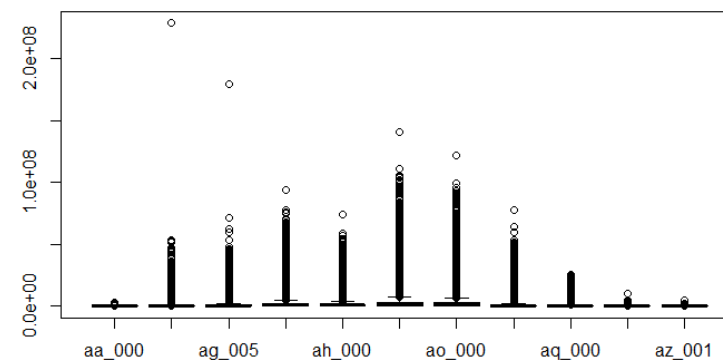
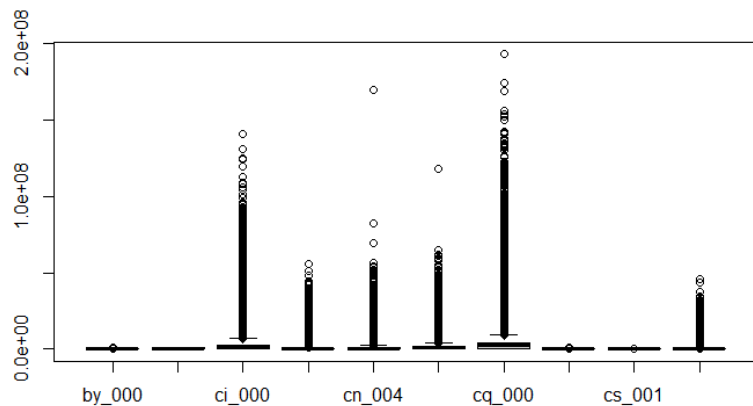
- As variáveis não apresentam distribuição normal;
- Muitas parecem sofrer interferência de valores discrepantes;







# ANÁLISE EXPLORATÓRIA





# ALGORITMOS APLICADOS

- REGRESSÃO LOGÍSTICA
- ARVORE DE DECISÃO
- (TENTATIVA) RANDOM FOREST
- USANDO O MS Azure:
  - Two-Class Decision Forest
  - Two-Class Support Vector Machine
  - Two-Class Logistic Regression (com PCA)
  - Two-Class Decision Jungle (com PCA)



# RESULTADOS ALCANÇADOS

- REGRESSÃO LOGÍSTICA
- NÍVEL DE ACURÁCIA – 97,2%
- MELHOR PARA PREVER O NÃO APRESENTA DEFEITO.

