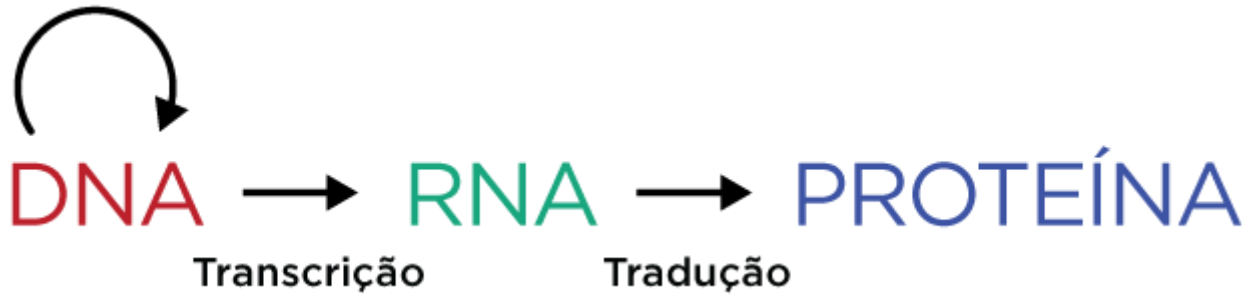
Projeto III – Introdução à modelagem de Big Data

Classificação de essencialidade gênica em organismos modelo

Iara Souza

Introdução

- Dogma central da biologia



Genes → Função

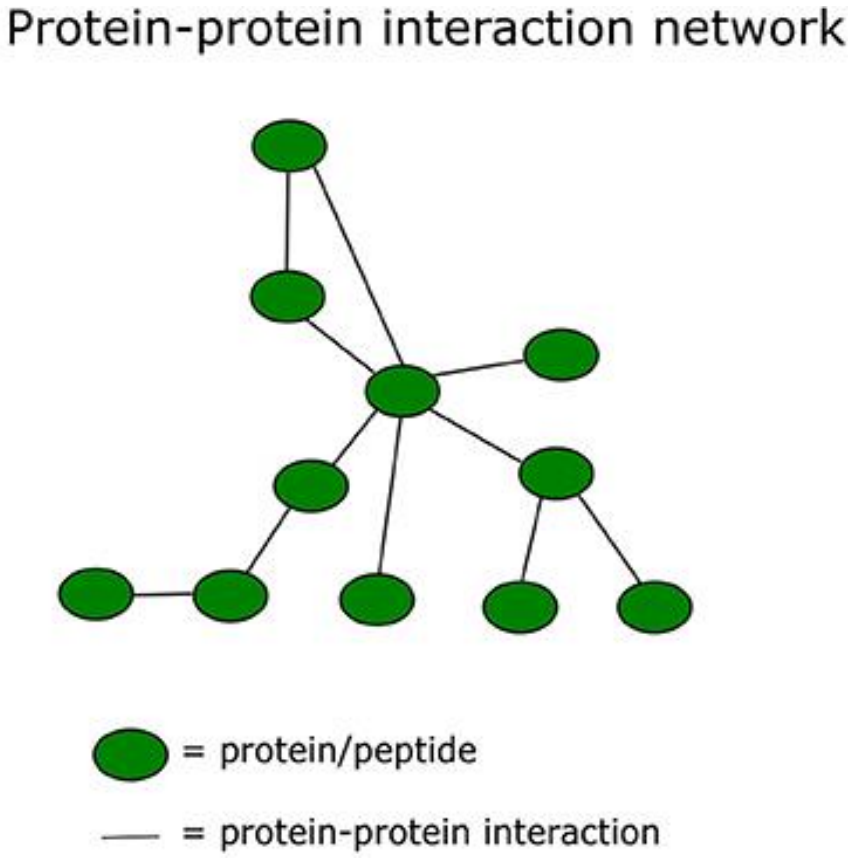
Genes essenciais

Genes essenciais —Alteração deletéria→ Morte do organismo

Genes não-essenciais —Alteração deletéria→ Viabilidade do organismo

Genes essenciais

- Genes essenciais são, em geral:
  - Mais antigos;
  - Interagem mais com outros genes;
  - Participam de funções cruciais na célula



Objetivo

- Classificar os genes de dois organismos modelo (camundongo e levedura) em essenciais ou não-essenciais com base em dados evolutivos e de propriedades de redes biológicas.

Coleta de dados

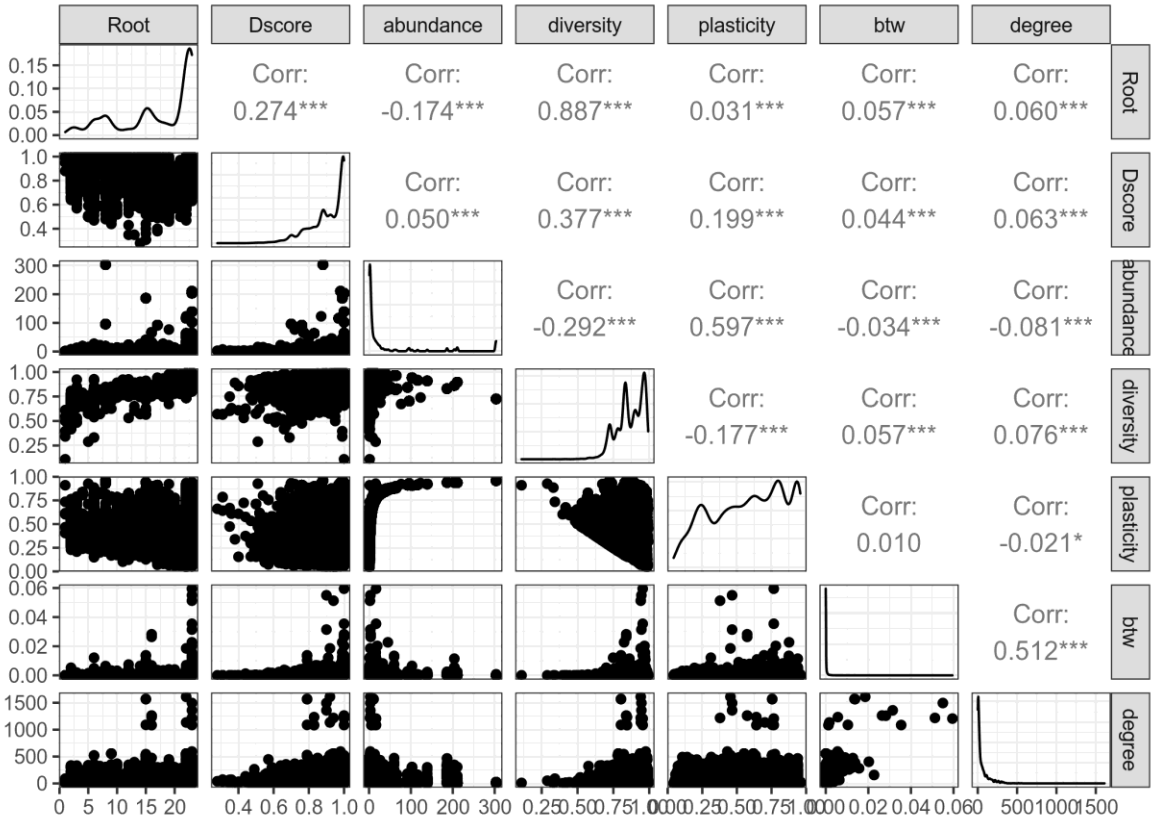
- A obtenção dos dados foi feita em múltiplas etapas:
  - 1. Anotação funcional dos genes
  - 2. Inferência das raízes evolutivas
  - 3. Construção das redes de interação proteína-proteína para cada organismo

