Relatório MC853 - Classificação de casos de Leptospirose

André Santos Rocha - 235887
Mariano Cho - 230797
Pedro da Rosa Pinheiro - 231081

Introdução

A leptospirose é uma infecção causada por bactérias frequentemente presentes na urina de animais contaminados. Quando um ser humano entra em contato com uma amostra contaminada desse fluido, há o perigo da bactéria entrar no corpo, seja através de feridas pré-existentes, mucosas ou outras áreas vulneráveis.

Sabe-se atualmente que a bactéria pode utilizar múltiplos tipos de mamíferos como hospedeiros, mas o principal fator de contaminação é a urina de ratos. Portanto, áreas com menor infraestrutura para descarte de lixo e condições piores de saneamento básico tendem a atrair e concentrar uma quantidade maior desses roedores. Quando somado à suscetibilidade à enchentes, tais áreas tornam-se focos de contaminação com a bactéria *letpospiras*, causadora da doença.

As melhores formas de evitar tais surtos é descartar corretamente o lixo e introduzir um grau de planejamento urbano mínimo suficiente para garantir saneamento básico e vazão correta de água em épocas de muita chuva, evitando enchentes. No entanto, por fatores socioeconômicos, tal descrição não faz parte da realidade de diversas áreas urbanas brasileiras, contribuindo diretamente para a contaminação, na maioria dos casos, de pessoas em condições socioeconômicas menos favorecidas. De acordo com o Instituto Butantan, a doença possui uma taxa de letalidade média de 10% no Brasil, permitindo reconhecer-se a importância de investir esforços em estudos que decompõem a doença e seus fatores e ajudam na conscientização relativa à doença.

Este projeto surge, portanto, como uma tentativa de estudar e avaliar fatores qualitativos e quantitativos em relação à leptospirose através do uso de *machine learning*. Considerando dados obtidos no DataSus, departamento governamental responsável por coletar, processar e custodiar dados de saúde pública brasileira, aqui é descrito, a princípio, toda a etapa relativa ao tratamento dos múltiplos *datasets* selecionados, abordando tópicos como pré-processamento dos dados.

Propõe-se, juntamente, um conjunto de modelos construídos com base em técnicas de aprendizado de máquina que visam prever o estado final de um indivíduo infectado pela bactéria, isto é, se seu caso termina em alta ou óbito. Utilizando as técnicas *K-Nearest-Neighbors, Random Forest* e *Regressão Logística*, implementou-se também os algoritmos *K-Fold* e *Grid Search* como ferramentas de auxílio no treinamento. Reconhecendo, porém, o caráter moral e social do tópico em análise, é de suma importância também avaliar os efeitos das abordagens e programas aqui relatados. Portanto, visando engrandecer a discussão e aproximá-la de um contexto real, discorreu-se também sobre como os modelos elaborados podem ter sido influenciados por vieses que, inevitavelmente, são inerentes aos dados, membros do grupo e abordagens. Avaliou-se,

também, fatores sociais dos pacientes e sua influência no desenvolvimento da doença, analisando quesitos como localização à nível de regiões brasileiras, idade e sexo.

Entrega 1 - Banco de dados e Pré-processamento

Tratamento de dados

Os dados foram obtidos através da plataforma do *DataSus*, onde foram selecionados os conjuntos referentes aos anos de 2023 e 2024, representando um total de 9153 amostras. No entanto, com um grande desbalanceamento entre as classes, vide, por exemplo, os casos de óbito representados por apenas 2% dos dados, estabeleceu-se como uma das primeiras tarefas balancear o conjunto. Juntamente, sem nenhum tratamento anterior, o *dataset* era composto, em parte, por colunas que não apresentavam uma grande relação com a tarefa alvo que se buscava e instâncias de casos incompletas ou erroneamente preenchidas. Tornou-se necessário, então, analisar e remover as *features* que, a priori, possuiríam menor efeito no modelo a ser construído e lidar com as amostras inconsistentes.

A princípio, foram excluídas diversas *features* não relacionadas à doença e ao seu desenvolvimento. Excluiu-se colunas

a) de identificação, referentes, por exemplo, aos pacientes e aos hospitais, pois em nada influenciam o desenvolvimento da doença;
b) referentes a datas, como data de nascimento e data de notificação da doença, pois os falham em contribuir à análise;
c) relacionadas a sintomas que não se aplicam à casos gerais da doença, seguindo a cartilha do Ministério da Saúde associada à leptospirose;
d) relacionadas aos processos administrativos de registro da notificação;
e) relacionadas a situações de risco às quais o paciente pode ter sido submetido antes do atendimento, dado que o local de contaminação não possui influência na evolução do quadro;
f) relacionadas a fatores socioeconômicos como, por exemplo, raça, escolaridade e renda;
    i) Importante ressaltar que tais fatores são relevantes para a predição de óbito ou não. Porém, idealmente essas questões não deveriam refletir em uma mudança no diagnóstico. Portanto, visando criar um modelo o mais objetivo possível, minimizando efeitos de uma desigualdade socioeconômica existente, o escolhido foi remover tais colunas.
    ii) Válido destacar também que, tal extração pode ocasionar uma piora nos valores finais das métricas de avaliação dos modelos construídos, já que estas colunas representam fatores influentes reais. No entanto, tal deterioramento da capacidade de predição é importante para impedir os modelos de se basearem ou até reforçarem conceitos falaciosos da sociedade.

g) referentes a exames laboratoriais, vide mais da metade de suas instâncias serem nulas. Assim, pouca diferença faria considerá-las e estratégias como preenchimento pela moda não gerariam resultados válidos estatisticamente.

