Análise de Pacientes com Suspeita de Dengue em 2016 da Região Metropolitana de Recife

Antônio Carlos Portela Rodrigues, Pedro Henrique Sousa de Moraes, Thiago Aquino dos Santos

Centro de Informática - UFPE

Abstract

A dengue é uma das arboviroses que que mais afligem a população brasileira. Por ser transmitida através de mosquitos, a população, que sofre pela falta de saneamento básico, é muito afetada pela doença. Este trabalho estuda os casos de dengue registrados em Recife no ano de 2016. Objetivamos estudar se locais com alta incidência de casos de dengue estão relacionados a uma quantidade maior ou menor de unidades de saúde de forma a verificar se elas exercem alguma influência nos casos. Além disso, estudaremos as relações entre os sintomas do caso com a gravidade da doença, para, a partir disso, verificar se existe um conjunto de sintomas específicos que está fortemente relacionado a gravidade da doença.

Keywords: Dengue, Recife, Análise, Classificação

1. Introdução

A dengue é uma das doenças transmitidas pelo mosquito *aedes aegypti*. Este inseto se reproduz um locais com água parada, e é bem comum na região metropolitana de Pernambuco, onde milhares de casos são registrados por ano. A investigação de grande quantidade de suspeitas de casos da doença não é suportada pelo município, apenas uma pequena parte da população faz os exames sorológicos para descobrir se estão doentes. Apesar disto, a prefeitura de Recife coleta dados[1] das pessoas que vão aos hospitais e unidades de atendimento médico em geral, catalogando informações pessoais e sintomas relatados.

[1] http://dados.recife.pe.gov.br/dataset

Estes dados estão disponíveis publicamente, e são o foco da investigação que será feita neste documento.

1.1. Objetivos

Este trabalho tem como objetivo, verificar possíveis fatores indicadores da dengue através de uma análise exploratória dos dados, e, com as informações obtidas, criar um modelo de aprendizagem de máquina para classificar pacientes de forma a verificar se estão com dengue baseado nas suas informações.

2. Obtenção dos Dados

A base de dados usada contem informações de pessoas que relataram sintomas da dengue[2] em Recife no ano de 2016. Os dados possuem ao todo 17389 registros de pessoas pessoas que se encaminharam ao sistema de saúde da cidade, cada um dos registros possui 121 colunas com informações de vários tipos.

Estas informações podem ser da pessoa, como idade, sexo, rua e bairro que reside e sintomas apresentados, podem ser também informações sobre a data e o local onde a pessoa se apresentou. Também existem dados que são adicionados posteriormente, como exames sorólogos, resultado da progressão dos sintomas, entre outros.

3. Pré-Processamento

O pré-processamento dos dados consistiu de duas etapas.

A primeira etapa foi a remoção de grande parte das colunas com informações desnecessárias para o objetivo do trabalho, como códigos de regiões ou de endereço postal, datas de coletas de exames, data do registro na unidade de saúde, código dos hospitais entre vários outros dados. Ao todo foram removidas 7 colunas de dados, restaram então 38 colunas.

A segunda etapa foi a troca dos valores de algumas colunas, isso foi feito para facilitar o entendimento dos dados, pois a representação numérica usada em quase todos os dados torna difícil sua compreensão. Dados de colunas

[2] http://dados.recife.pe.gov.br/dataset/2a9b1c39-0700-4ddf-9a10-b3c8d5d9396c/resource/2a2ef847-7063-462e-bf76-a49ebc9a6d13/download/casos-dengue2016.csv

que descreviam sintomas, resultados de exames, dados dos pacientes e informações do caso foram trocados por seus rótulos correspondentes, outros dados não sofreram modificações. Das 38 colunas restantes após a primeira etapa, 33 tiveram seus valores trocados.