	não	sim
não	97,5% 15229	2,5% 396
sim	9,9% 37	90,1% 338

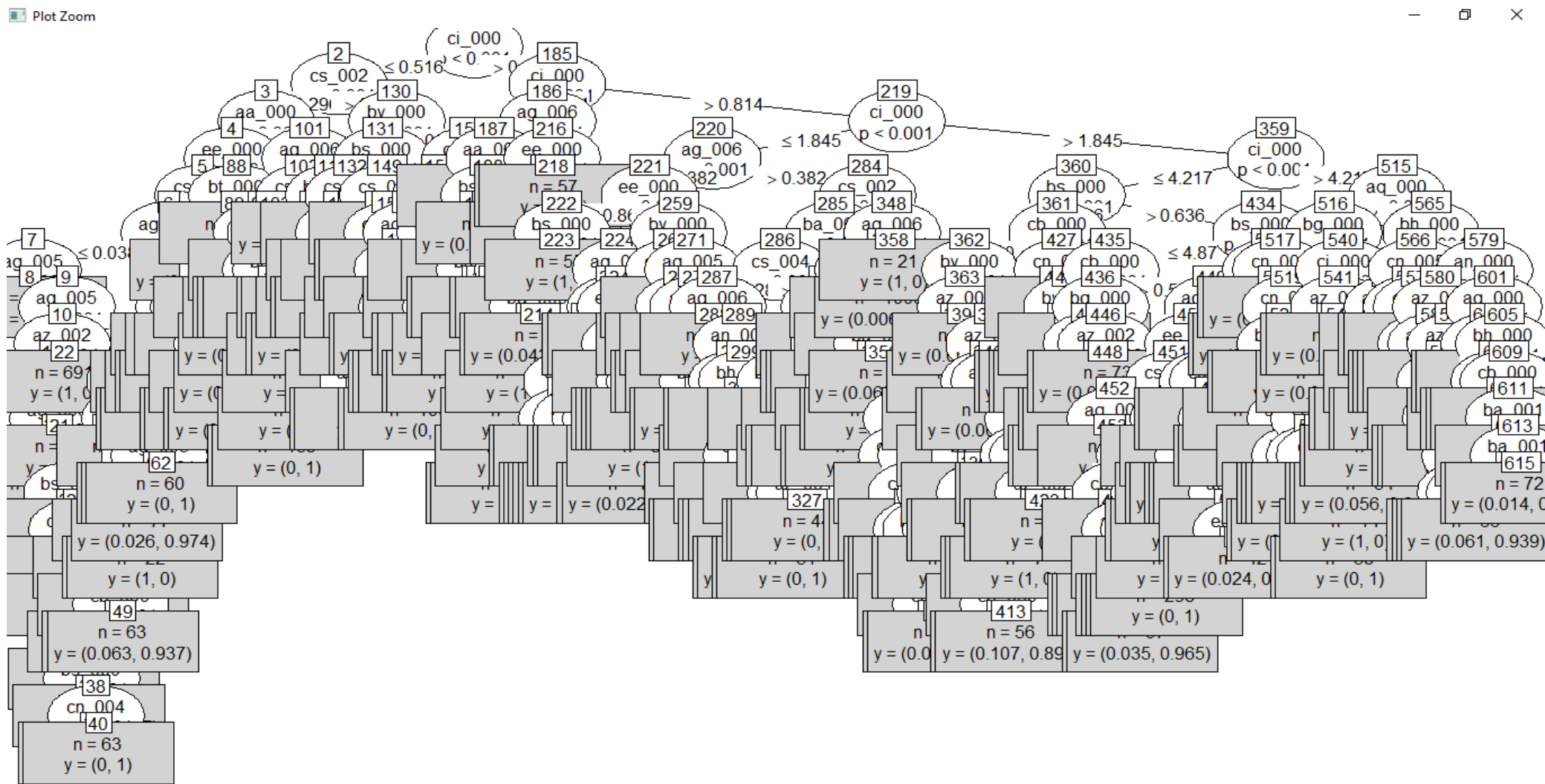


# RESULTADOS ALCANÇADOS

- ÁRVORE DE DECISÃO
- NÍVEL DE ACURÁCIA – 97,4%
- MUITO RUIM PARA PREVER O NÃO APRESENTA DEFEITO.

	não	sim
não	99,4% 15298	0,6% 99
sim	54,2% 327	45,8% 276

# RESULTADOS ALCANÇADOS





# RESULTADOS ALCANÇADOS

- TENTATIVA RANDOM FOREST

The screenshot displays the RStudio interface with two main windows. The left window shows the R console and environment pane. The right window shows the R script editor.