Colunas escolhidas

Foi preciso juntamente analisar as *features* presentes no *dataset* e decidir quais seriam as mais adequadas para a construção dos modelos. Segue abaixo um breve comentário sobre cada uma das colunas escolhidas.

1) **NU_IDADE_N**

Referente à idade do paciente no momento em que surgiram os primeiros sintomas. Os dados podem estar em anos, meses, dias ou horas.

2) **CS_SEXO**

Referente ao sexo do paciente. Os valores apresentados são: Masculino, Feminino ou Ignorado.

3) **CLI_FEBRE**

Referente à presença ou não de sintomas de febre. Os valores apresentados são: Sim, Não ou Ignorado.

4) **CLI_MIALGI**

Referente à presença ou não de sintomas de mialgia. Os valores apresentados são: Sim, Não ou Ignorado.

5) **CLI_CEFALE**

Referente à presença ou não de sintomas de cefaleia. Os valores apresentados são: Sim, Não ou Ignorado.

6) **CLI_ PROST**

Referente à presença ou não de sintomas de prostração. Os valores apresentados são: Sim, Não ou Ignorado.

7) **CLI_CONGES**

Referente à presença ou não de sintomas de congestão conjuntival. Os valores apresentados são: Sim, Não ou Ignorado.

8) **CLI_VOMITO**

Referente à presença ou não de sintomas de vômito. Os valores apresentados são: Sim, Não ou Ignorado.

9) **CLI_DIARRE**

Referente à presença ou não de sintomas de diarreia. Os valores apresentados são: Sim, Não ou Ignorado.

10) **CLI_ICTERI**

Referente à presença ou não de sintomas de icterícia. Os valores apresentados são: Sim, Não ou Ignorado.

11) CLI_RENAL

Referente à presença ou não de sintomas de insuficiência renal. Os valores apresentados são: Sim, Não ou Ignorado.

12) CLI_RESPIR

Referente à presença ou não de sintomas de alterações respiratórias. Os valores apresentados são: Sim, Não ou Ignorado.

13) CLI_CARDIA

Referente à presença ou não de sintomas de alterações cardíacas. Os valores apresentados são: Sim, Não ou Ignorado.

14) CLI_MENING

Referente à presença ou não de sintomas de meningismo. Os valores apresentados são: Sim, Não ou Ignorado.

15) ATE_UF

Referente à Unidade Federativa onde o paciente foi internado. Os valores apresentados são as siglas das Unidades Federativas brasileiras.

16) EVOLUCAO

Referente à presença ou não de sintomas de febre. Os valores apresentados são: Cura, Óbito por leptospirose, Óbito por outras causas ou Ignorado.

Remoção e tratamento das colunas

Primeiramente, analisando a coluna **EVOLUCAO (17)**, removeu-se todas as instâncias que possuíam valor nulo, pois dados sem *label* não agregam ao processo de aprendizado supervisionado, e, portanto, não seriam úteis. Em seguida, foram retiradas as *labels* de valor 3, representativas de óbito por outras causas, já que distanciam-se do objetivo original de descobrir se o paciente será curado ou não; e de valor 9, associadas à "Ignorado", que, assim como os casos nulos, não são úteis. Analisando os valores associados à *feature* **Unidade Federativa (16)** viu-se que possuía uma quantidade pequena de casos nulos, que, portanto, também foram removidos.

Retirou-se também a coluna de **Tipo de Notificação (TP_NOT)**, já que, embora teoricamente útil, 100% dos seus dados possuíam valor 2 (Notificação individual), fato esse que faz com que essa *feature* não agregue em nada ao modelo. Incluir **SG_UF_NOT** também não faria sentido, já que representaria informações redundantes, dado que a unidade federativa já era informada pela coluna **16**.

Analisando a coluna que indica se a pessoa é **gestante** ou **não (CS_GESTANT)**, descobriu-se que apenas duas instâncias que resultaram em óbito eram classificadas como grávidas, classificadas então como *outliers.* Com isso, a coluna era homogênea, com todas as amostras classificadas como "não gestantes", justificando, portanto, a remoção da *feature*. Na mesma lógica, a coluna **ATE_HOSP**, que indica se o paciente foi hospitalizado

ou não, possuía apenas duas instâncias associadas ao segundo caso, justificando sua remoção.

