

Data Mining

Trabalho Prático

Adson Leal
Caio Cargnin Cardoso
Diego Marzarotto





Objetivo

- desenvolver um modelo de classificação capaz de inferir o desfecho do caso
- proposta inicial: inferir o tratamento indicado para cada caso
- problema: múltiplos tratamentos para um mesmo caso
- classes como combinações dos tratamentos aplicados para cada caso
- inviável pela explosão combinatória de classes
- decidiu-se por usar o desfecho do caso como classe (11 classes)



Motivação

- o modelo pode ser utilizado para realizar uma pré-triagem dos casos
- inferir o desfecho poderia ser útil para priorizar o atendimento
- detalhes do caso de uso dependem de melhor conhecimento do domínio
- exemplo: como priorizar um caso que o desfecho provável seja óbito?



Análise Exploratória

- "classificacao_gravidade" (01_caso_intoxicacao)
- "manifestacao_clinica" (01_caso_intoxicacao)
- "idade" (04_paciente)
- "especificacao_idade" (04_paciente)
- "periodo_gestacao" (04_paciente)
- "peso" (04_paciente)
- "sexo" (04_paciente)
- "intensidade_exposicao" (05_exposicao)
- "tempo_decorrido" (05_exposicao)
- "especificacao_tempo_decorrido" (05_exposicao)
- "via_exposicao" (05_exposicao)
- "circunstancia_exposicao" (05_exposicao)
- "classe_agente" (06_agente_intoxicante)
- "subclasse_agente" (06_agente_intoxicante)
- "grupo_agente" (06_agente_intoxicante)
- "manifestacao_apresentada" (07_manifestacao)
- "classificacao_manifestacao" (07_manifestacao)



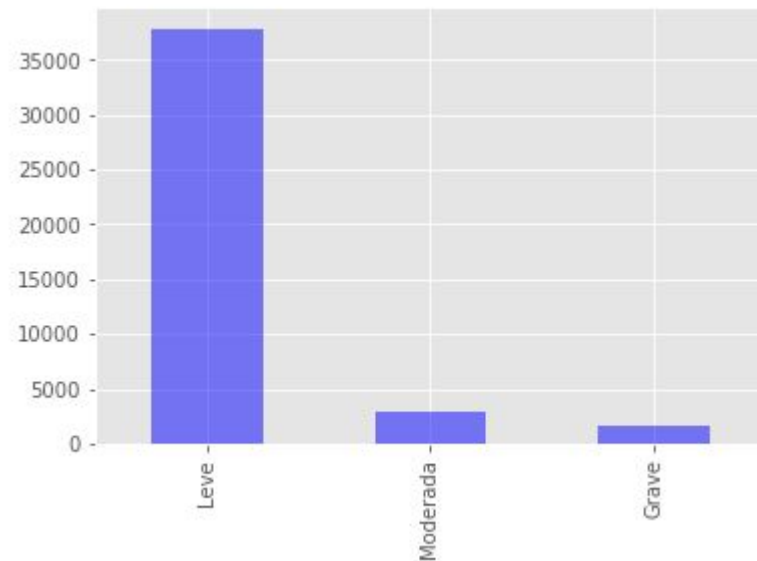
Tabela 01_caso_intoxicacao

- um registro por caso
- "classificacao_gravidade" (01_caso_intoxicacao)
- "manifestacao_clinica" (01_caso_intoxicacao)