Variáveis do conjunto de dados

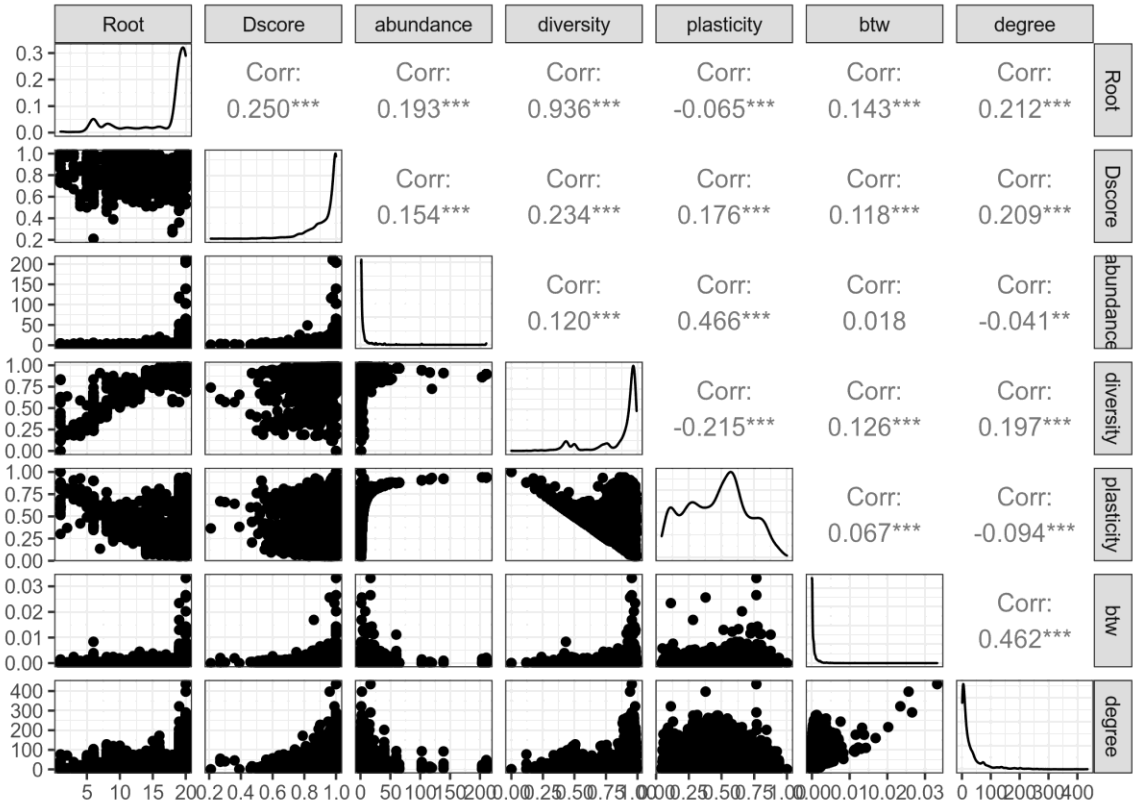
- `ensembl_peptide_id`: identificador da proteína (gene);
- `cog_id`: identificador do grupo de ortólogos ao qual o gene pertence;
- `Root`: raiz evolutiva inferida. Quanto maior, mais antigo;
- `Dscore`: escore de consistência para a inferência da raiz;
- `Pvalue` e `AdjPvalue`: p-valor e p-valor ajustado para a inferência da raiz evolutiva;
- `abundance`: medida de abundância dos grupos de ortólogos;
- `diversity`: medida de diversidade dos grupos de ortólogos;
- `plasticity`: índice de plasticidade evolutiva;
- `ancestry`: ancestralidade do gene. Transformação da variável `Root` (escala de 0 - mais novo - a 1 - mais antigo);
- `btw`: centralidade do nó na rede;
- `degree`: grau do nó na rede.