dt_nascimento	tp_sexo	tp_gestante	tp_raca_cor	tp_escolaridade	no_bairro_residencia	febre	mialgia
31-10-1955	M	ignorado	ignorado	ignorado	MADALENA	sem resposta	sem resposta
08-06-1977	F	não se aplica	ignorado	ignorado	IPUTINGA	sem resposta	sem resposta
24-12-1944	M	ignorado	parda	ignorado	AGUA FRIA	sem resposta	sem resposta
17-06-1999	F	não se aplica	ignorado	ignorado	CASA AMARELA	sem resposta	sem resposta
22-09-1964	M	ignorado	ignorado	ignorado	TORROES	sem resposta	sem resposta
05-01-1951	M	ignorado	ignorado	ignorado	NOVA DESCOBERTA	sem resposta	sem resposta
31-01-2002	F	não se aplica	ignorado	ignorado	SAN MARTIN	sim	sim
10-08-1988	F	2º trimestre	branca	ignorado	SAO JOSE	sim	sim

Figura 1: Dados após processamento

4. Análise Exploratória

Para buscar possíveis relacionamentos entre os dados que pudessem estar ligados a dengue, fizemos uma análise exploratória dos dados, onde foram verificadas várias propriedades dos pacientes em busca de informações correlacionadas com a doença.

4.1. Análise dos Bairros

A primeira informações que verificamos foi se a região está relacionada de alguma forma com a quantidade de casos de dengue para isso fizemos um processamento na nossa base de dados original e agrupamos os casos por bairro. Esse processamento é problemático pois ele poderia variar com o tamanho da população de cada bairro.

Dado o problema da população, decidimos então verificar o percentual de casos de dengue de todos os casos registrados por bairro, isso nos dá uma percepção dos bairros com maior incidência de dengue dentre os pacientes que se encaminharam às unidade de saúde.

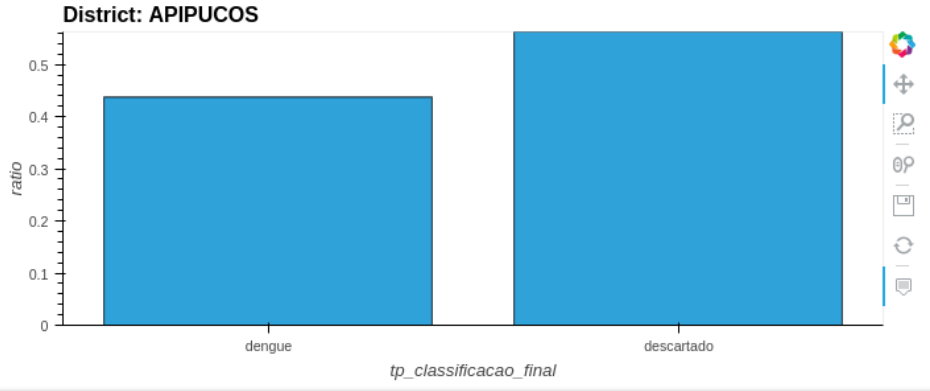

Figura 2: Bairro com proporção baixa de casos de dengue

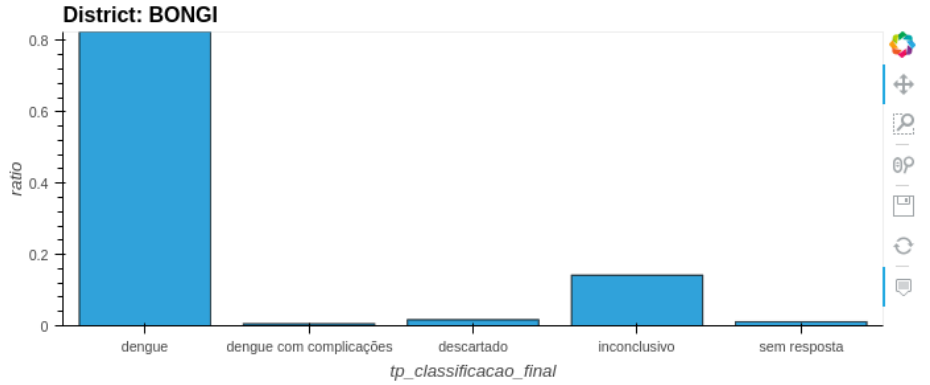

Figura 3: Bairro com proporção alta de casos de dengue

4.2. Gênero dos Pacientes

Uma das informações verificadas foi se o gênero dos pacientes estava relacionado com a quantidade de casos.