No que se refere à coluna de **idade (1)**, foi preciso realizar um tratamento de dados. Os valores estavam codificados com quatro dígitos, sendo o primeiro dígito, 1, 2, 3 ou 4, indicando se a idade estava representada em hora, dia, mês ou ano, respectivamente. Converteu-se, portanto, todas as idades para um único valor numérico, considerando anos como unidade de medida.

Para formatar a coluna da **Unidade Federativa (16)**, até então tida como *String*, foi utilizada a técnica de *"one hot encoding"*, já que uma codificação numérica, introduziria uma relação de cardinalidade entre elas. Assim, foram criadas colunas mais amplas, englobando os estados conforme sua **região**, resultando em cinco novas colunas: *regiao_Norte*, *regiao_Nordeste*, *regiao_Centro-Oeste*, *regiao_Sudeste* e *regiao_Sul*.

A coluna de **Sexo (2)**, bem como as referentes aos sintomas (do formato **CLI_XXX**) foram tratadas com *label encoding*. Na primeira, 0 foi atribuído ao sexo feminino e 1 ao sexo masculino. Nas outras mencionadas, 0 aos casos com ausência do sintoma e 1 nos casos com sua presença. Além disso, substituiu-se os valores nulos e "Ignorados" dessas classes pela sua moda, processo esse válido de ser realizado já que representavam no máximo 5% das instâncias totais para cada sintoma.

Por fim, conforme a proposta do projeto, dividiu-se o *dataset*, já tratado, em bases de treino e de teste, considerando um critério geográfico. Observando a distribuição original dos dados, percebeu-se que, embora com menos estados, a região Sudeste possuía uma quantidade um pouco menor de instâncias do que a soma de todas as outras. Optou-se, portanto, por separar as instâncias dessa primeira região para teste e o restante para treino.

Entretanto, as bases estavam desbalanceadas, com o óbito representando apenas 5% das amostras. Optou-se, portanto, por remover aleatoriamente instâncias representativas de casos de cura até que os dados associados à "cura" e "óbito" fossem equivalentes a 75% e 25% dos dados, respectivamente, em ambos *datasets*.

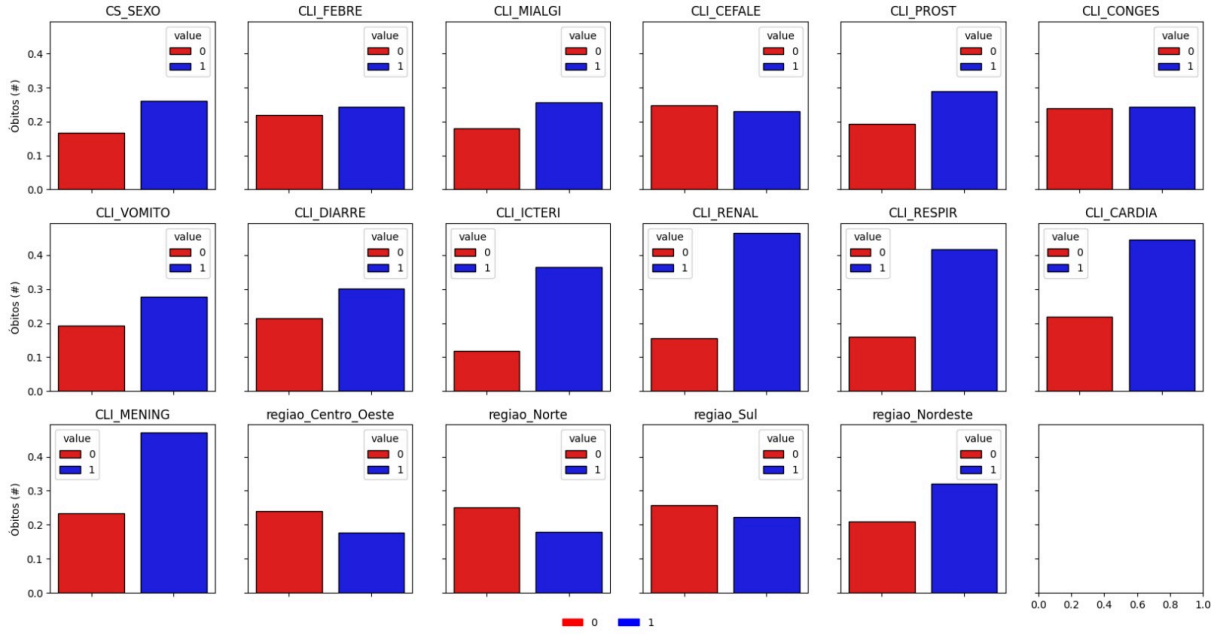

Figura 1: Distribuição dos dados de acordo com cada *feature*

Entrega 2 - Construção e análise dos modelos

Na segunda etapa do projeto, 3 modelos de aprendizado de máquina foram treinados e avaliados, a fim de escolher o que obtivesse o melhor desempenho na nossa tarefa de predição de óbito. O que alcançasse o melhor resultado de acordo com a métrica a ser utilizada seria o selecionado. Com os dados de treino e teste importados, foram definidos os classificadores a serem treinados: KNN, *Random Forest* e *Logistic Regression*.