Análise exploratória

Camundongo

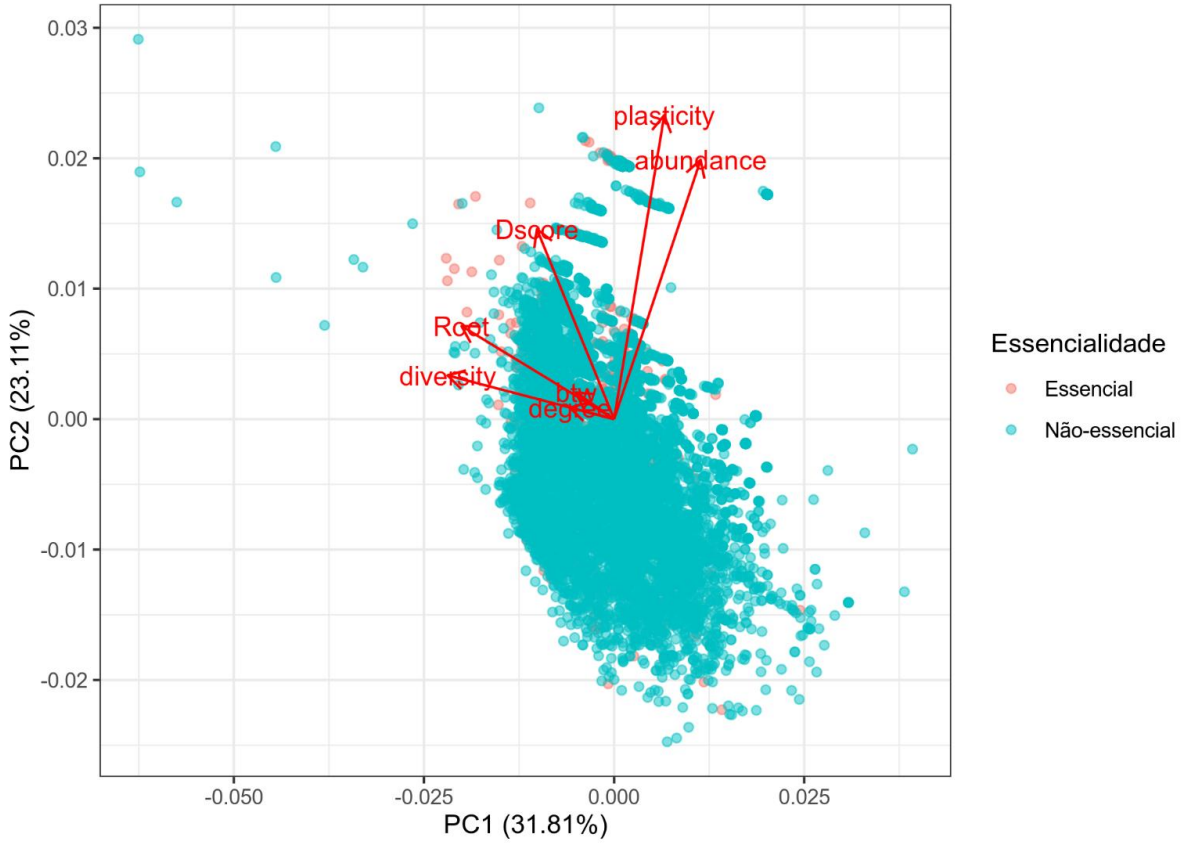


Levedura

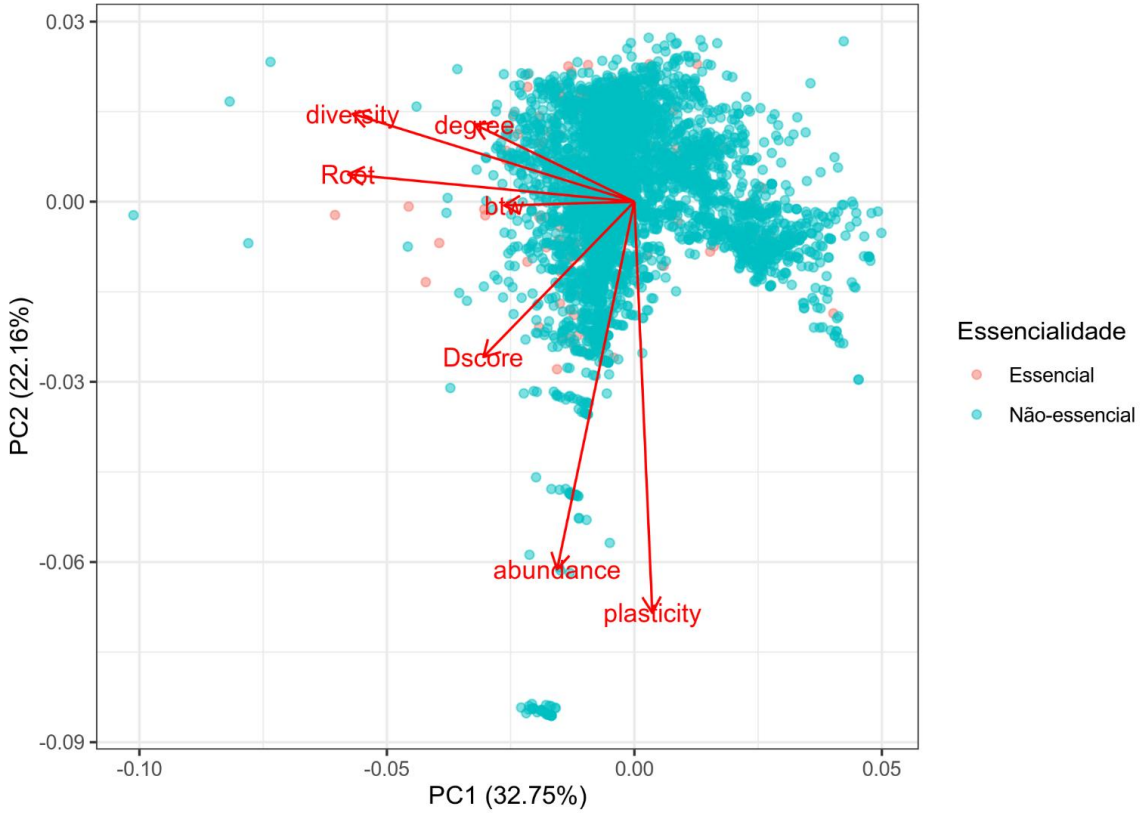


Análise exploratória

Camundongo



Levedura



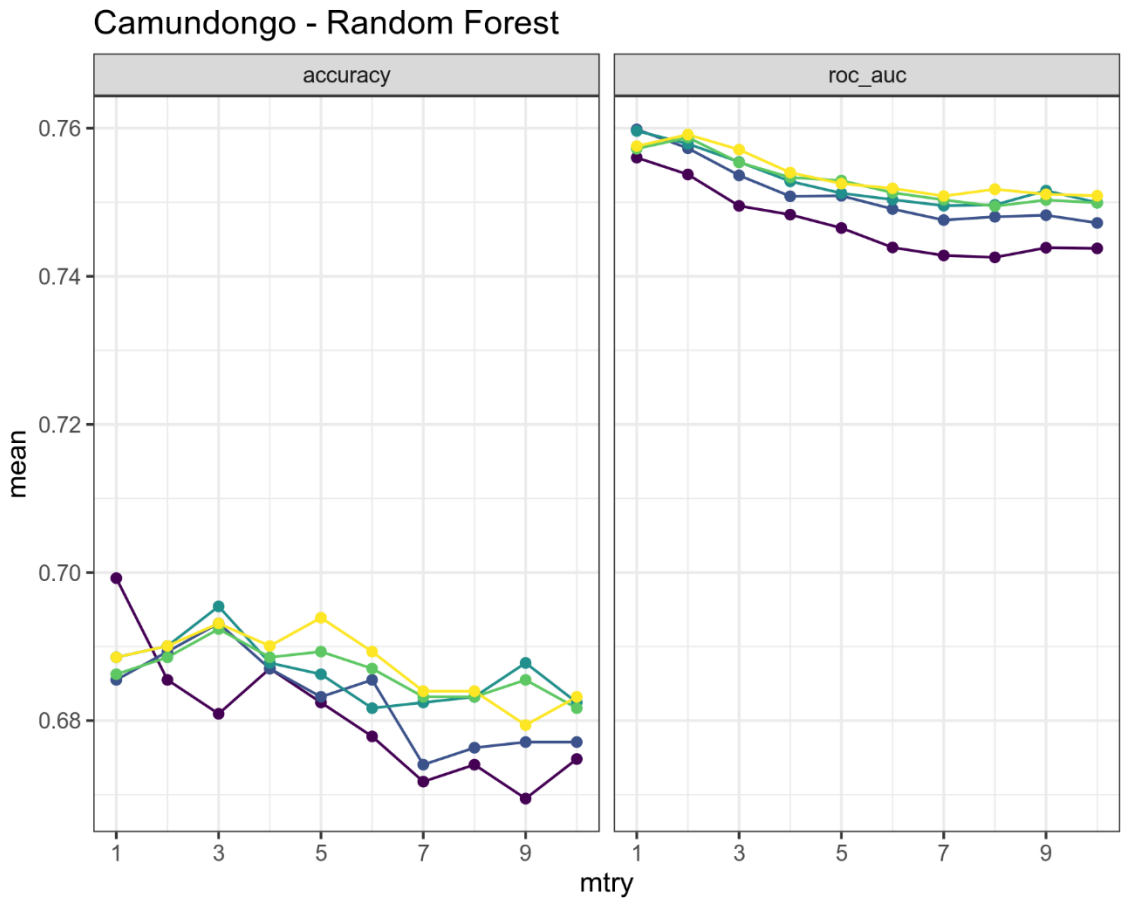
Resultados – Algoritmos usados

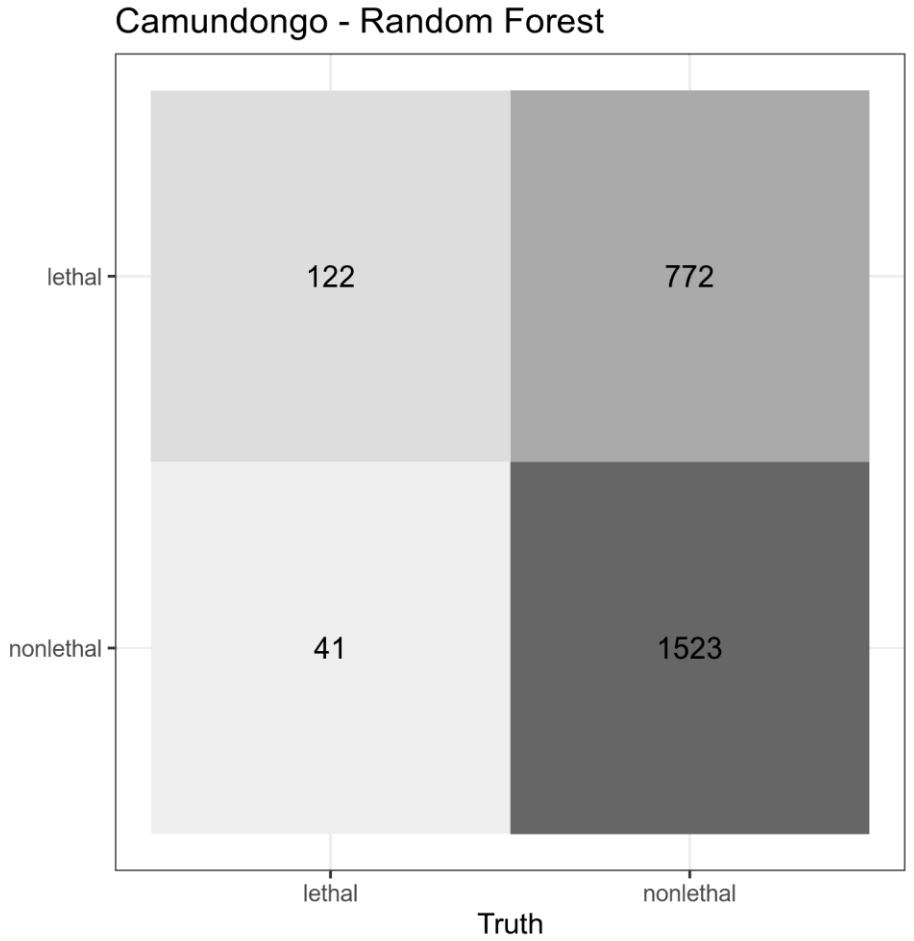