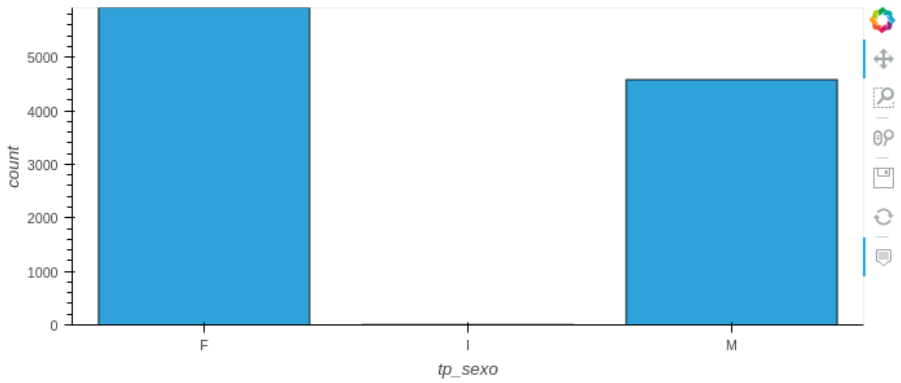

Figura 4: Contagem dos pacientes por gênero

A diferença vista é apenas um reflexo do descuido com a saúde, mais comum com os homens, que deixam de ir a hospitais quando estão doentes.

4.3. Nível de Escolaridade

Procuramos informações relacionando o nível de escolaridade dos pacientes e os casos da doença para verificar se pessoas com diferentes graus de escolaridade estão mais ou menos relacionados aos casos de dengue. Devido a falta de informações providas, já que eram ignoradas, não pudemos encontrar qualquer tipo de relacionamento com este dado.

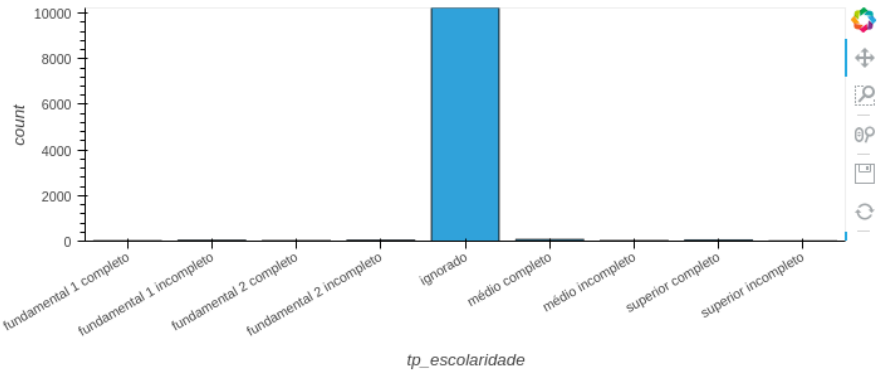

Figura 5: Notável falta de dados da escolaridade dos pacientes

4.4. Idade

Verificamos a idade dos pacientes em busca de faixas etárias que estavam mais propensas a ter dengue, com os resultados iniciais, onde foram

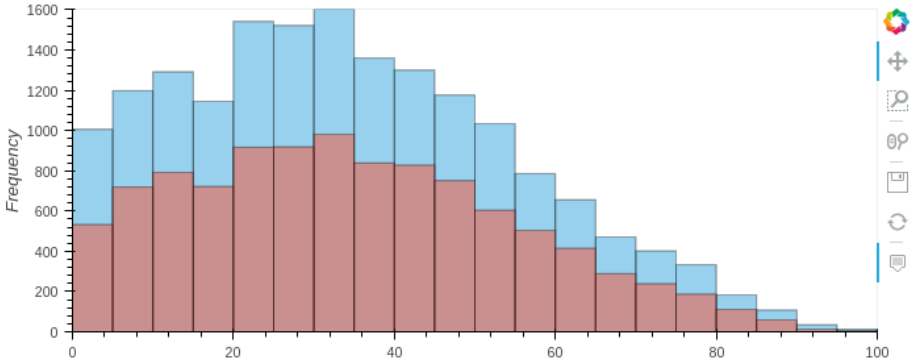

Figura 6: Proporção da idade de todos os pacientes (azul) e dos pacientes com dengue (vermelho)

analisados todos os casos, sendo comprovadamente de doença ou não.