Uma análise importante realizada também neste momento está relacionada à visualização da correlação entre os dados do nosso *dataset*. Vemos, pela Tabela 1, proposta logo abaixo, que, embora algumas variáveis aparentam ter relações menos diretas, isto é, valor *p* próximo de 0, com a nossa *feature* de predição "*EVOLUCAO*", alguns sintomas revelam-se ainda menos atuantes no estado final do paciente. **Sintomas respiratórios (12)** e **sintomas renais (11)**, por exemplo, apresentam correlação próxima à -0.3, representando, numa análise imediata, um distanciamento entre um paciente falecer e tais sintomas. Válido ressaltar, porém, que a correlação direta individual entre variáveis e a *feature* alvo da predição não é suficiente para se chegar em algum tipo de conclusão. É necessário também avaliar a relação entre a variável resposta e combinações de conjuntos das variáveis explicativas.

Coluna	Correlação
NU_IDADE_N	-0,205755
CS_SEXO	-0,085674
CLI_FEBRE	-0,029928
CLI_MIALGI	-0,080563
CLI_CEFALE	0,015946
CLI_PROST	-0,109684
CLI_CONGES	-0,000272
CLI_VOMITO	-0,105879
CLI_DIARRE	-0,099776
CLI_ICTERI	-0,287506
CLI_RENAL	-0,310248
CLI_RESPIR	-0,283925
CLI_CARDIA	-0,122663
CLI_MENING	-0,072052
ATE_HOSP	-0,025036

regiao_Sul	0,045418
Regiao_Centro-Oeste	0,027033
regiao_Nordeste	-0,111205
regiao_Norte	0,057107

Tabela 1: Valor-p para correlação entre *features* do conjunto de teste e a *feature* meta

Antes de começar o treinamento, foi definido o método *balance* para manter a proporção de 1,5:1 entre as classes, com a intenção de reduzir o desbalanceamento sem comprometer a qualidade estatística dos dados ou aumentar o risco de *overfitting*. As técnicas utilizadas nesse processo foram: *smote* e *near miss*.

O *smote* é uma técnica de *oversampling* utilizada para aumentar a classe minoritária em 20% por meio da geração de instâncias sintéticas. Ele seleciona uma amostra da classe minoritária, identifica seus vizinhos mais próximos com *KNN*, calcula a distância entre eles e gera novas amostras em pontos aleatórios ao longo dessas distâncias, promovendo diversidade e evitando duplicações. Já o *near miss* é uma técnica de undersampling que utilizada para reduzir a classe majoritária até que ela represente 60% do total, selecionando apenas as amostras mais próximas da classe minoritária — aquelas mais difíceis de classificar e próximas à fronteira de decisão — para manter o modelo focado nas regiões mais desafiadoras do espaço de decisão.

Já na etapa de treinamento dos modelos, utilizou-se a função *StratifiedKFold* para dividir os dados em 5 subconjuntos, preservando a proporção das classes em cada partição. Após essa divisão, construiu-se um modelo para cada *fold*, de modo a fazer uso de todo o *dataset*, mas com base em subdivisões. Para isso, aplicou-se a função *balance* aos dados do subconjunto em questão, com o objetivo de mitigar possíveis desequilíbrios entre as classes. Em seguida, empregou-se o algoritmo *GridSearchCV* para buscar os melhores hiperparâmetros, utilizando validação cruzada interna com k=5 para garantir uma estimativa robusta do desempenho dos modelos, parâmetros esses que foram utilizados em seguida para o treinamento. Os hiperparâmetros avaliados podem ser visualizados abaixo:

Parâmetro	**Valores avaliados**
n_neighbors	[1,5,10]
p	[1,2]
weights	[‘uniform’, ‘distance’]

Tabela 2: Parâmetros avaliados para KNN

Parâmetro	**Valores avaliados**
class_weight	‘balanced’, {0:1, 1:2},

	{0:1, 1:3}

Tabela 3: Parâmetros avaliados para Regressão Logística

Parâmetro	**Valores avaliados**
n_estimators	[10, 100, 200]
max_depth	[10, 50]
min_samples_split	[2, 10, 30]

Tabela 4: Parâmetros avaliados para *Random Forest*

Desse modo, com 3 algoritmos diferentes e 5 *folds*, foram construídos 15 modelos diferentes, cada um com seu respectivo desempenho baseado em *f1-score*, desempenho esse que pode ser visualizado pela Tabela 2.