- Random Forest:
  - `mtry`: 10 níveis, de 1 a 10.
  - `min_n`: 5 níveis, de 10 a 50.

- SVM polinomial:
  - `cost`: 10 níveis, de -10 a 10
  - `degree`: 2 níveis (1 e 3)

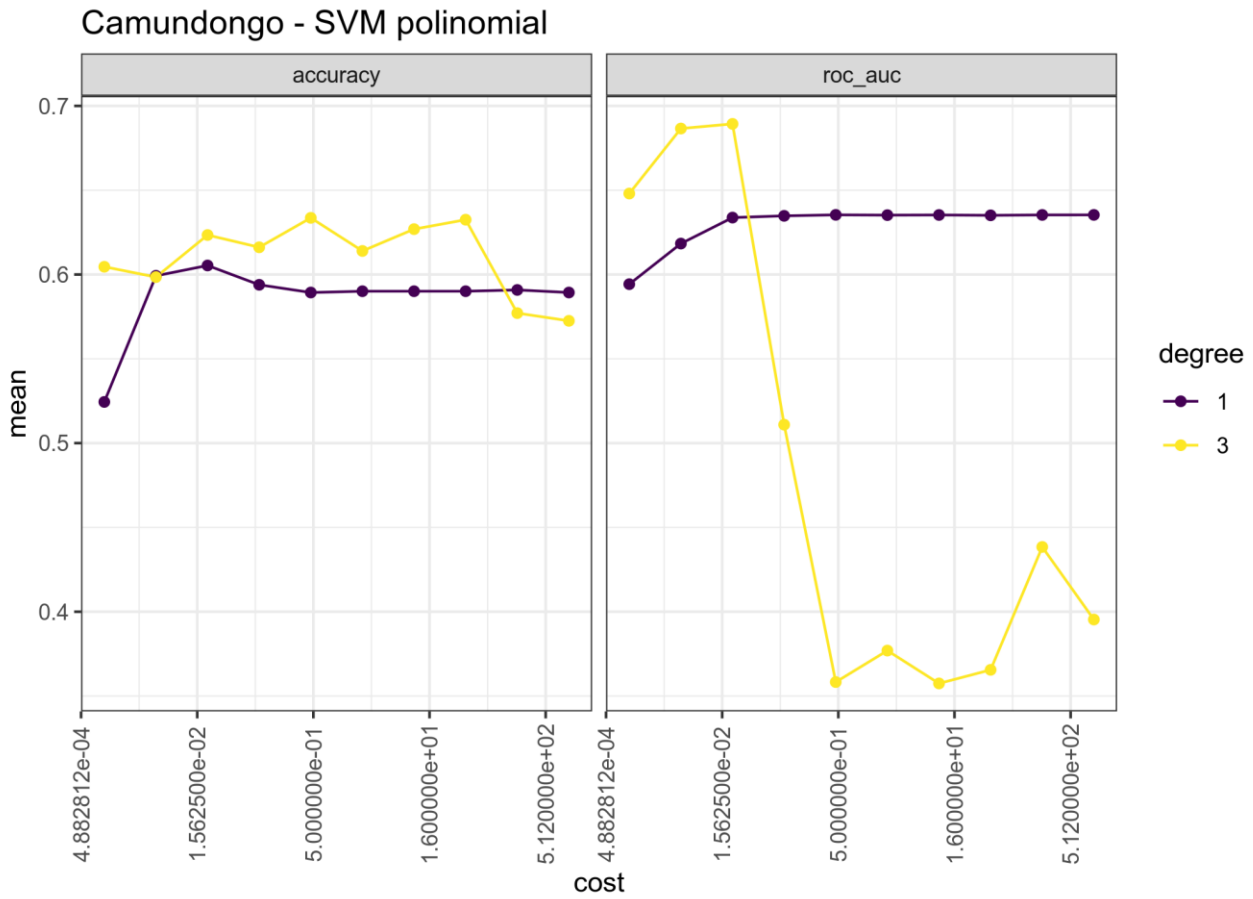
- SVM radial:
  - `cost`: 10 níveis, de -10 a 10.
  - `rbf_sigma`: 5 níveis