Notamos que pessoas de 20 a 35 anos apresentam maior frequência de casos registrados, apesar disso esta faixa etária também é a que possui a maior população, de forma que a quantidade de pessoas por faixa etária que se apresentaram nas unidades de saúde é proporcional a população.

Então investigamos nesses grupos existe algum que possui mais casos de dengue dentre os todos os casos relatados, o grupo de pessoas da faixa etária com dengue também é proporcional a todo o grupo da faixa etária, indicando que a dengue não está relacionada com a idade.

4.5. Exames sorológicos

Vários tipos de exames sorológicos podem ser feitos para identificar casos de dengue, pensando nisto, fizemos uma análise dos resultados dos exames sorológicos para procurar correlações com outros dados. Ao iniciar a análise dos dados percebemos que esses exames são raramente feitos nos pacientes.

tp_classificacao_final	Tp_result_NS1	
dengue	inconclusivo	1
	não reagente	5
	não realizado	2654
	reagente	2
	sem resposta	7813
dengue com complicações	não realizado	15
	sem resposta	18
dengue hemorrágica	sem resposta	3

Figura 7: Um dos exames sorológicos para detectar dengue, quase nenhum paciente o faz

4.6. Sintomas

Os sintomas são ótimos dados para encontrar correlações com doenças, pensando nisso, decidimos analisar os sintomas apresentados pelos pacientes, verificando se estão correlacionados com os casos de dengue, os sintomas apresentaram várias correlações com casos de dengue, e são uns dos dados mais uteis para a os métodos de classificação pois podem ser obtidos do paciente de imediato. Vários sintomas como febre, mialgia, cefaleia, exantema, vômito, náusea, etc foram analisados.