Atributo *classificacao_gravidade*

- Leve 37837
- Moderada 2913
- Grave 1575





Atributo *manifestacao_clinica*

- Sim 35414
- Não 7237
- Ignorada 28

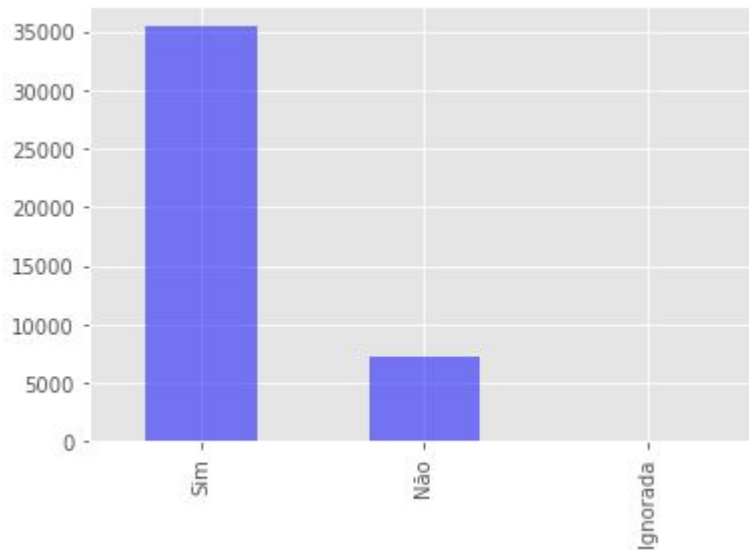


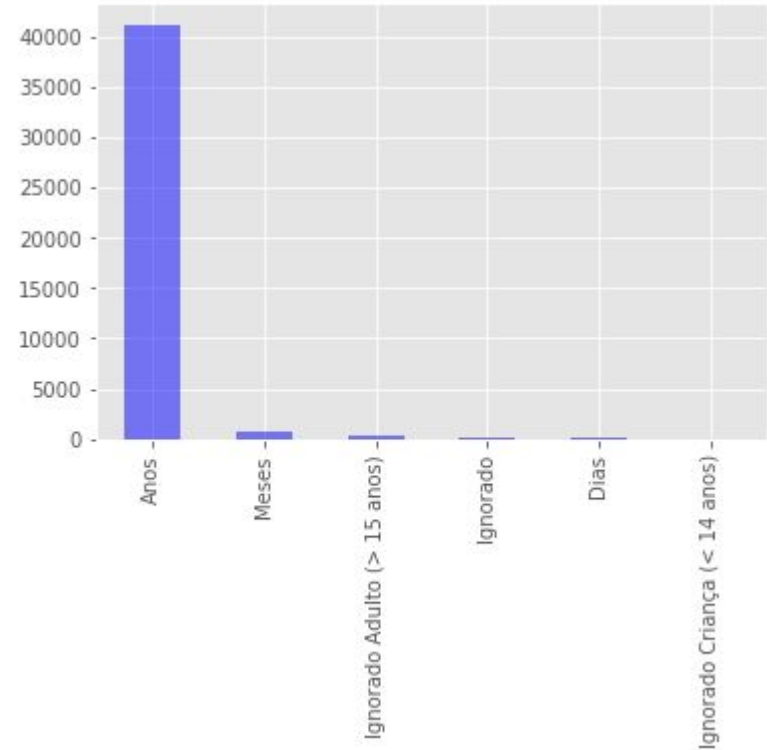


Tabela 04_paciente

- um registro por caso
- "idade" (04_paciente)
- "especificacao_idade" (04_paciente)
- "periodo_gestacao" (04_paciente)
- "peso" (04_paciente)
- "sexo" (04_paciente)

Atributo *especificacao_idade*

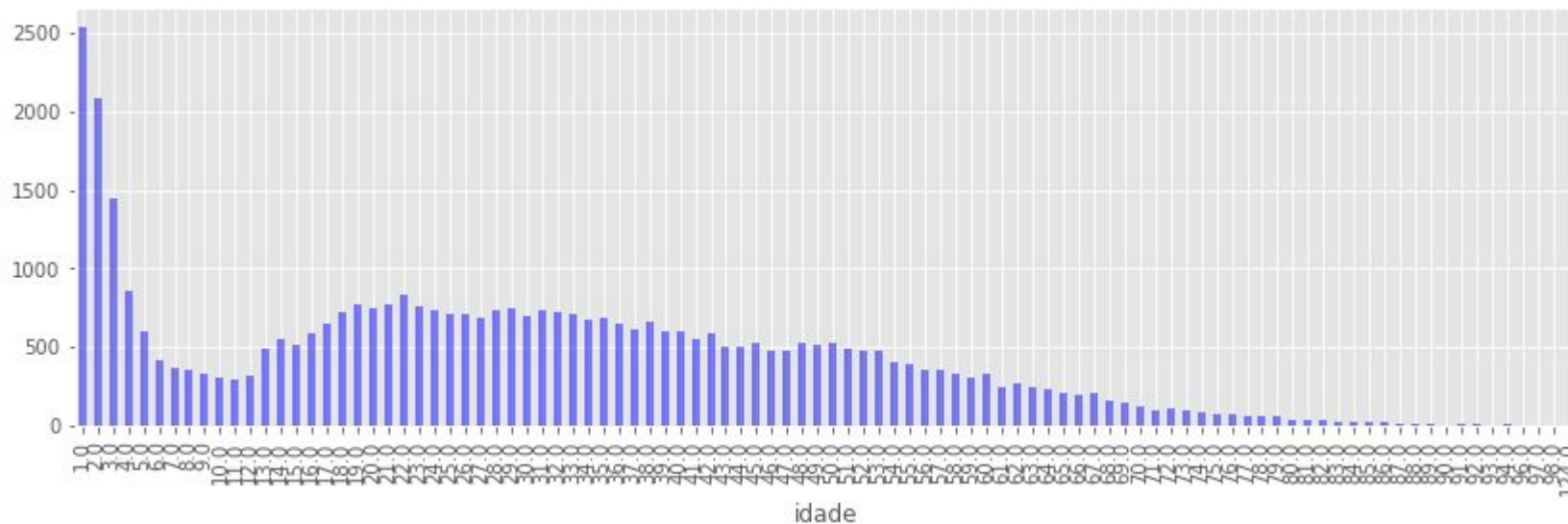
- Anos 41204
- Meses 813
- Ignorado Adulto (> 15 anos) 312
- Ignorado 113
- Dias 46
- Ignorado Criança (< 14 anos) 18