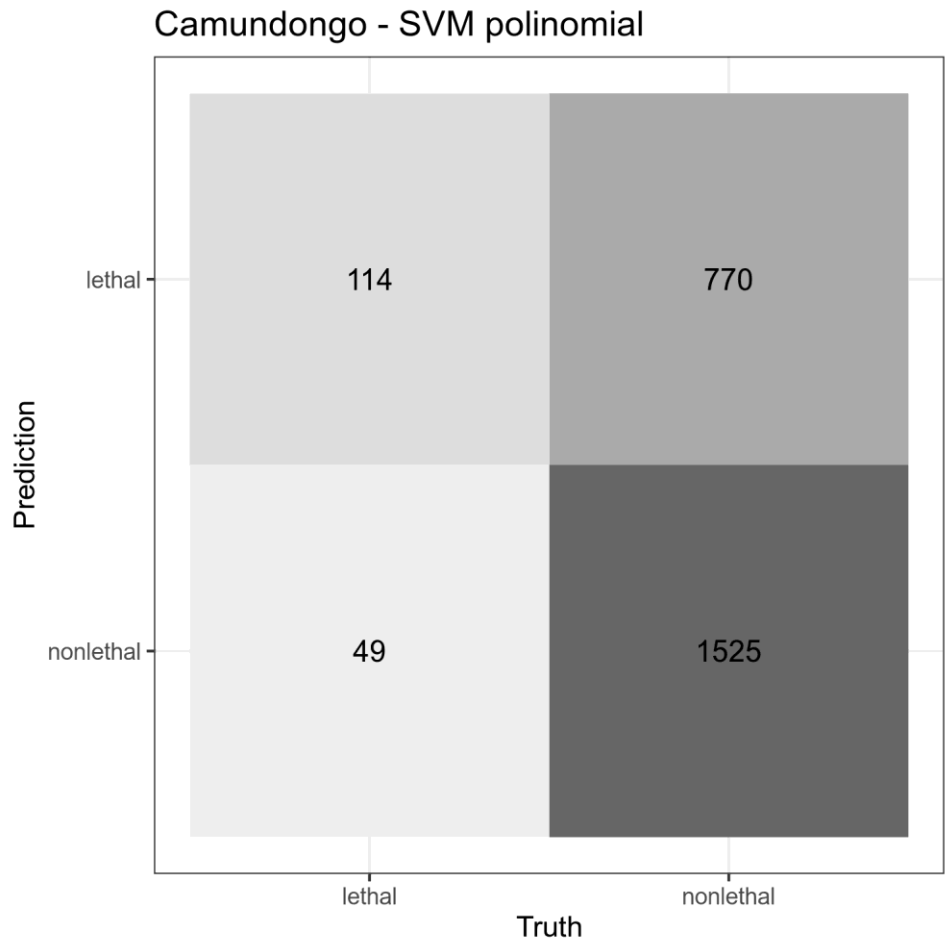
Resultados


Camundongo - Random Forest
accuracy
roc_auc
mean
0.76
0.74
0.72
0.70
0.68
1
3
5
7
9
mtry
min_n
10
20
30
40
50

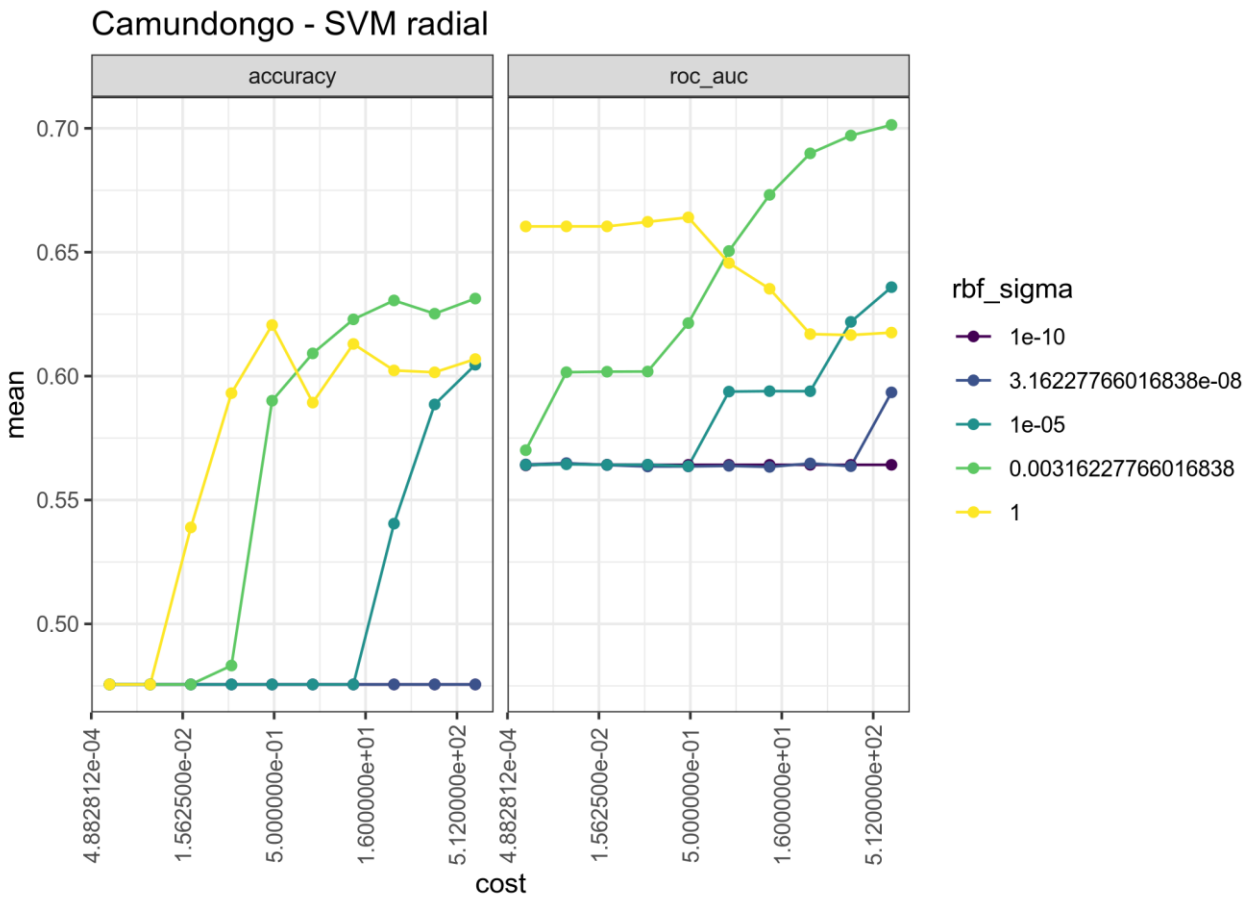

Camundongo - Random Forest
Prediction
lethal
nonlethal
122
772
41
1523
lethal
nonlethal
Truth

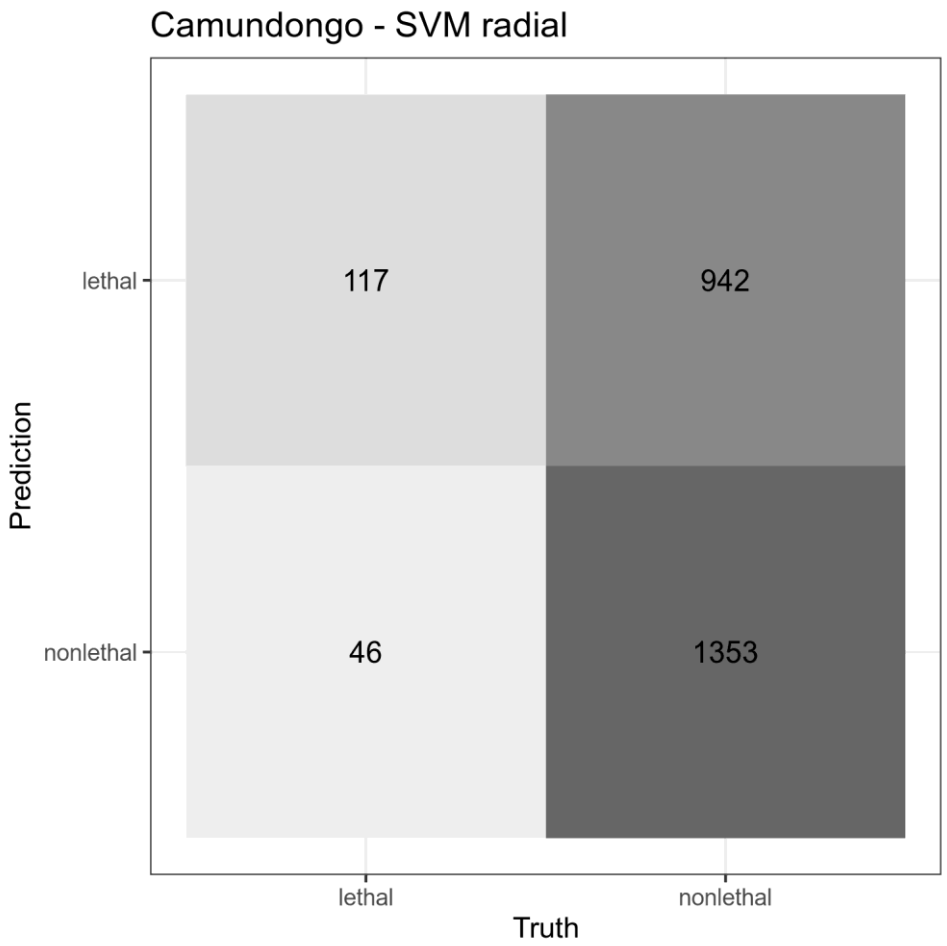
Resultados


Camundongo - SVM polinomial
accuracy
roc_auc
mean
0.7
0.6
0.5
0.4
4.882812e-04
1.562500e-02
5.000000e-01
1.600000e+01
5.120000e+02
cost
degree
1
3