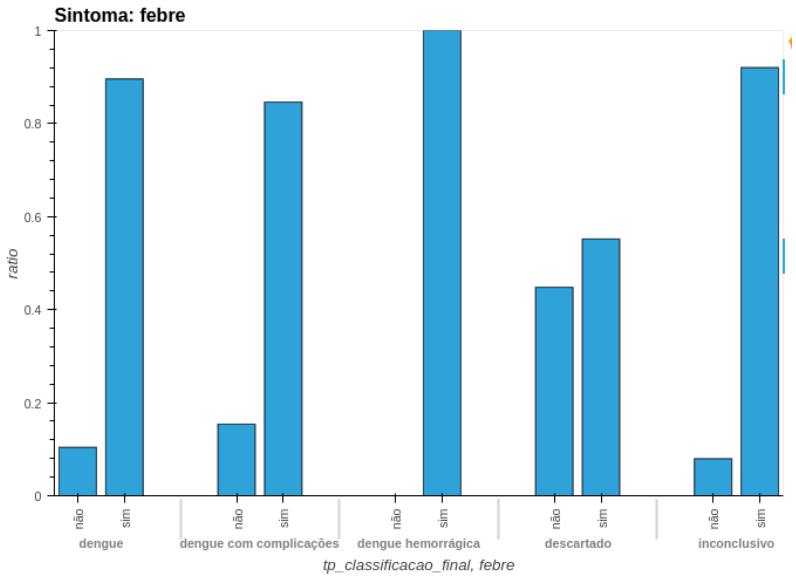

Figura 8: Sintoma da febre para cada tipo de dengue ou para casos descartados

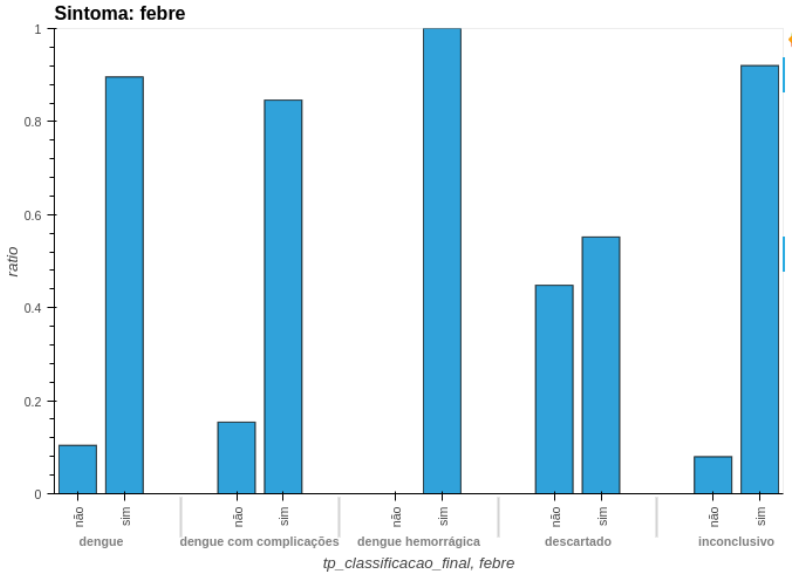

Figura 9: Sintoma da dor nas costas para cada tipo de dengue ou para casos descartados

5. Classificação

Com várias informações obtidas da análise exploratória, construímos um modelo de aprendizagem de máquina para classificação dos casos de dengue.

Usamos como entrada os dados sobre os sintomas dos pacientes e agrupamos algumas classes como os possíveis tipos de dengue em apenas uma classe, e os outros casos como não relacionados a dengue.

5.1. Estratégia de Treinamento

Para encontrar bons modelos de classificação para os dados, testamos alguns modelos conhecidos sobre variações de configurações. A estratégia de divisão dos dados usada foi a *k-fold cross validation* com 6 *folds*.

Os *folds* gerados foram re-amostrados com o algoritmo *Random Over Sampler*, aplicados apenas nos subconjuntos de trainamento de cada *fold*, dessa forma não há sobreposição de amostras em diferentes *folds*.

5.2. Modelos de Aprendizagem e Configurações

Quatro modelos com estratégias de aprendizagem diferentes foram testados, são eles:

- *Random Forest*: Modelo baseado em múltiplos classificadores de árvore;
- *Multilayer Perceptron*: Grupo das redes neurais;
- *K-Nearest Neighbors*: Baseado no aprendizado por instância;
- *Naive Bayes Gussiano* Gaussiano: Modelo da família dos classificadores probabilísticos.

5.2.1. Random Forest

O modelo *Random Forest* foi testado com as seguintes configurações:

- Configuração 1: Número de estimadores=10, profundidade máxima dos estimadores=ilimitada;
- Configuração 2: Número de estimadores=100, profundidade máxima dos estimadores=4;
- Configuração 3: Número de estimadores=200, profundidade máxima dos estimadores=2;
- Configuração 4: Número de estimadores=500, profundidade máxima dos estimadores=1.

5.2.2. Multilayer Perceptron

Este modelo foi testado com as seguintes configurações:

- Configuração 1: Neurônios escondidos: 10, função de ativação: função unitária linear retificada, treinamento: adam[3], número máximo de épocas: 400;
- Configuração 1: Neurônios escondidos: 40, função de ativação: sigmoide logística, treinamento: adam, número máximo de épocas: 400;
- Configuração 1: Neurônios escondidos: (10, 10)[4], função de ativação: função unitária linear retificada, treinamento: adam, número máximo de épocas: 400;
- Configuração 1: Neurônios escondidos: (40, 10), função de ativação: sigmoide logística, treinamento: adam, número máximo de épocas: 400.

5.2.3. K-Nearest Neightbors

Para este modelo usamos as seguintes configurações:

- Configuração 1: K: 1, ponderação: contagem;
- Configuração 1: K: 3, ponderação: inverso da distância;
- Configuração 1: K: 5, ponderação: inverso da distância;
- Configuração 1: K: 10, ponderação: inverso da distância;

5.2.4. Naive Bayes Gaussiano

Neste modelos não foram usados parâmetros especiais, pois ele possui apenas um parâmetro de rebalanceamento de classes, que nesse caso não é necessário devido a aplicação do algoritmo de *over sampling*.

[3]Variação do Gradiente Descendente Estocástico

[4]Duas camadas de neurônios escondidas

5.3. Medidas

Com a aglutinação das dos tipos de dengue, restaram apenas duas classes, isso permitiu a aplicação de várias medidas mais comuns. As medidas dependentes de classe foram avaliadas para casos positivos de dengue.