**Left Window (Console and Environment):**

```
R 4.1.0 · C:/JOBS_R/Bix_Tecnologia/
ci_000 +
+ cn_004 + cn_005 + cq_000 + cs_002 + cs_003 + cs_004 +
+ cs_005 +
+ ee_000 + ee_002, dfbix_treino)
> plot(tree, type='simple')
> pred_tree <- predict(tree, dfbix_teste)
> print("Confusion Matrix Para Decision Tree")
[1] "Confusion Matrix Para Decision Tree"
> table(Predicted = pred_tree, Actual = dfbix_teste$class)
      Actual
Predicted 0      1
neg 15298    99
pos   327   276
> B1 <- predict(tree, dfbix_teste)
> tab1 <- table(Predicted = B1, Actual = dfbix_teste$class)
> tab2 <- table(Predicted = pred_tree, Actual = dfbix_teste$class)
> print(paste("Decision Tree Accuracy", sum(diag(tab2))/sum(tab2)))
[1] "Decision Tree Accuracy 0.973375"
> rfModel <- randomForest(class ~ ., data = dfbix_treino)
Error: cannot allocate vector of size 900.3 Mb
>
```

**Environment Pane:**

Name	Type	Length	Size	Value
B1	factor	16000	63 KB	Factor w/ 2 levels "neg",...
dfbix_teste	data.frame	52	6.4 MB	16000 obs. of 52 variab...
dfbix_treino	tbl_df	52	46.4 MB	118000 obs. of 52 varia...
fitted.results	numeric	16000	1.1 MB	Large numeric (16000 elem...
Gctorture	logical	1	56 B	FALSE
LogModel_v1	glm	30	0 B	List of 30
misclasificEr...	numeric	1	56 B	0.0278125
pred_tree	factor	16000	63 KB	Factor w/ 2 levels "neg",...
tab1	table	4	1.3 KB	'table' int [1:2, 1:2] 15...
tab2	table	4	1.3 KB	'table' int [1:2, 1:2] 15...
tree	BinaryTree	1	919.4 MB	Large BinaryTree ( 964 ...