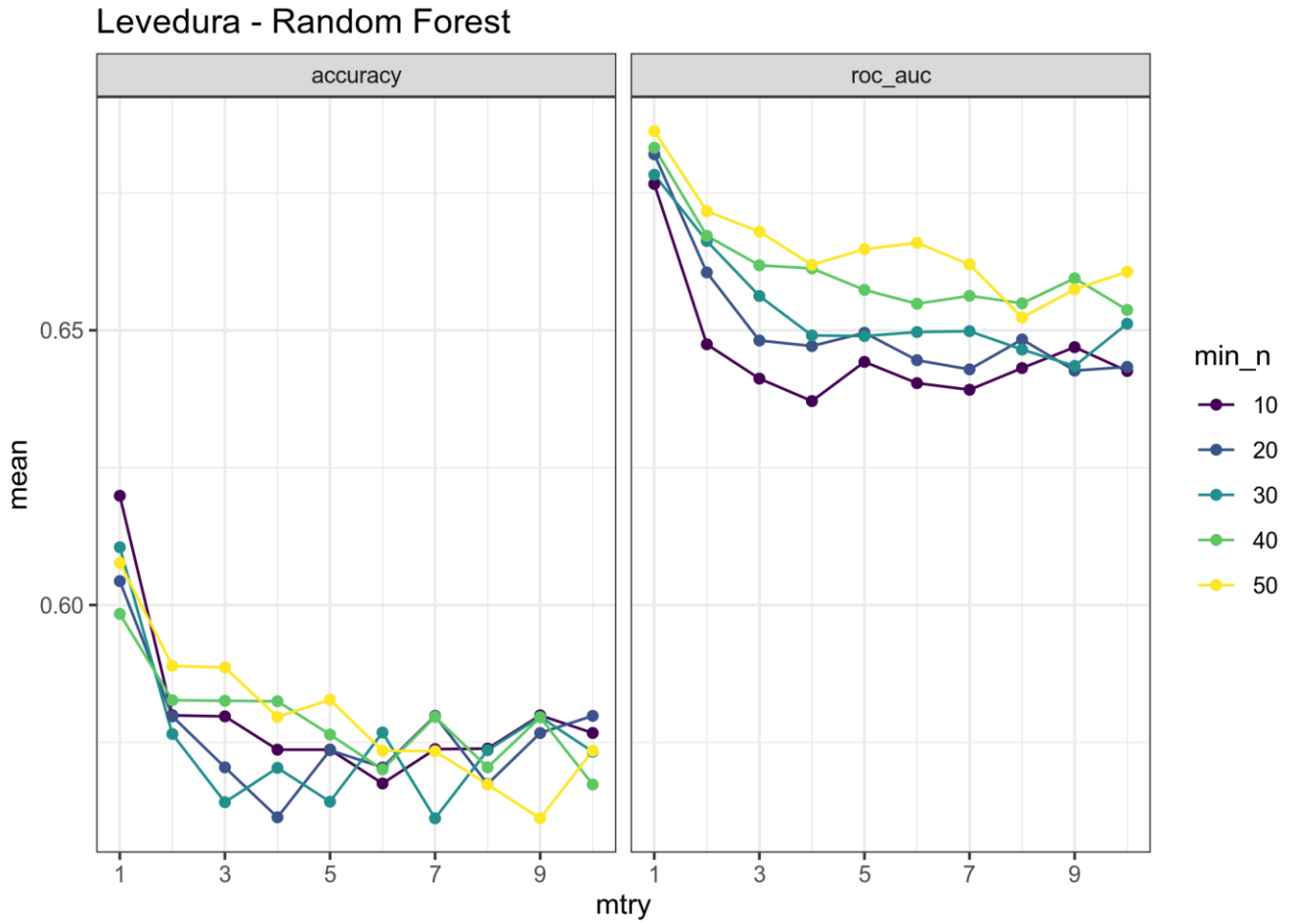

Camundongo - SVM polinomial
Prediction
lethal
nonlethal
114
770
49
1525
lethal
nonlethal
Truth

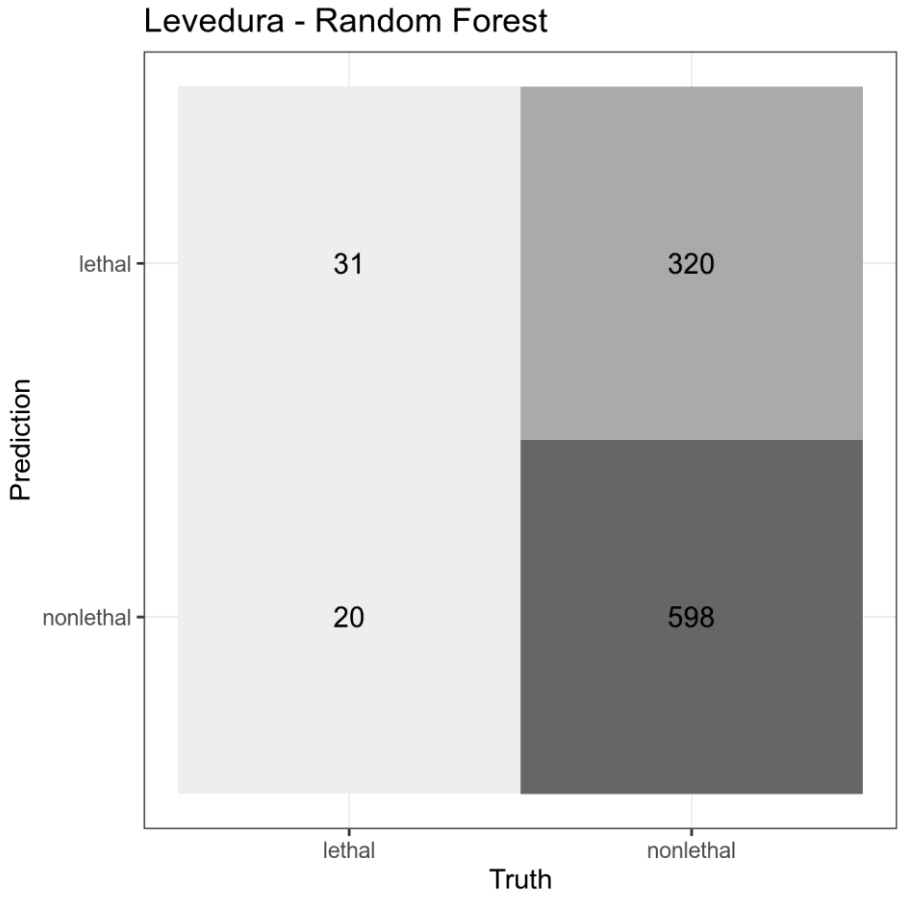
Resultados


Camundongo - SVM radial
accuracy
roc_auc
mean
0.70
0.65
0.60
0.55
0.50
4.882812e-04
1.562500e-02
5.000000e-01
1.600000e+01
5.120000e+02
cost
rbf_sigma
1e-10
3.16227766016838e-08
1e-05
0.00316227766016838
1