Originalmente, o critério de escolha do melhor modelo, tanto para a definição dos hiperparâmetros quanto para a definição do modelo final, era o seu nível de *recall*. No entanto, viu-se após a construção, treinamento e teste dos primeiros modelos que, ao invés de aprenderem a tarefa em questão, estavam classificando a maioria das amostras como finalizadas em óbito. Dessa forma, acertava a maior parte dos casos positivos, comportamento esse que pode ser visualizado pela Figura 2. Mas ao se analisar seu desempenho em relação à, por exemplo, precisão, percebeu-se que os modelos apresentavam um desempenho não satisfatório, já que a maior parte dos casos preditos como “positivos”, isto é, óbito, não eram. Buscando minimizar tal problema e tentar evitar que os modelos assumissem uma estratégia dominante caracterizada por predição de óbito constante, o grupo alterou sua métrica de performance para *f1-score*. Essa, por definição, combina precisão e *recall* em um único valor, possibilitando um comportamento mais equilibrado dos modelos.

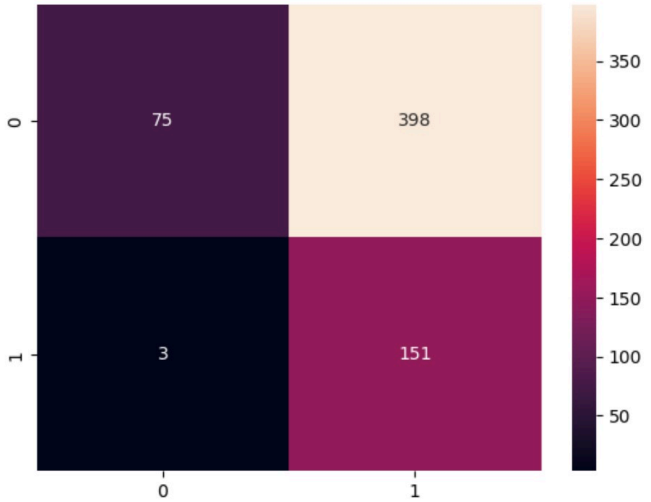

Figura 2: Matriz de confusão com modelo baseado em *recall*

Juntamente, realizou-se o cálculo dos prováveis melhores hiperparâmetros, permitindo o início do processo de treinamento dos modelos. A princípio, na primeira etapa, a quinta regressão logística (*Logistic Regression_4*) apresentou o melhor desempenho no que se refere à *recall*, alcançando um valor de aproximadamente 94%. No entanto, como

proposto anteriormente, a métrica alvo de performance foi, posteriormente, alterada. Então todo o processo de cálculo dos hiperparâmetros, em conjunto com o treinamento e teste dos modelos, foi refeito. No final, o selecionado foi a quarta versão de regressão logística (*Logistic Regression_3* na tabela abaixo), com 54,5% de *f1-score*. Para alcançar tal valor, utilizaram-se os hiperparâmetros *class_weight*={0:1, 1:2}. Na tabela 5 abaixo é possível visualizar o desempenho de todos os modelos treinados.

Modelo	**Acurácia Balanceada**	**Precisão**	***Recall***	***f1-score***
KNN_0	0,631	0,429	0,447	0,438
KNN_1	0,578	0,328	0,426	0,370
KNN_2	0,601	0,370	0,426	0,396
KNN_3	0,640	0,449	0,458	0,454
KNN_4	0,595	0,391	0,375	0,383
Random Forest_0	0,685	0,466	0,574	0,514
Random Forest_1	0,652	0,397	0,574	0,470
Random Forest_2	0,643	0,414	0,511	0,457
Random Forest_3	0,672	0,450	0,563	0,500
Random Forest_4	0,634	0,418	0,479	0,447
Logistic Regression_0	0,653	0,323	0,872	0,471
Logistic Regression_1	0,692	0,416	0,681	0,516
Logistic Regression_2	0,660	0,363	0,702	0,478
Logistic Regression_3	0,716	0,429	0,750	0,545
Logistic Regression_4	0,680	0,379	0,750	0,503

Tabela 5: Desempenho de todos os modelos treinados

Posteriormente, os modelos foram avaliados no *dataset* de testes, obtendo aproximadamente 70,1% de acurácia balanceada, 39,0% de precisão e 81,8% de *recall*.

Com a construção da matriz de confusão, visualizada pela Figura 3, obteve-se uma visão mais detalhada dos resultados. Embora ainda esteja predizendo 1, isto é, óbito, erroneamente com grande frequência, vide os 197 casos de falsos positivos, enxerga-se uma melhora significativa em relação ao modelo anterior baseado em *recall*, observado pela Figura 2. Nesse momento, tem-se um modelo mais balanceado, com maior capacidade de predição correta dos casos que resultam em alta. Além disso, é possível analisar que, embora ainda esteja errando com alta frequência de falsos positivos, tal formato de erro tende a ser priorizado em relação a erros de falsos negativos. Isso, pois no tocante à saúde individual, é de maior importância focar em cuidar mais de pessoas que possivelmente vão receber alta do que deixar de prestar atenção a casos mais sérios.