As medidas avaliadas foram as seguintes:

- Acurácia;
- Precisão;
- Cobertura;
- F-Measure;
- Matriz de Confusão;
- Curva de Característica de Operação do Receptor (ROC);

5.4. Resultados Obtidos

O modelos de classificação foram representados com siglas para melhor visualização dos dados, são elas:

- *Random Forest*: rfc;
- *Multilayer Perceptron*: mlp;
- *K-Nearest Neighbors*: knn;
- *Naive Bayes Gussiano* gnb.

5.4.1. Acurácia

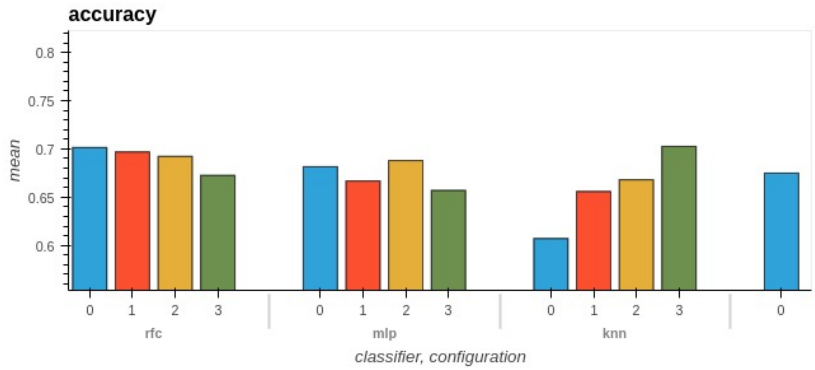

Figura 10: Acurácia

5.4.2. Precisão

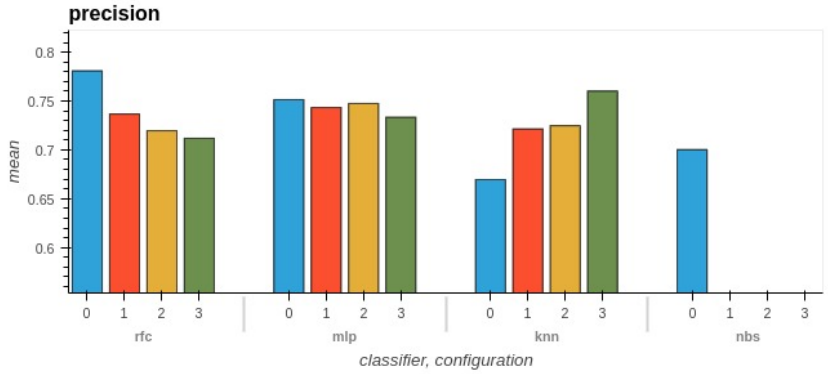

Figura 11: Precisão

5.4.3. Cobertura

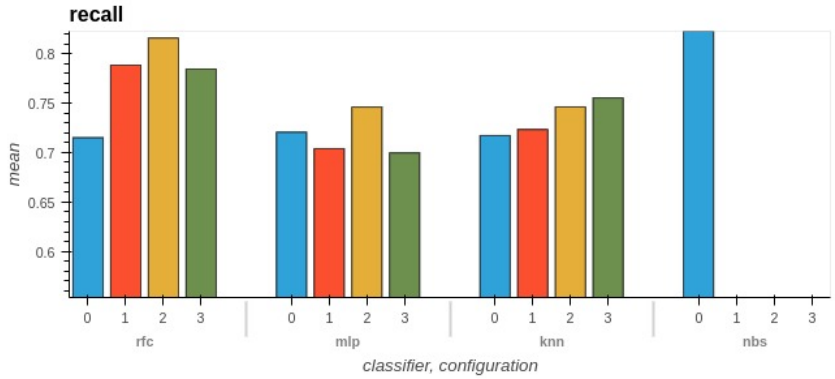

Figura 12: Cobertura

5.4.4. F-Measure

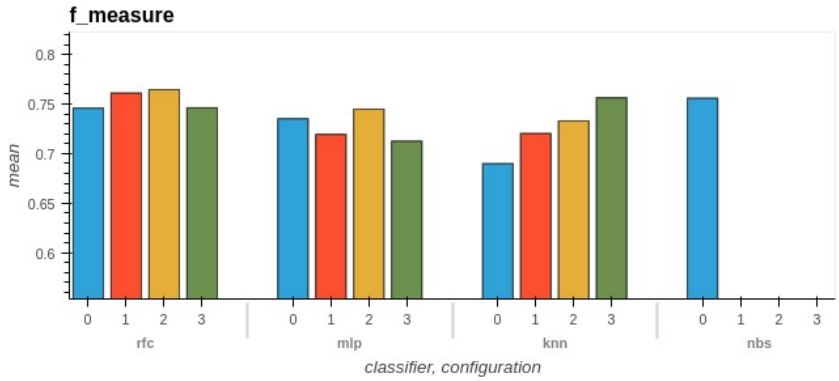

Figura 13: F-Measure

5.4.5. Matrizes de Confusão

Os dados das matrizes de confusão estão normalizados levando em conta a quantidade total dos elementos por classe.