**Right Window (Script Editor):**

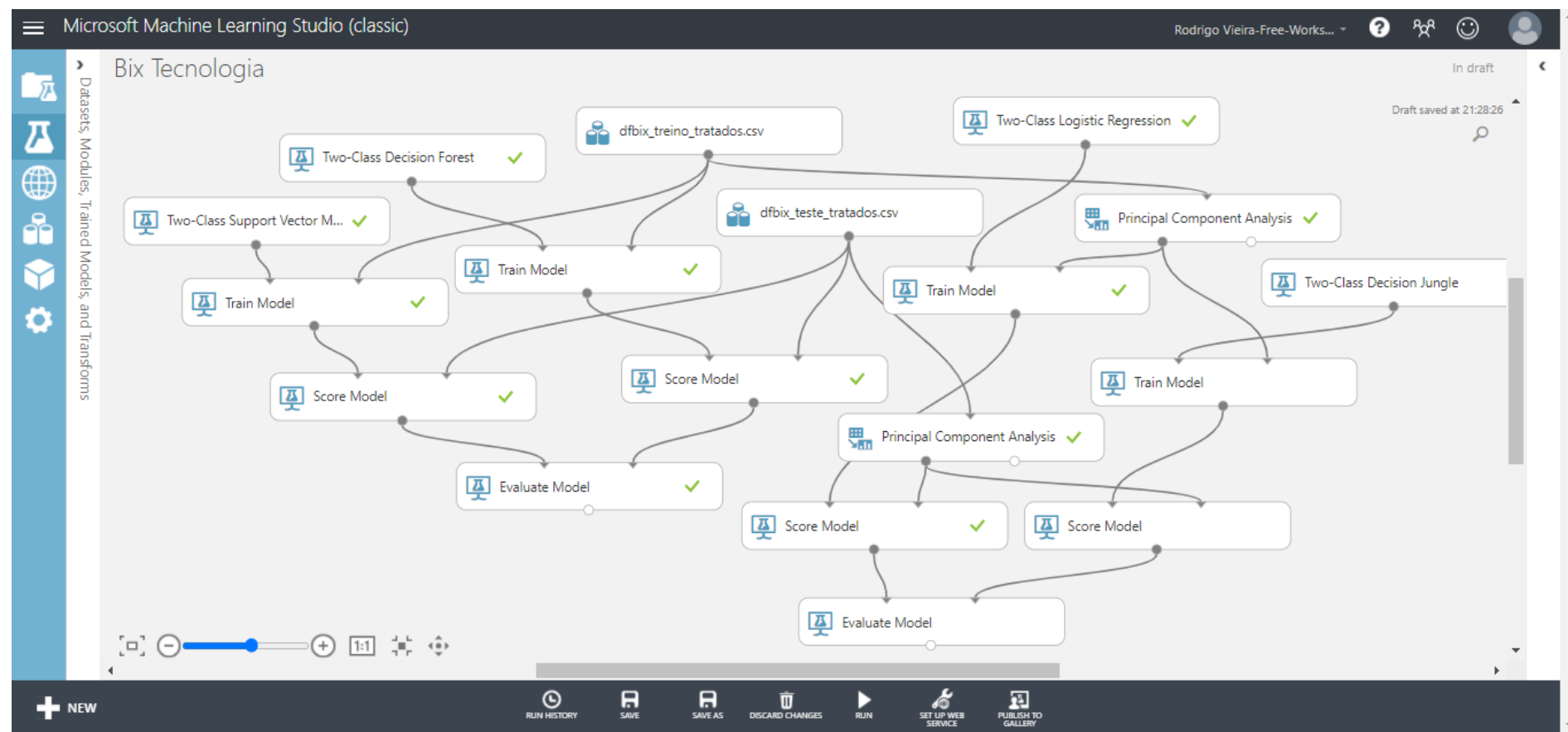
```
Desafio_BIX-v2.R*
Source on Save
Run Source
Project: (None)

420 cn_004 + cn_005 + cq_000 + cs_002 + cs_003 + cs_004 + cs_005 +
421 ee_000 + ee_002, dfbix_treino)
422 plot(tree, type='simple')
423
424 # A variável "ci_000" aparece no topo da árvore para prever o "defeito",
425 # no sistema de freio da frota.
426 # Matriz de Confusão da Árvore de Decisão.
427 pred_tree <- predict(tree, dfbix_teste)
428 print("Confusion Matrix Para Decision Tree")
429 table(Predicted = pred_tree, Actual = dfbix_teste$class)
430
431 # Precisão da árvore de decisão
432 # Esse modelo apresentou acurácia ligeiramente maior (97,3%).
433 # mas foi muito ruim para prever o sim, apenas 45,8%.
434 B1 <- predict(tree, dfbix_teste)
435 tab1 <- table(Predicted = B1, Actual = dfbix_teste$class)
436 tab2 <- table(Predicted = pred_tree, Actual = dfbix_teste$class)
437 print(paste("Decision Tree Accuracy", sum(diag(tab2))/sum(tab2)))
438
439 ##### Random Forest #####
440 # minha máquina não de conta... kkkk
441 # Apresentou erro: "cannot allocate vector of size 900.3 Mb"
442 # Deixo o código, na esperança de funcionar nas poderosas máquinas da Bix!!
443 rfModel <- randomForest(class ~ ., data = dfbix_treino)
444 print(rfModel)
445 plot(rfModel)
446
447 # Prevendo valores com dados de teste
448 pred_rf <- predict(rfModel, testing)
449
450 # Confusion Matrix
451 print("Confusion Matrix Para Random Forest"); table(dfbix_treino$class,
452 pred_rf)
453
454 # Variáveis mais importantes
455 varImpPlot(rfModel, sort=T, n.var = 8, main = 'Top 8 Feature Importance')
456
```