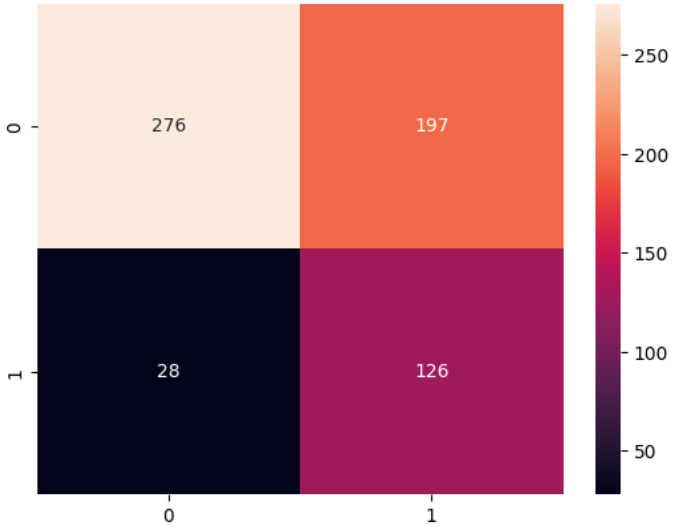

Figura 3: Matriz de confusão do modelo de melhor desempenho

Como conclusão, dada a natureza crítica do problema proposto, enxerga-se que o modelo construído até então não é ideal para uso amplo, apresentando dificuldades em uma tarefa tão sensível quanto predizer o resultado do desenvolvimento de uma doença infecciosa como a leptospirose. No entanto, percebe-se uma evolução desde o início das propostas implementadas, reflexo de ajustes ao longo de diferentes etapas da análise e do projeto.

Reconhece-se, também, que o conjunto de dados selecionados como alicerce para o treinamento e teste dos modelos apresentava erros que dificultaram a tarefa desde seu início. Porém, através do pré-processamento e análise feitos na primeira etapa, foi possível se distanciar um pouco dos problemas que eles introduziram.

Além disso, analisar um modelo baseando-se em apenas uma métrica pode disfarçar problemas internos, vide, por exemplo, como uma revocação de 81,8% esconde um comportamento do modelo de priorizar predições de óbito para a maioria dos casos. No entanto, percebe-se que é exatamente por essa análise que foi e ainda é possível adaptar cada vez mais o modelo para o problema. Como concretização dessa ideia, vê-se a evolução da acurácia e da precisão entre as duas gerações de modelos propostos, métricas essas que passaram de 57% e 27% na primeira tentativa para 70,1% e 39%, respectivamente, na aqui proposta.

É de suma importância, portanto, seguir estudando e explorando o modelo proposto, de modo a buscar resultados satisfatórios dado o problema original. Por mais que fora do

contexto e possibilidade de atuação do projeto, é possível afirmar que o processo seria mais proveitoso caso sujeito a conjuntos maiores e mais diversos de dados, capazes de prover informações sobre casos mais genéricos e complexos simultaneamente.

Entrega 3 - *Fairness*

Exploração das métricas

Inicialmente, escolheu-se a métrica oportunidade igualitária para avaliar equidade na variável “sexo”. Baseada nesta variável, conduzimos uma análise exploratória para entender a métrica de equidade mais pertinente ao nosso modelo. As métricas sugeridas para uso foram: paridade demográfica, oportunidade igualitária e probabilidades equalizadas.

A paridade demográfica é uma métrica que assegura proporção igualitária de previsões positivas entre os grupos. A oportunidade igualitária é uma métrica que visa equalizar as taxas de verdadeiros positivos entre os diferentes grupos ou, no nosso caso, entre homens e mulheres. A probabilidade equalizada é uma métrica que garante que as taxas de verdadeiros positivos e de falsos negativos sejam iguais entre os grupos.

Sendo assim, o procedimento adotado foi separar dois conjuntos de dados: um formado apenas por instâncias do sexo masculino e outro formado apenas por instâncias do sexo feminino. Utilizou-se o nosso melhor modelo para predizer a evolução do caso em cada um desses conjuntos. Ao avaliar métricas relevantes para a detecção de equidade nos grupos, obteve-se os seguintes resultados:

Grupo	**Taxa de falsos positivos**	**Precision**	**Recall**	**Instâncias classificadas como positivas**
Homens	0.421	0.417	0.82	0.528
Mulheres	0.403	0.296	0.808	0.473
Diferença absoluta	**0.018**	**0.121**	**0,012**	**0.055**

Tabela 6: Métricas para conjuntos de dados de homens e mulheres antes do *oversampling*

Ao analisá-los, nota-se que a maior diferença entre os modelos consiste na diferença entre as precisões. Diante disso, a métrica com a qual optamos trabalhar foi a probabilidade equalizada, pois pode contribuir para a redução da discrepância entre as precisões na medida em que reduza a diferença entre as taxas de falsos positivos.

Ademais, o estudo das matrizes de confusão de cada grupo também aponta na direção da probabilidade equalizada enquanto métrica adequada de *fairness*. É possível perceber, por exemplo, na Figura 4 que o *recall,* métrica mais próxima ao grupo masculino, custou ao modelo a predição de, aproximadamente, 47% do *dataset* como óbito. Enquanto isso, na Figura 5, nota-se que o modelo no grupo masculino, apesar de predizer 53% das instâncias como positivas, é mais preciso.