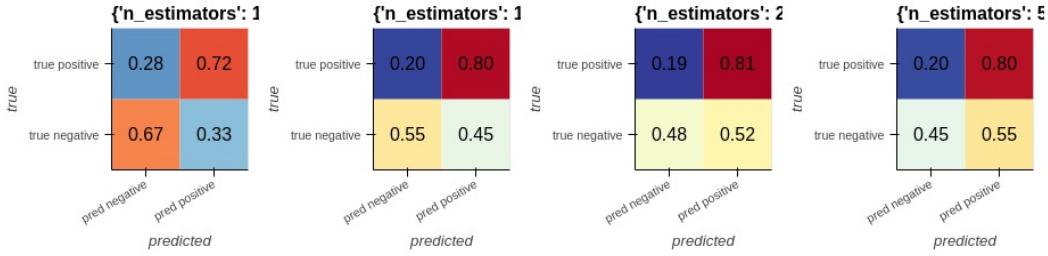

Figura 14: Matrizes de confusão das configurações do *Random Forest*

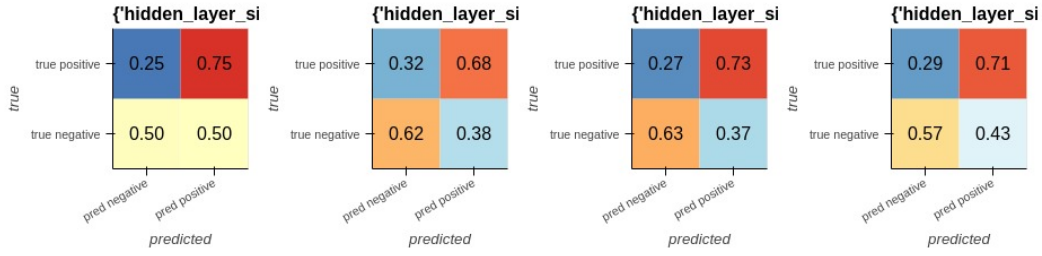

Figura 15: Matrizes de confusão das configurações do *Multilayer Perceptron*

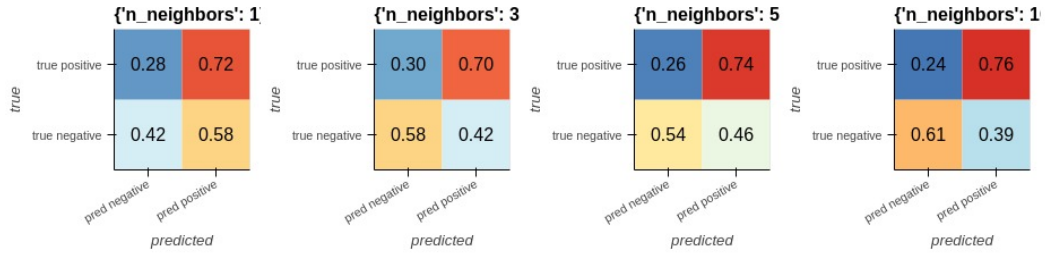

Figura 16: Matrizes de confusão das configurações do *K-Nearest Neighbors*

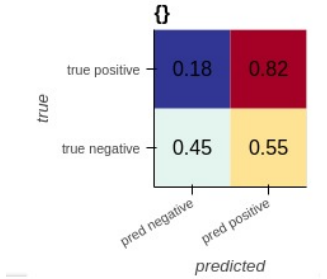

Figura 17: Matriz de confusão da configuração do *Naive Bayes Gaussiano*

5.4.6. Curvas ROC

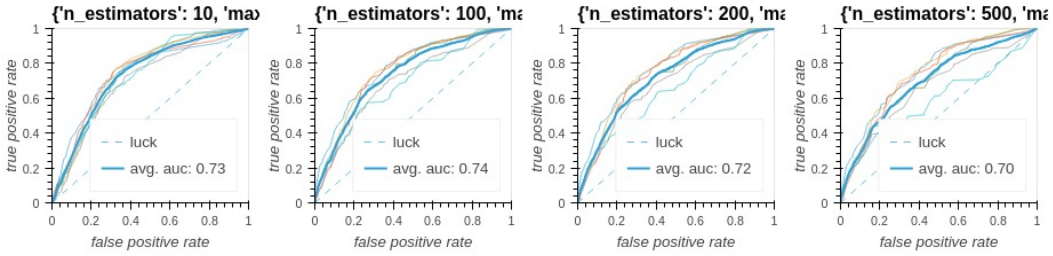

Figura 18: Curvas ROC das configurações do *Random Forest*

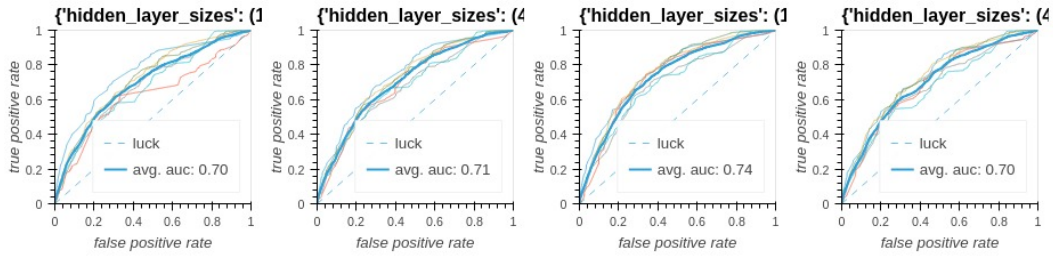

Figura 19: Curvas ROC das configurações do *Multilayer Perceptron*

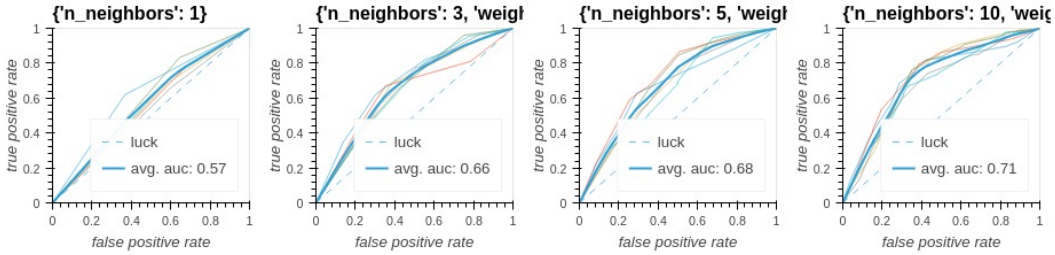

Figura 20: Curvas ROC das configurações do *K-Nearest Neightbors*

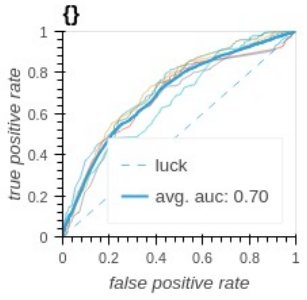

Figura 21: Curvas ROC configuração do *Naive Bayes Gaussiano*

6. Conclusão

Após a análise exploratória dos dados e testes com vários modelos de aprendizagem e configurações, obtivemos valores muito aproximados para a maioria das medidas, com pequenos destaques. Isso é causado pela quantidade de testes feita, pois devido ao tempo curto do projeto, não foi possível avaliar mais modelos de aprendizagem e configurações.

Os resultados dos classificadores podem ser uteis em muitos casos, pois a maioria da população não faz exames sorológicos, e quando fazem, há uma espera de 2 a 3 dias para obter o resultado, que é em alguns casos inconclusivo.

Referências