# RESULTADOS ALCANÇADOS

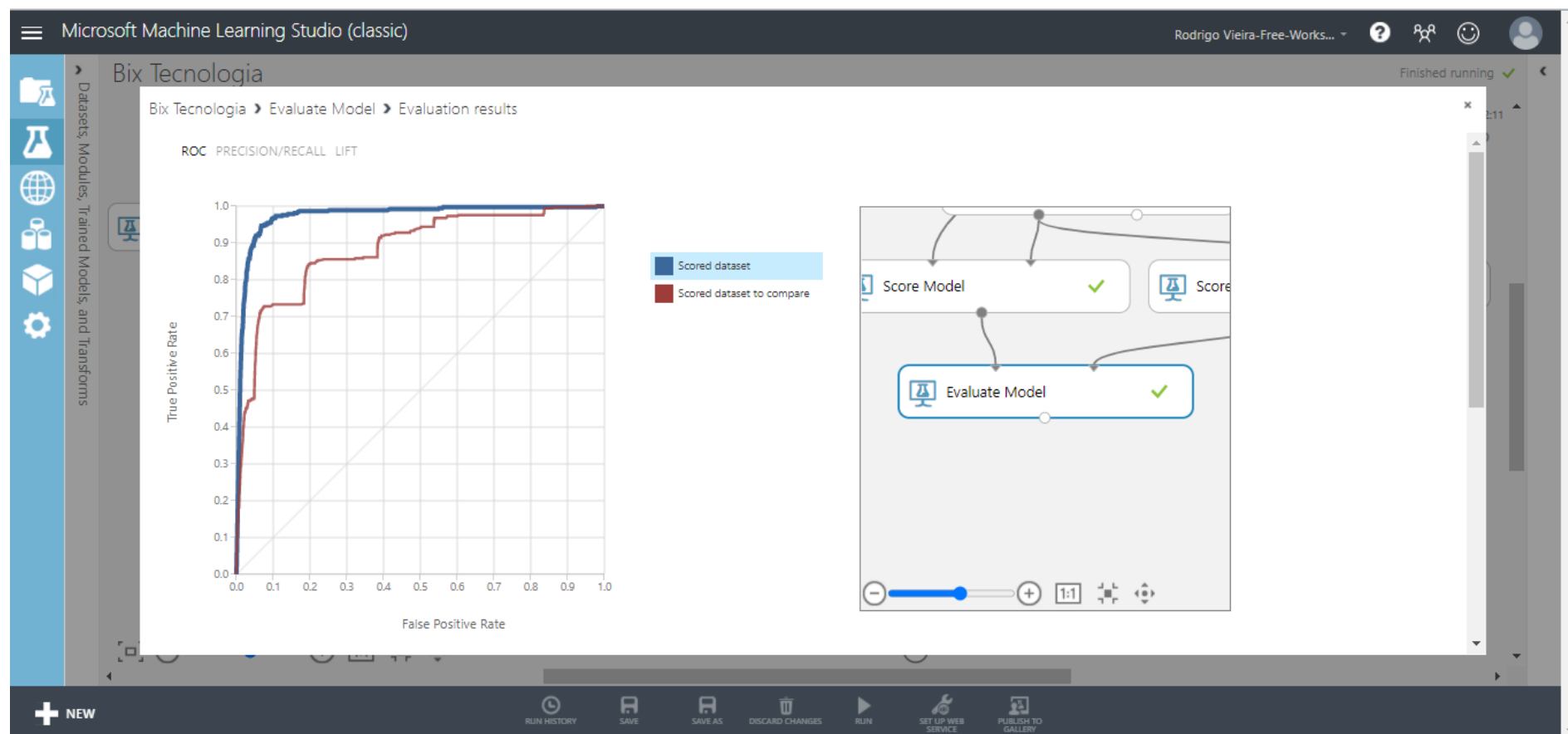
- USANDO O MSAZURE





# RESULTADOS ALCANÇADOS

- USANDO O MSAZURE