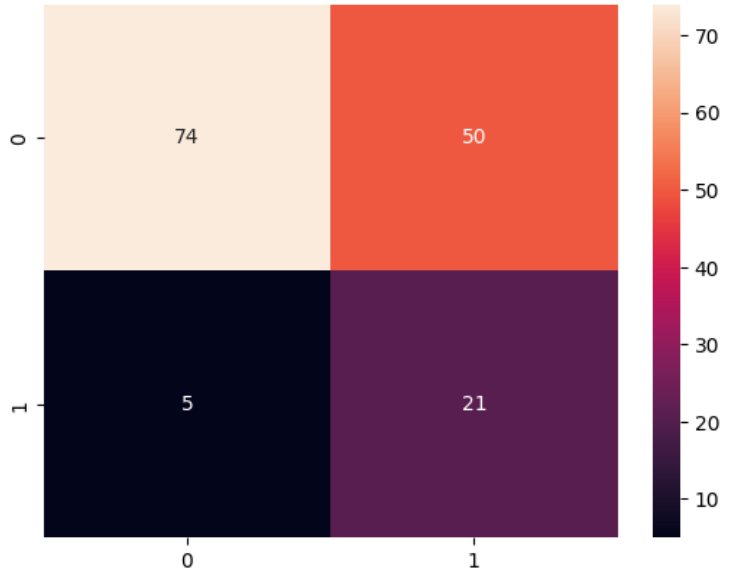

Figura 4: Matriz de confusão para o conjunto de dados formado por mulheres

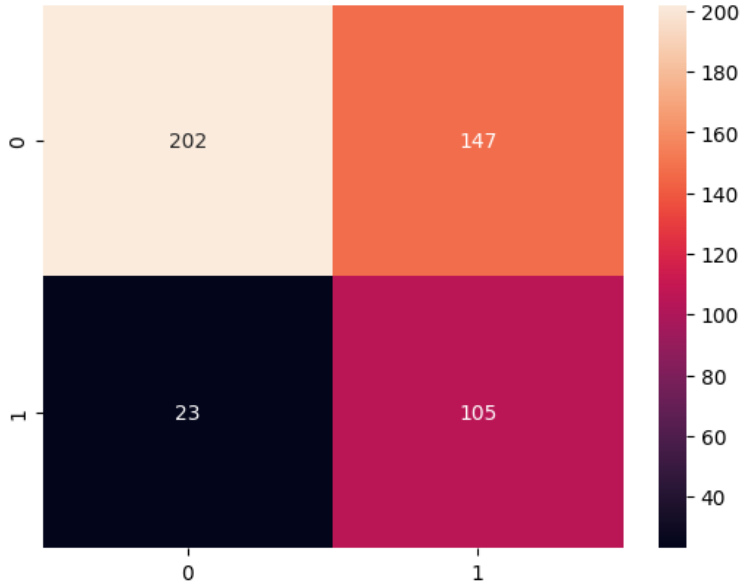

Figura 5: Matriz de confusão para o conjunto de dados formado por homens

Portanto, acredita-se que a probabilidade equalizada é a melhor métrica para avaliar equidade no conjunto estudado, pois, assim, equilibrar-se-ão os verdadeiros positivos e também os falsos positivos, amenizando a discrepante diferença de precisão (12 pontos percentuais) do nosso modelo para lidar com mulheres.

Implementação de estratégia para *fairness*

Para alcançar a equidade no nosso modelo, adotou-se uma estratégia de *oversampling.* No conjunto de dados original, havia a seguinte proporção entre homens e mulheres:

Grupo	Quantidade
Homens	756
Mulheres	239

Tabela 7: Distribuição de homens e mulheres antes de *oversampling*

Aplicou-se, então, a técnica *SMOTE* para *oversampling,* de modo que a classe minoritária (mulheres) atingisse 50%, 75% e 99% da classe majoritária (homens). Com isso, realizou-se um novo treinamento, obtendo-se os seguintes resultados:

Modelo	Oversample	Acurácia Balanceada	Precisão	*Recall*	*f1-score*
Logistic Regression 3	50%	0,686	0,358	0,896	0,511
Logistic Regression 3	75%	0,672	0,345	0,903	0,499
Logistic Regression 3	99%	0,615	0,299	**0,974**	0,457
Logistic Regression 3	Antes do Oversampling	**0,716**	**0,429**	0,750	**0,545**

Tabela 8: Resultados do treinamento após *oversampling*

Percebe-se que o oversampling prejudica o desempenho dos modelos em grande parte das métricas. Isso se dá, em grande parte, devido à adição de ruído e alteração da distribuição da variável-alvo, já que o *oversampling* foi feito com base na variável “sexo”.