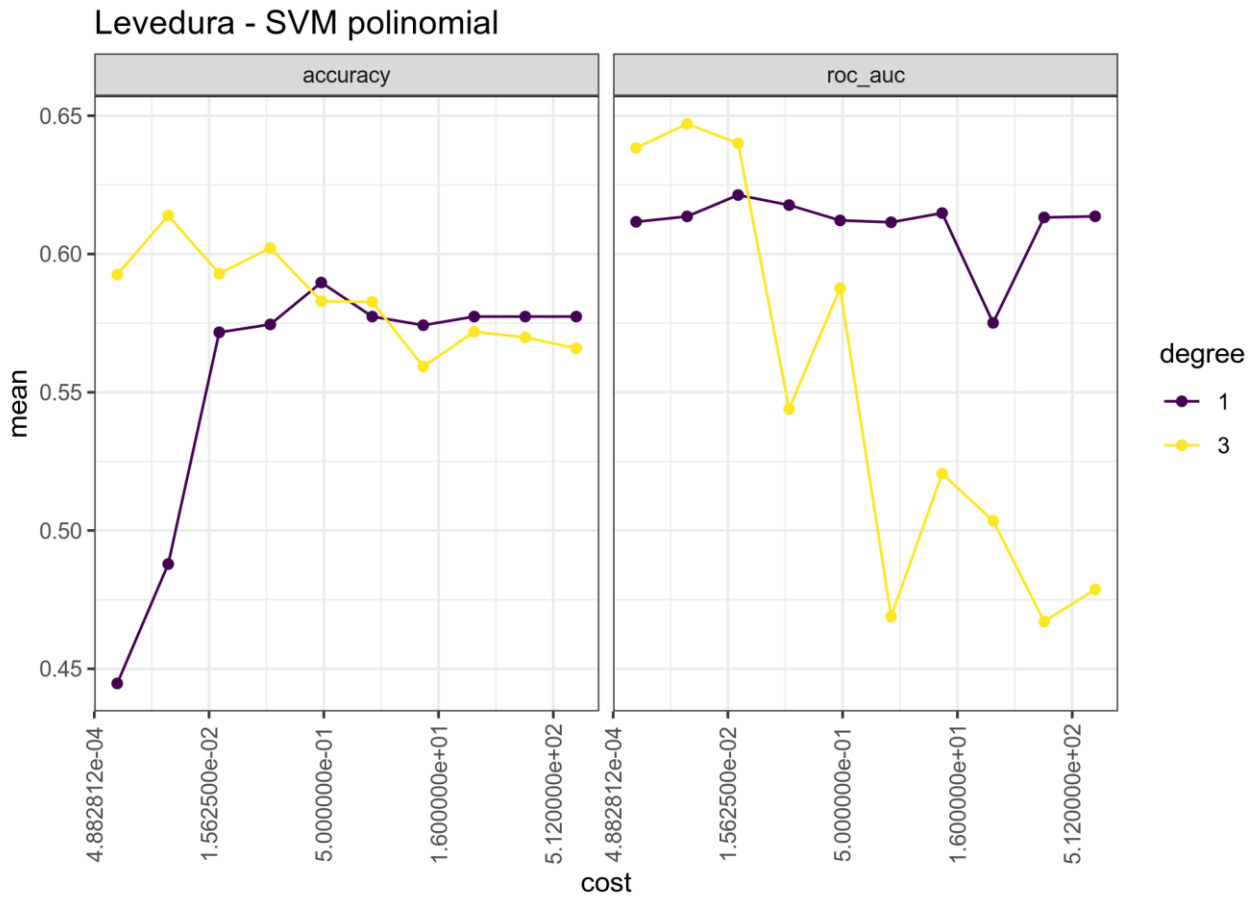

Camundongo - SVM radial
Prediction
lethal
nonlethal
117
942
46
1353
lethal
nonlethal
Truth

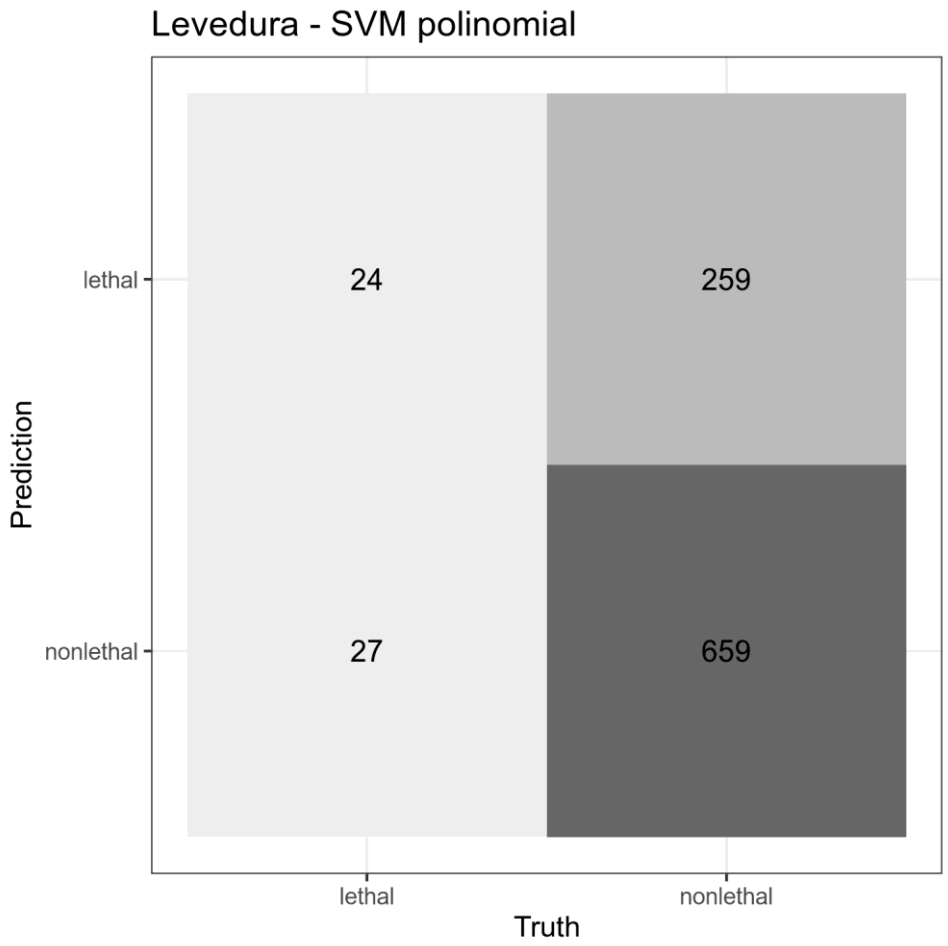
Resultados


Levedura - Random Forest
accuracy
roc_auc
mean
0.65
0.60
1
3
5
7
9
mtry
min_n
10
20
30
40
50