Atributo *idade*

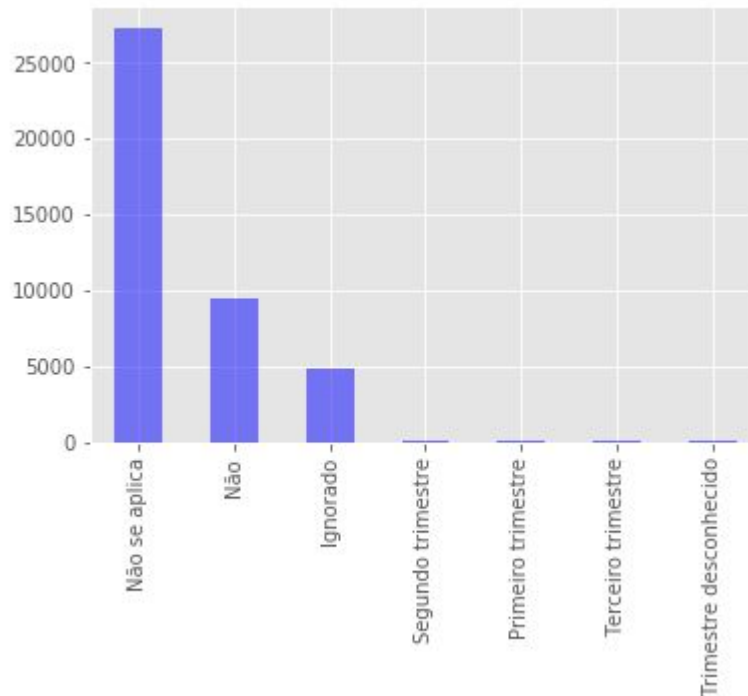
- bebê (até 2 anos)
- criança (2 até 10 anos)
- pré-adolescente (10 até 13 anos)
- adolescente (13 até 16 anos)
- jovem (18 até 30 anos)
- adulto (30 até 60 anos)
- idoso (acima de 60 anos)





Atributo *periodo_gestacao*

- Não se aplica 27308
- Não 9485
- Ignorado 4873
- Segundo trimestre 135
- Primeiro trimestre 119
- Terceiro trimestre 104
- Trimestre desconhecido 34



* Atributo generalizado (gestante: 0/1)

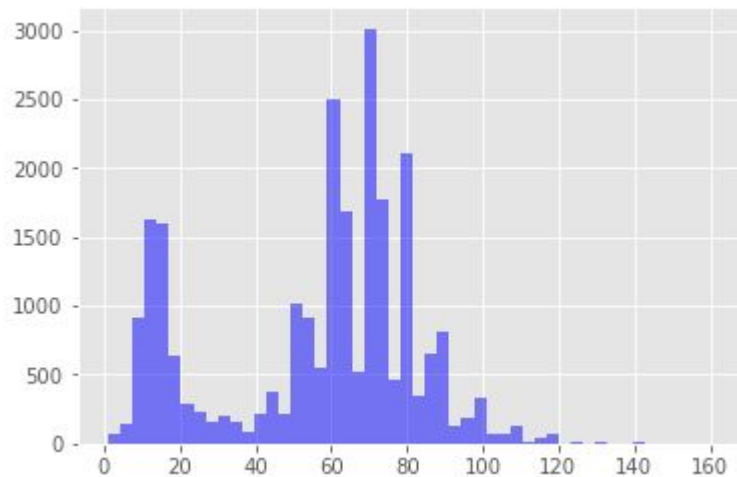


Atributo *peso*

- 500.0
- 500.0
- 200.0
- 200.0
- 162.0
- 150.0
- ...

- até 5 kg
- 5 até 15 kg
- 15 até 25 kg
- 25 até 35 kg
- 35 até 45 kg
- 45 até 60 kg
- 55 até 70 kg
- 65 até 80 kg
- 75 até 90 kg
- 85 até 95 kg
- acima de 95kg

* Outliers removidos (> 200)