Para avaliar a equidade, testou-se o modelo para homens e mulheres separadamente, obtendo os seguintes valores:

Modelo	Estratégia	Grupo	Taxa de falsos positivos	Precisão	Recall	Instâncias classificadas como positivas
Logistic Regression 3	50%	Homens	0,540	0,264	0,923	0,607
		Mulheres	0,519	0,386	0,891	0,618
		Diferença Absoluta	**0,022**	**0,123**	**0,032**	**0,012**
Logistic Regression 3	75%	Homens	0,579	0,271	0,885	0,567
		Mulheres	0,519	0,365	0,906	0,667

		Diferença Absoluta	**0,060**	**0,094**	**0,022**	**0,100**
Logistic Regression 3	99%	Homens	0,710	0,228	0,100	0,760
		Mulheres	0,756	0,320	0,097	0,813
		Diferença Absoluta	**0,047**	**0,092**	**0,003**	**0,053**
Logistic Regression 3	Antes do oversampling	Homens	0.421	0.417	0.82	0.528
		Mulheres	0.403	0.296	0.808	0.473
		Diferença absoluta	**0.018**	**0.121**	**0,012**	**0.055**

Tabela 9: Resultados de comparação do modelo antes e a após estratégia de *oversampling*

Ao observar a Tabela 9, nota-se que o ganho de equidade não foi atingido, já que não foi possível atenuar as somas das diferenças entre os *recalls* e as taxas de falsos positivos. Acredita-se que a introdução dos novos dados artificiais não foi suficiente para tornar o modelo mais justo devido à provável adição de dados ruidosos. Uma possibilidade a ser explorada futuramente poderia ser a coleta de dados reais para aumentar a quantidade de dados femininos.

Vale ressaltar que se a métrica zhouvesse sido a paridade demográfica, os modelos treinados com o*versampling* de 99% e 50%, haveriam ambos progredido em equidade. Ademais, se a métrica fosse oportunidades igualitárias, o modelo com *oversampling* 99% haveria progredido também.

Assim, cumpre-se o propósito inicial de avaliar a equidade do modelo sob a ótica das probabilidades equalizadas, bem como a estratégia de *oversampling* enquanto técnica para atingir equidade. Concluímos que a técnica pode ser limitada em casos com poucos dados, mas que é preciso seguir buscando alternativas para garantir equidade nos modelos de *machine learning*.

Considerações Éticas

O projeto aqui proposto discorre sobre a predição da conclusão de um caso de leptospirose em indivíduos brasileiros, isto é, se, dada uma pessoa infectada, essa receberá alta ou irá à óbito. No entanto, é fundamental ressaltar que, embora os modelos responsáveis por essa classificação sejam construídos baseados em dados reais, não se deve levar suas inferências como veredito para tomada de decisões médicas.

O estudo e aplicação de *machine learning* em áreas da saúde vem crescendo aceleradamente, abrangendo tópicos desde automatização de tarefas administrativas dentro de ambientes hospitalares quanto suporte à diagnósticos e tratamentos. No entanto, tarefas adjacentes ou diretamente relacionadas à gestão e cuidado da saúde humana possuem necessidades intrínsecas à relações interpessoais, características essas naturais de um domínio diferente daquele de atuação do aprendizado de máquina.

Nesse sentido, vê-se que a ferramenta aqui proposta, embora dedique-se a contribuir para o estudo de diagnósticos, é uma tentativa de modelagem que ainda sim

dispõe de um grau de imprecisão. Assim, tanto os modelos desenvolvidos neste trabalho quanto quaisquer outros associados ao tópico em questão devem ser sempre utilizados enquanto acompanhados de especialistas e servem apenas como um norte no combate às doenças. Basear ações práticas para com outro indivíduo utilizando as predições dos modelos como argumento não apenas é, atualmente, falacioso como também perigoso, já que substitui o caráter humano necessário para tais ações e assumem um risco com a vida alheia.

Além disso, por mais que já proposto, é imprescindível retomar a ideia por trás do abandono de algumas das *features* associadas à características sociais dos pacientes. Colunas referentes à, por exemplo, raça, escolaridade e renda, foram removidas do *dataset* aqui utilizado. A ideia que sustenta essa decisão vem da concepção de que tais fatores distanciam o modelo de um julgamento objetivo, alvo principal deste estudo. É falacioso afirmar que essas *features* não possuem efeito nos casos e amostras, já que, tratando-se do contexto brasileiro, o alto grau de desigualdade social implica em desigualdade no tratamento, desigualdade no acesso à medicamentos e exames e desigualdade geral no contexto de vida, resultando em contato maior ou menor com situações de risco à doença. No entanto, em um cenário ideal, esses fatores não deveriam influenciar no resultado do diagnóstico ou sequer estarem presentes. Então essa remoção foi feita objetivando tirar a influência dessas características dos modelos, de forma a evitar que essas ferramentas pudessem considerar que, por exemplo, um indivíduo, por ser de uma camada socioeconômica mais baixa, estaria automaticamente mais sujeito à óbito.