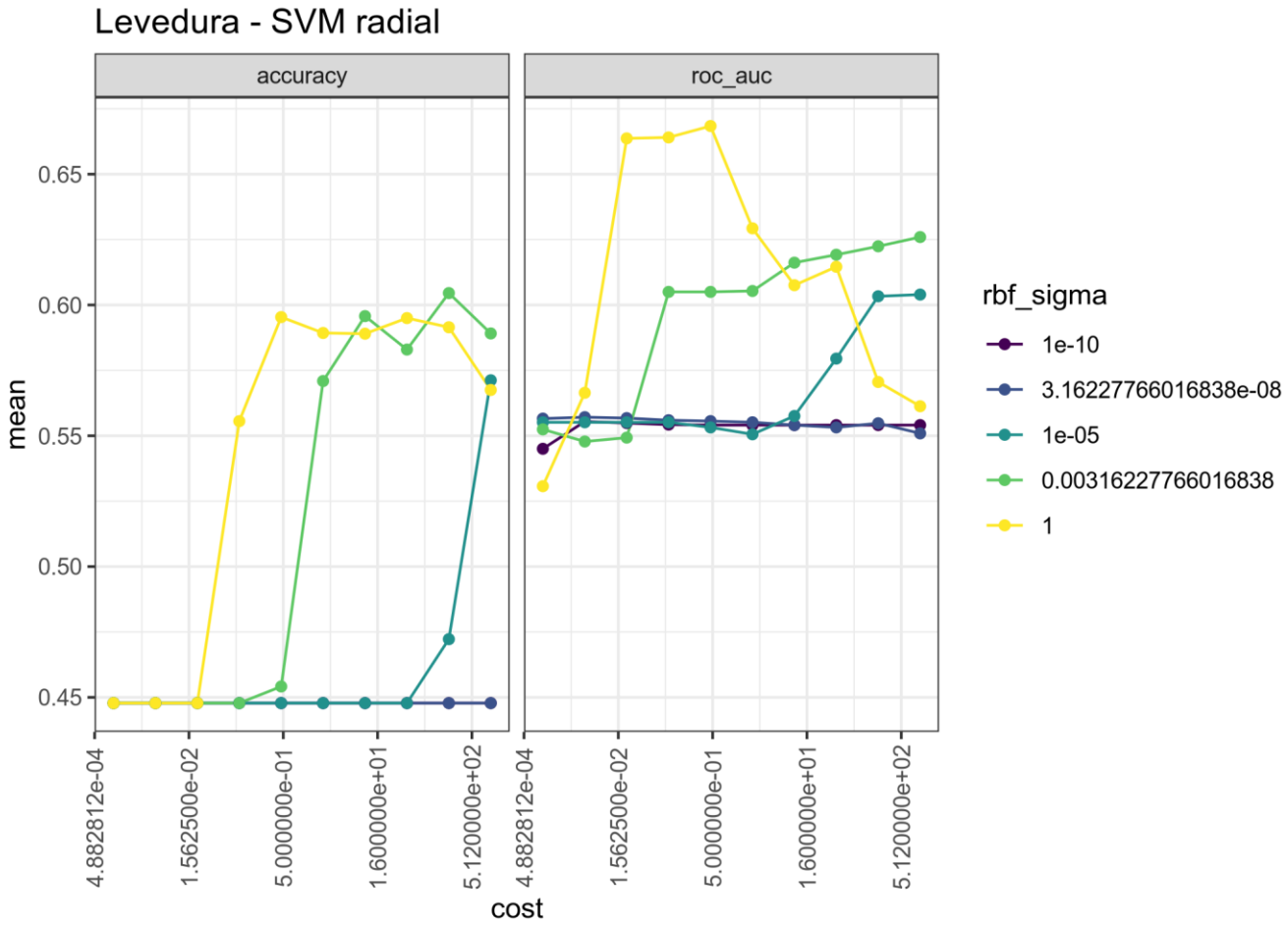

Levedura - Random Forest
Prediction
lethal
nonlethal
31
320
20
598
lethal
nonlethal
Truth

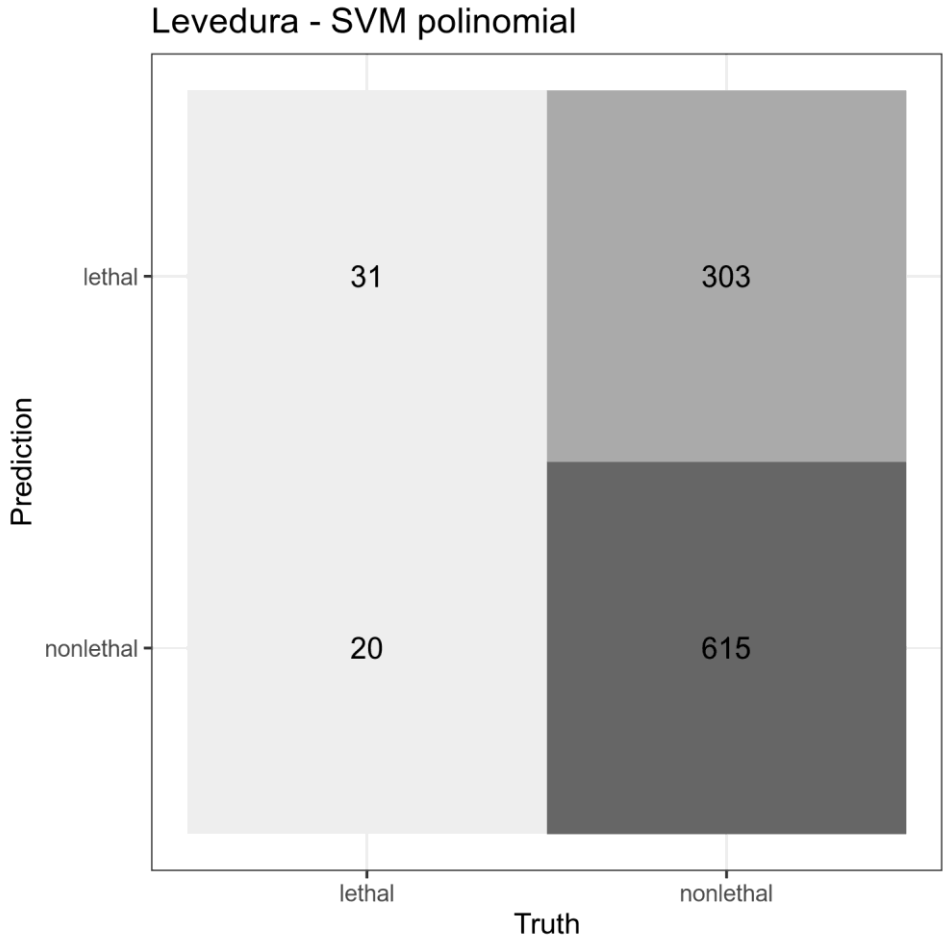
Resultados


Levedura - SVM polinomial
accuracy
roc_auc
mean
0.65
0.60
0.55
0.50
0.45
4.882812e-04
1.562500e-02
5.000000e-01
1.600000e+01
5.120000e+02
cost
degree
1
3


Levedura - SVM polinomial
Prediction
lethal
nonlethal
24
259
27
659
lethal
nonlethal
Truth

Resultados


Levedura - SVM radial
accuracy
roc_auc
mean
0.65
0.60
0.55
0.50
0.45
4.882812e-04
1.562500e-02
5.000000e-01
1.600000e+01
5.120000e+02
cost
rbf_sigma
1e-10
3.16227766016838e-08
1e-05
0.00316227766016838
1


Levedura - SVM polinomial
Prediction
lethal
nonlethal
31
303
20
615
lethal
nonlethal
Truth

Considerações finais

- Nenhum dos algoritmos produziu um bom modelo;
  - Todos abaixo de 70% de acurácia;

- Isto era esperado, devido a grande variabilidade das classes e ao fato de a essencialidade ser uma característica complexa;

- Os modelos podem ser melhorados com mais variáveis preditoras