Atributo *sexo*

- Feminino 21777
- Masculino 20651
- Ignorado 81

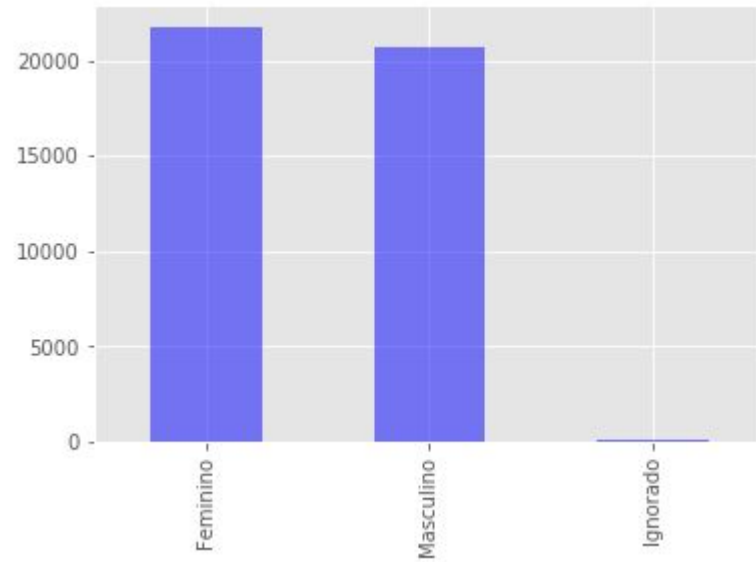




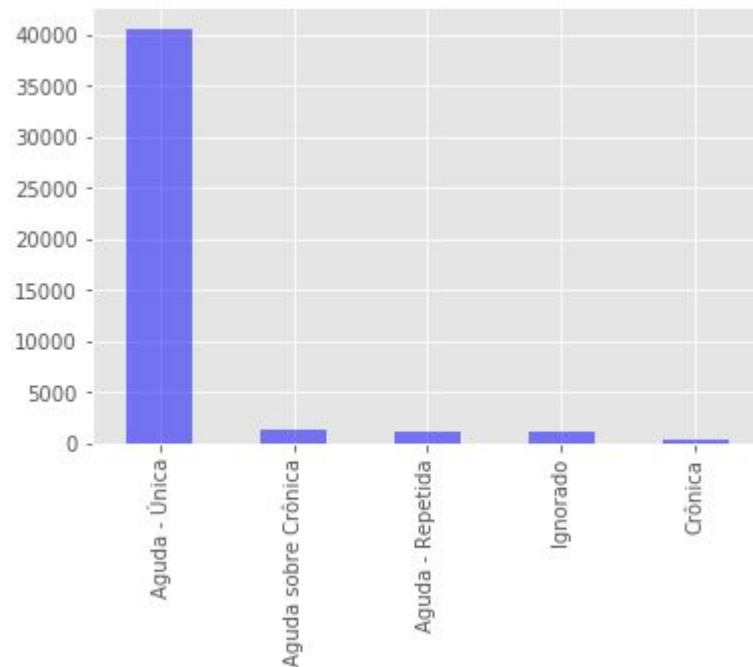
Tabela 05_exposição

- múltiplos registro por caso
- "especificacao_tempo_decorrido" (05_exposicao)
- "via_exposicao" (05_exposicao)
- "circunstancia_exposicao" (05_exposicao)



Atributo *intensidade_exposicao*

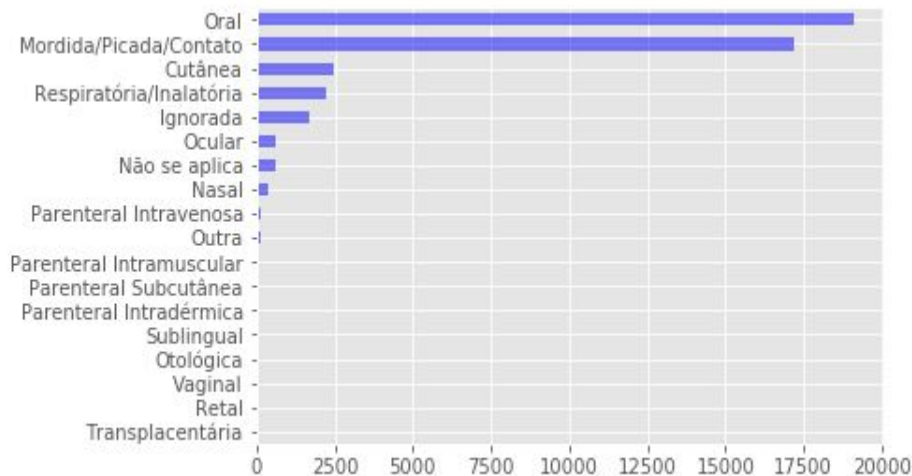
- Aguda - Única 40592
- Aguda sobre Crônica 1238
- Aguda - Repetida 1145
- Ignorado 1143
- Crônica 351





Atributo *via_exposicao*

| | |
|----------------------------|-------|
| • Oral | 19087 |
| • Mordida/Picada/Contato | 17204 |
| • Cutânea | 2451 |
| • Respiratória/Inalatória | 2239 |
| • Ignorada | 1707 |
| • Ocular | 630 |
| • Não se aplica | 587 |
| • Nasal | 365 |
| • Parenteral Intravenosa | 116 |
| • Outra | 111 |
| • Parenteral Intramuscular | 71 |
| • ... | |





Atributo *circunstancia_exposicao*

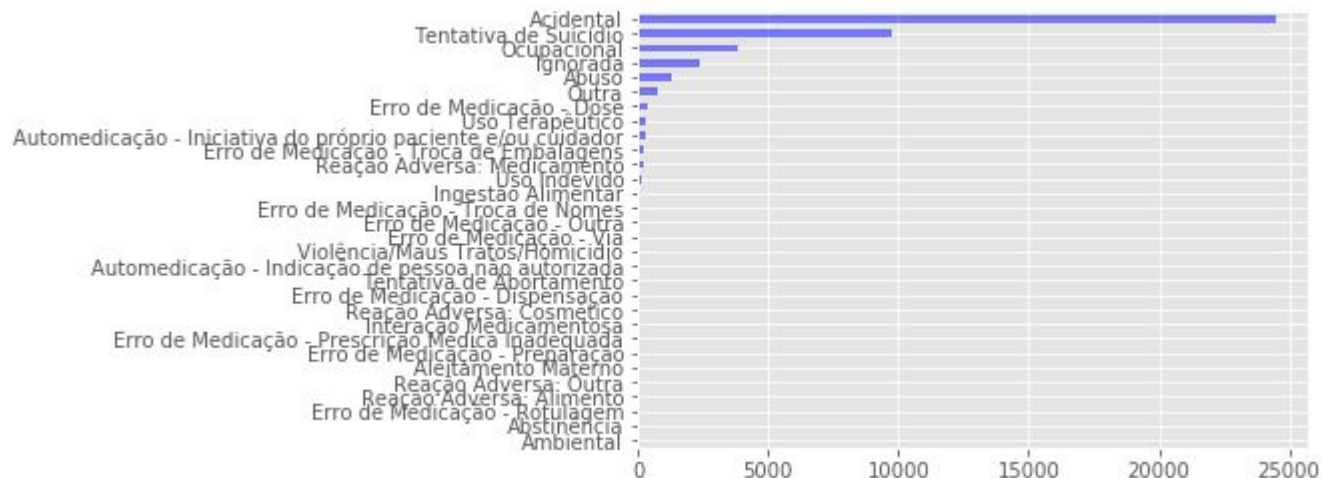




Tabela 06_agente_intoxicante

- múltiplos registro por caso
- "classe_agente" (06_agente_intoxicante)
- "subclasse_agente" (06_agente_intoxicante)
- "grupo_agente" (06_agente_intoxicante)