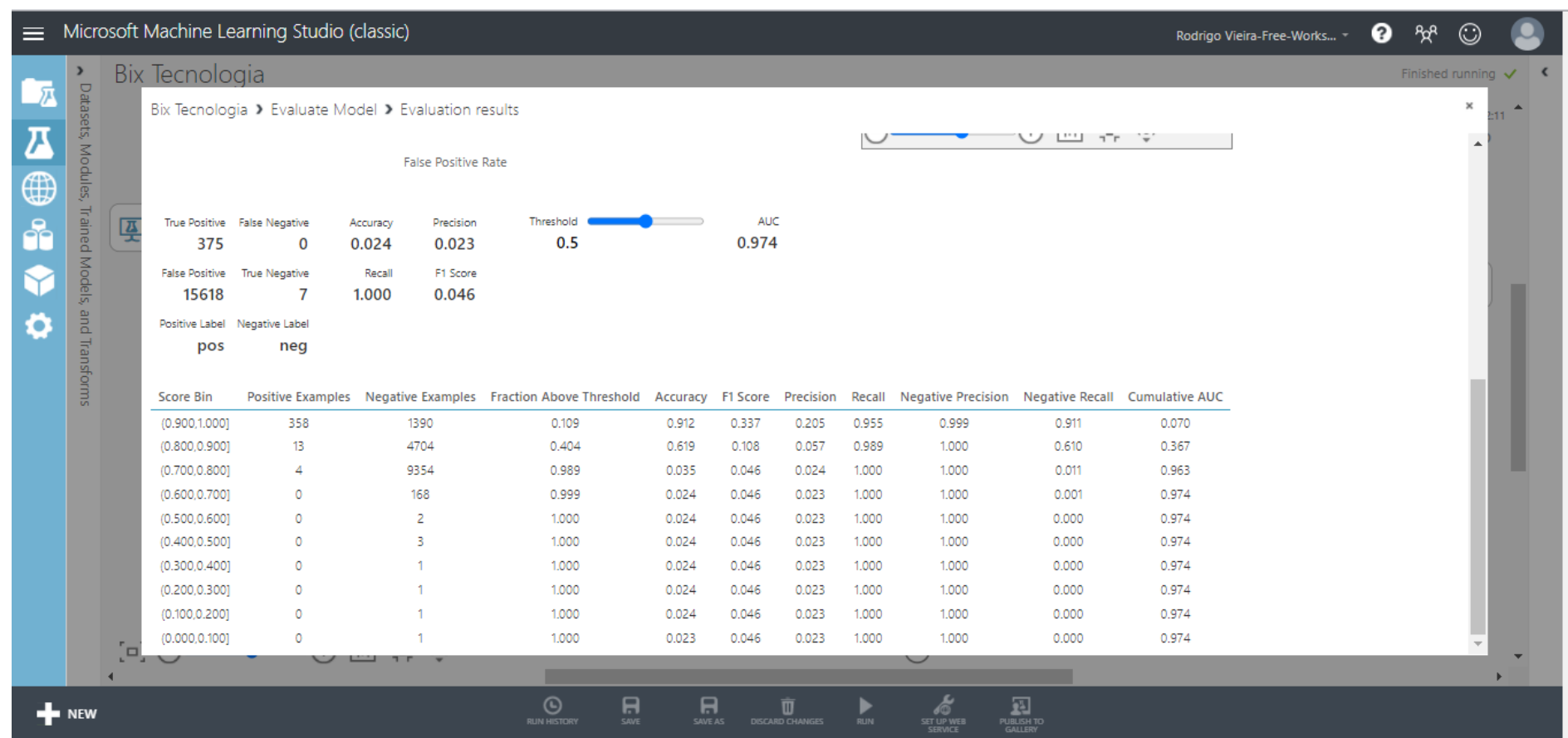






# RESULTADOS ALCANÇADOS

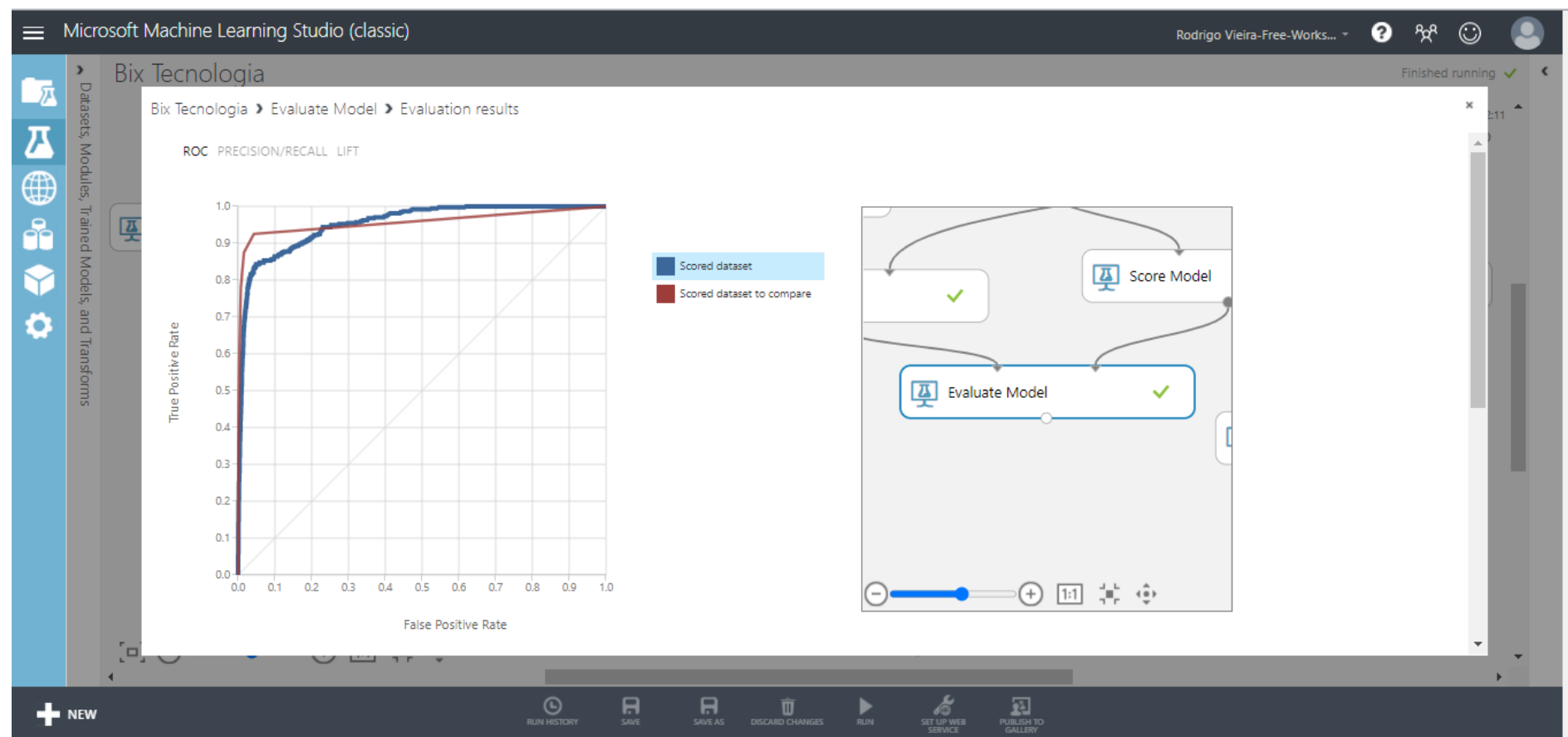