Atributo *grupo_agente*





Atributo *classe_agente*

- 226 valores possíveis



Atributo *subclasse_agente*

- 612 valores possíveis

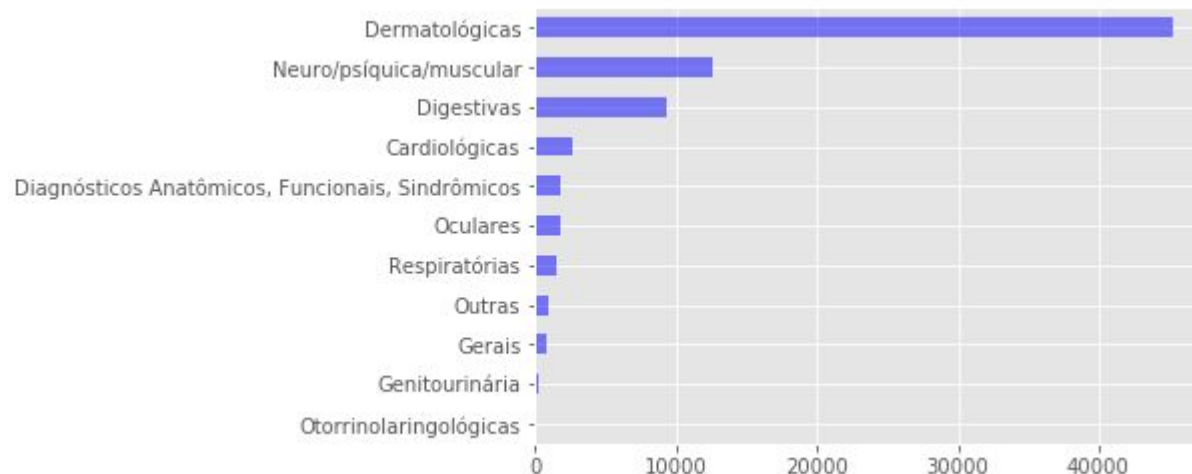


Tabela 07_exposicao

- múltiplos registros por caso
- "manifestacao_apresentada" (07_manifestacao)
- "classificacao_manifestacao" (07_manifestacao)



Atributo *classificacao_manifestacao*





Atributo *manifestacao_apresentada*

- 208 valores possíveis



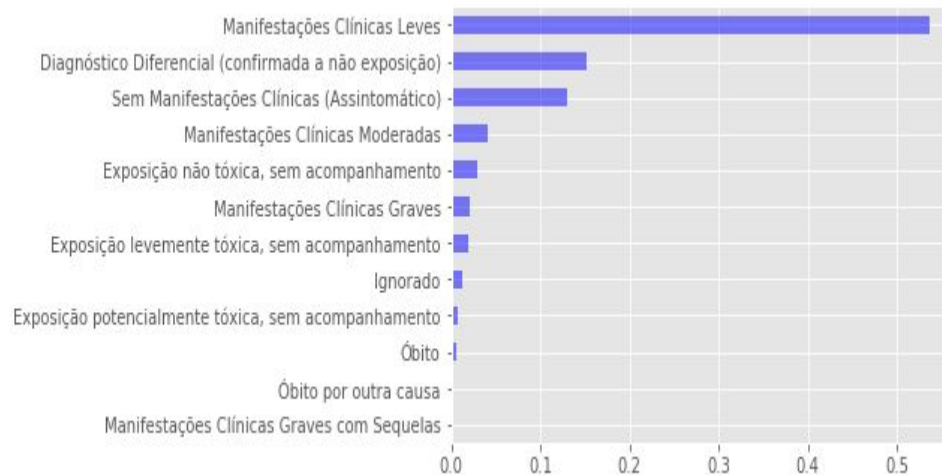
Tabela 12_encerramento

- um único registro por caso
- "manifestacao_apresentada" (07_manifestacao)
- "classificacao_manifestacao" (07_manifestacao)



Atributo *desfecho*

| | |
|--|----------|
| • Manifestações Clínicas Leves | 0.536333 |
| • Diagnóstico Diferencial (confirmada a não exposição) | 0.151111 |
| • Sem Manifestações Clínicas (Assintomático) | 0.129828 |
| • Manifestações Clínicas Moderadas | 0.041427 |
| • Exposição não tóxica, sem acompanhamento | 0.029045 |
| • Manifestações Clínicas Graves | 0.019891 |
| • Exposição levemente tóxica, sem acompanhamento | 0.018667 |
| • Ignorado | 0.012424 |
| • Exposição potencialmente tóxica, sem acompanhamento | 0.007129 |
| • Óbito | 0.005273 |
| • Óbito por outra causa | 0.001561 |
| • Manifestações Clínicas Graves com Sequelas | 0.000823 |





Limpeza / Seleção / Transformação

- todos os atributos binarizados
- categorização do atributo “idade”
- categorização do atributo “peso”
- generalização do atributo “periodo_gestacao”



Conjunto de atributos #1

- 40311 exemplos
- número de atributos reduzido
- 107 atributos binários
- descarta os seguintes atributos:
 - “classe_agente”
 - “manifestacao_apresentada”
 - “subclasse_agente”



Conjunto de atributos #2

- 40311 exemplos
- número de atributos intermediário
- 541 atributos binários
- descarta os seguintes atributos:
 - “classe_agente”
 - “manifestacao_apresentada”



Conjunto de atributos #3

- 40311 exemplos
- número de atributos ampliado
- 1153 atributos binários
- utiliza todos os atributos



Classes

- valores do atributo “desfecho”
- 11 classes possíveis
- somente exemplos com desfecho



Proposta

- 3 modelos treinados com número distintos de atributos
- primeiro modelo: número menor de atributos (107)
- segundo modelo: número intermediário de atributos (541)
- terceiro modelo: número maior de atributos (1153)
- conjuntos de dados formado por 40311 exemplos
- exemplos separados em 3 conjuntos distintos (treinamento, validação e teste)
- conjunto de dados separado na proporção 80/20 em treinamento e testes
- conjunto de treinamento separado na proporção 80/20 em treinamento e validação
- modelo final em cada experimento formado por um comitê de modelos
- hiperparâmetros selecionados de acordo com acurácia no conjunto de validação
- modelo com melhor resultado no conjunto de validação será o modelo escolhido



Modelos

- naive bayes multinomial
- naive bayes binomial
- random forest
- gradient boosted tree
- regressão logística
- svm (linear)
- svm (kernel)
- ensemble (comitê top-5)



Naive Bayes

```
('naive bayes (multinomial)', MultinomialNB(), [{  
    'alpha': [0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99, 0.999, 1]  
}]),  
  
('naive bayes (binomial)', BernoulliNB(), [{  
    'alpha': [0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99, 0.999, 1]  
}]),
```



Árvore de decisão (CART)

```
('árvore de decisão (cart)', DecisionTreeClassifier(random_state=SEED), [{  
    'criterion': ['gini', 'entropy'],  
    "min_samples_split": [2, 5, 10, 25, 50],  
    "max_depth": [None, 2, 5, 10, 25, 50],  
    "min_samples_leaf": [2, 5, 10, 25, 50],  
    "max_leaf_nodes": [None, 5, 10, 25, 50]  
}]),
```



Random Forest

```
('random forest', RandomForestClassifier(random_state=SEED), [{  
    'n_estimators': [1, 2, 3, 4],  
    'max_depth': [2, 5, 10, 25, 50],  
    "min_samples_split": [2, 5, 10, 25, 50]  
}]),
```



Gradient Boosted Tree

```
('gradient boosted tree', GradientBoostingClassifier(random_state=SEED), [{  
    'n_estimators': [1, 2, 3, 4],  
    'learning_rate': [0.1, 0.3, 0.5],  
    'max_depth': [2, 5, 10, 25, 50],  
}]),|
```



Regressão Logística

```
('regressão logística', SGDClassifier(random_state=SEED), [{  
    'loss': ['log'],  
    'penalty': ['l2'],  
    'alpha': [0.00001, 0.0001, 0.001, 0.01, 0.01, 0.1],  
    'eta0': [0.00001, 0.0001, 0.001, 0.01, 0.01, 0.1],  
}]),
```



SVM

```
('svm (linear)', LinearSVC(random_state=SEED) ,[{  
    'loss': ['hinge', 'squared_hinge'],  
    'multi_class': ['ovr'],  
    'penalty': ['l2'],  
    'C': [0.01, 0.1, 1, 10, 100, 1000]  
}]),
```