- USANDO O MSAZURE





# RESULTADOS ALCANÇADOS

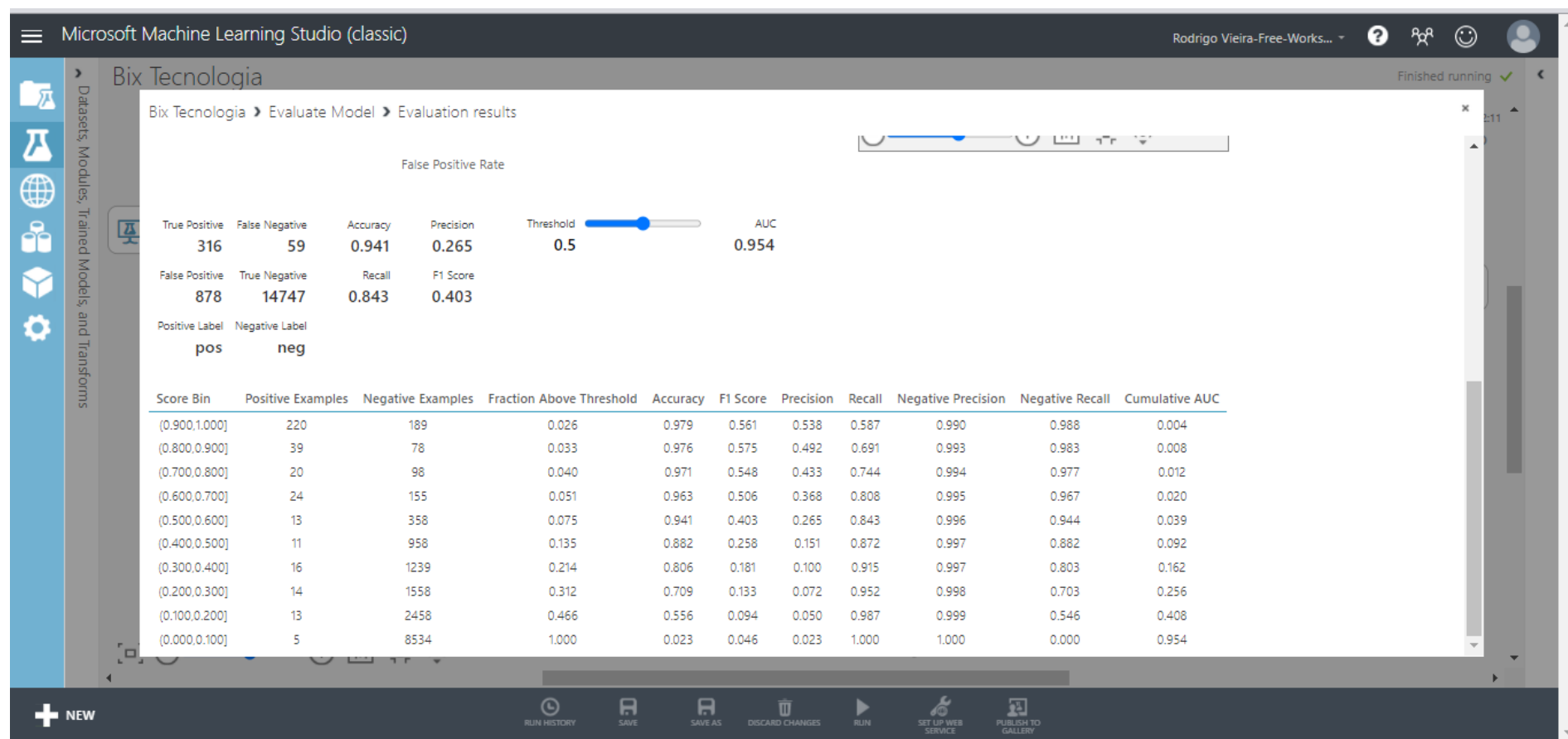
- USANDO O MSAZURE





# RESULTADOS ALCANÇADOS

- USANDO O MSAZURE





# SUGESTÕES E RECOMENDAÇÕES

- REVER PROCESSO DE COLETA DE DADOS PARA REDUZIR PERDAS DE REGISTROS.