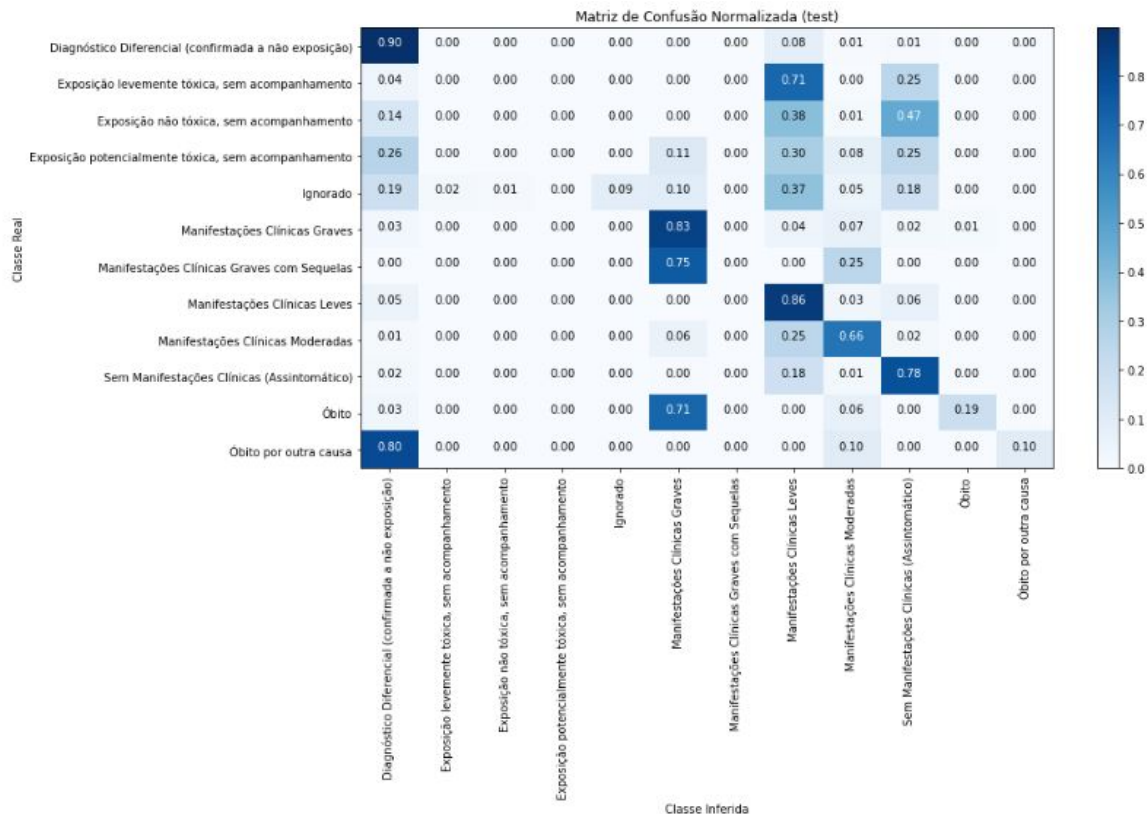
```
('svm (kernel)', SVC(random_state=SEED), [  
    {  
        'kernel': ['poly'],  
        'C': [0.01, 0.1, 1, 10, 100, 1000],  
        'decision_function_shape' : ['ovr'],  
        'degree' : [2, 3, 4, 5]  
    },  
    {  
        'kernel': ['rbf'],  
        'C': [0.01, 0.1, 1, 10, 100, 1000],  
        'decision_function_shape' : ['ovr']  
    }  
])
```



Modelo #1

- Acurácia (comitê): 0.781

| | |
|---------------------------|----------|
| svm (linear) | 0.782946 |
| árvore de decisão (cart) | 0.782791 |
| svm (kernel) | 0.781860 |
| gradient boosted tree | 0.781085 |
| random forest | 0.779070 |
| regressão logística | 0.777829 |
| naive bayes (multinomial) | 0.746822 |
| naive bayes (binomial) | 0.742636 |

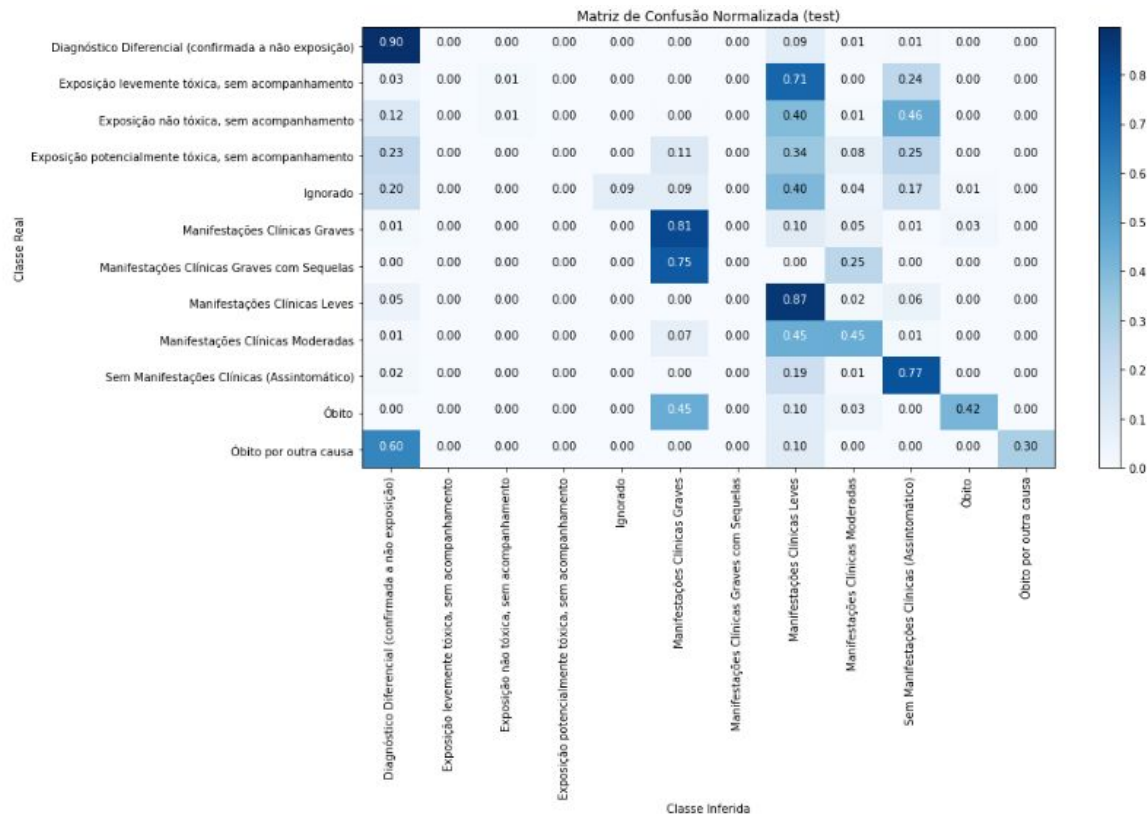




Modelo #2

- Acurácia (comitê): 0.7747

| | |
|---------------------------|----------|
| svm (linear) | 0.784186 |
| árvore de decisão (cart) | 0.783256 |
| svm (kernel) | 0.783256 |
| gradient boosted tree | 0.783101 |
| regressão logística | 0.778450 |
| random forest | 0.777209 |
| naive bayes (multinomial) | 0.704496 |
| naive bayes (binomial) | 0.700310 |

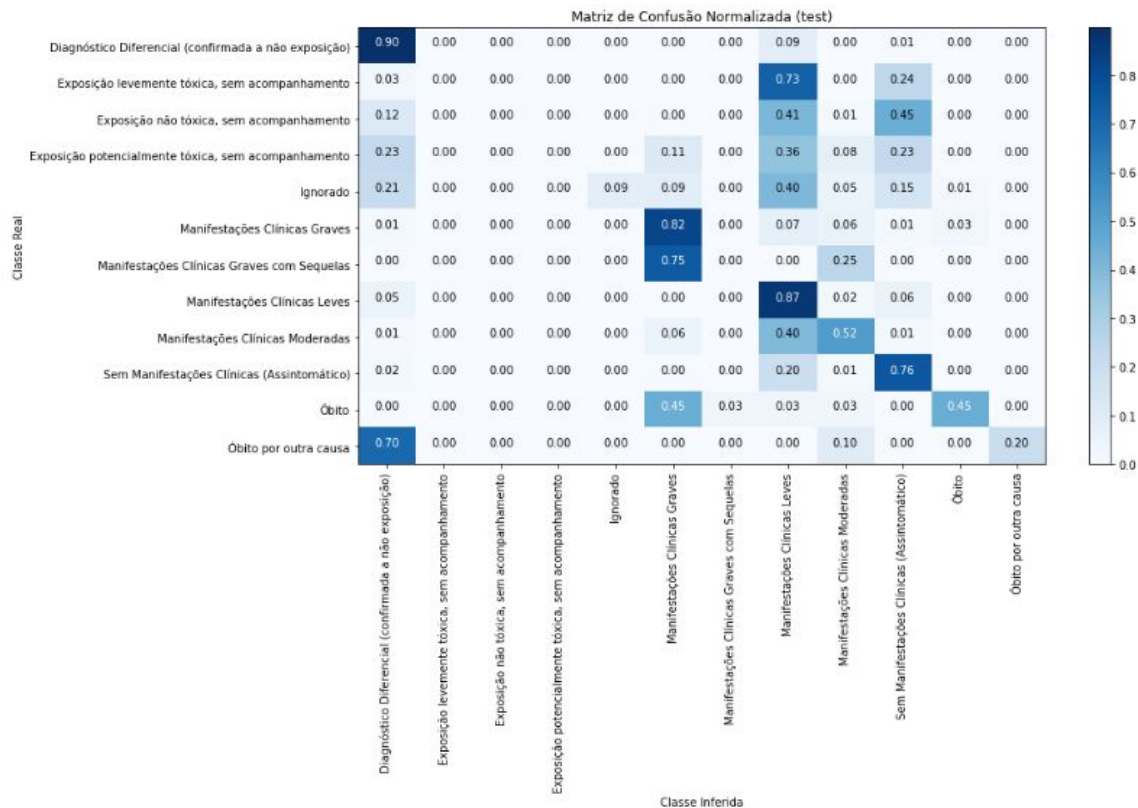




Modelo #3

- Acurácia (comitê): 0.7758

| | |
|---------------------------|----------|
| svm (linear) | 0.784341 |
| gradient boosted tree | 0.783101 |
| árvore de decisão (cart) | 0.782946 |
| svm (kernel) | 0.782481 |
| regressão logística | 0.780620 |
| random forest | 0.766977 |
| naive bayes (multinomial) | 0.692403 |
| naive bayes (binomial) | 0.691783 |





Conclusões

- modelos mais simples pioraram com um número maior de atributos
- modelos mais complexos melhoraram com um número maior de atributos
- comitê teve um desempenho pior que alguns modelos sozinhos
- melhor modelo: 0.78308 acurácia no conjunto de testes (svm linear + conjunto #3)
- modelo 1: 0.77911 acurácia no conjunto de testes (comitê)
- modelo 2: 0.7775 acurácia no conjunto de testes (comitê)
- modelo 3: 0.77936 acurácia no conjunto de testes (comitê)



Possíveis Melhorias

- abordagens diferentes para codificar atributos contínuos (peso, idade)
- identificar atributos relevantes através de algum algoritmo de seleção de atributos
- criar modelo utilizando somente atributos mais relevantes
- utilizar outros modelos de classificação no comitê (redes neurais)
- utilizar pesos baseados na performance de cada modelo no comitê
- agrupar classes de acordo com um melhor entendimento do domínio
- utilizar métrica f1 na avaliação, mais indicada para classes desbalanceadas
- continuar busca de hiperparâmetros

Fim

Notebook - Parte 1

Notebook - Parte 2

