Interpretabilidade

Bruna Almeida Osti* e Fabio Fogliarini Brolesi†
Instituto de Computação, Universidade Estadual de Campinas - UNICAMP
Campinas, SP
Email: *b231024@dac.unicamp.br, †brolesi@gmail.com

***Resumo*—Este trabalho busca abordar a interpretabilidade de modelos, utilizando diversas abordagens de pré-processamento, in-processamento e pós-processamento.**

***Index Terms*—Trustworthy, AI, Ethical**

I. INTRODUÇÃO

Os modelos prognósticos usam vários fatores em combinação para prever o risco de resultados clínicos futuros em pacientes. Um bom modelo deve (i) fornecer previsões precisas que informam os pacientes e seus cuidadores, (ii) apoiar a pesquisa clínica e (iii) permitir decisões para melhorar os resultados dos tratamentos aos pacientes, segundo [1]. Um modelo prognóstico tem três fases principais: desenvolvimento do modelo (incluindo validação interna), validação externa e investigações de impacto na prática clínica. Embora muitos modelos prognósticos sejam propostos, poucos são atualmente usados na prática, conforme mostra [1].

Mas há sérios riscos em aprender padrões com exemplos, não é um processo de simplesmente memorização. Envolve generalizar e aprimorar os detalhes que são característicos de uma classe em geral, não apenas indivíduos específicos que aparecem nos exemplos. Este é o processo de indução: extrair regras gerais de exemplos específicos — regras que efetivamente respondem por casos passados, mas também se aplicam a casos futuros, ainda não vistos [2].

O fato de ser baseado em evidências não garante que os modelos levaram a decisões precisas, confiáveis ou justas. Nossos exemplos históricos dos resultados relevantes quase sempre refletirão preconceitos históricos contra certos grupos sociais, esteriótipos culturais predominantes e desigualdades demográficas existentes, e aprender padrões sobre esses dados significa geralmente replicar esses mesmos preconceitos [2].

As abordagens de aprendizado de máquina não são de entendimento trivial para a grande maioria das pessoas, pois possuem formas funcionais complexas. A interpretabilidade e a explicabilidade são formas de interação entre a máquina e o humano, especificamente a comunicação da máquina para o humano, que permite que a máquina e o humano colaborem na tomada de decisões, a partir do entendimento humano dos julgamentos feitos pela máquina e que levaram a determinado resultado. O presente trabalho tem por objetivo apresentar formas de interpretar e comunicar os modelos de modo a trazer maior segurança a quem depende das respostas do mesmo, na medida em que pode-se entender em algum grau o que o modelo está executando de forma menos abstrata.

II. TRABALHOS RELACIONADOS

No trabalho de [3], é oferecido uma proposta de modelo explicável para internações de unidade de terapia intensiva, afim de avaliar os fatores de risco que afetam a mortalidade dos pacientes internados nestas unidades. Os modelos estatísticos de aprendizado de máquina podem ser utilizados para análise e inferência de relações entre as características e os resultados com os pacientes. Ainda assim, esses modelos podem ser difíceis de usar ou seus resultados são difíceis de interpretar em um ambiente clínico, pois muitas vezes exigem condições específicas para serem aplicadas. Ao contrário do modelo "caixa preta", o modelo explicável permite a relação quantitativa entre parâmetros clínicos e previsões de resultados a serem visualizados pelo usuário, o que, segundo [3] permite aos médicos interpretar melhor como as alterações nos parâmetros de risco afetam o resultado.

Também [4], utilizando redes neurais convolucionais, diz que elas aprendem de maneira orientada por dados de e que os dados de eletrocardiogramas podem fornecer informações clínicas valiosas. Desta forma, entendendo que existe valor no resultado da rede neural e também nos resultados clínicos, aplicou-se o a técnica de explicações agnósticas de modelos interpretáveis locais (*LIME*) para destacar quais pontos do exame direcionam diagnósticos específicos.

Já [5] diz que a interpretabilidade de modelos de predição complexos na área de saúde são necessários dada a natureza desta área. Os modelos de aprendizado de máquinas são utilizados na área de saúde, como previsão do risco de doença do paciente, na probabilidade de readmissão do paciente e na previsão da necessidade de cuidados. Para entender e aceitar ou rejeitar essas predições, o usuário final e os profissionais de saúde devem entender o raciocínio por trás do modelos estatísticos. [5] também diz que a falta de interpretabilidade de modelos é um fator que limita a adoção de aprendizado de máquina em saúde.

O trabalho de [6], por sua vez, utilizou a eliminação recursiva de features (recursive feature elimination - RFE) com algoritmos de random forest para classificar variáveis preditoras candidatas com base na importância relativa. O uso de RFE é um método para otimizar a seleção de features dentre um conjunto maior de variáveis candidatas, calculando iterativamente a importância para cada variável. Posteriormente, os preditores menos importantes foram removidos por meio de eliminações até que um subconjunto de 10 preditores que otimizou o desempenho foi identificados para o desenvolvimento

do modelo final.

Já a proposta de [7] usa modelos de inteligência artificial explicáveis existentes em em conjunto com o conhecimento clínico para obter mais benefícios em sistemas baseados em inteligência artificial. A abordagem proposta é a que seguinte:

- Aplicativos de saúde inteligentes capturam as informações de saúde de indivíduos e usam os modelos de inteligência artificial já treinados.
- As predições junto com os dados de saúde são usadas por métodos de inteligência artificial explicável para gerar explicações.
- Essas explicações podem ser analisadas com a ajuda de conhecimento de um clínico e essa análise permitirá validação das predições feitas pelo modelo de inteligência artificial por médicos para permitir a transparência.
- Se as predições estiverem corretas, então as explicações junto com conhecimento clínico pode ser usado para gerar insights e recomendações.
- Se as predições estiverem incorretas, então a contradição entre as explicações e o conhecimento do clínico pode ser usada para rastrear fatores para predições imprecisas e permitir melhoria no modelo de inteligência artificial implantado.

III. METODOLOGIA

A interpretabilidade e a explicabilidade são necessárias para tornar as predições confiáveis, e superar vieses cognitivos no problema de comunicação entre o modelo de aprendizado e o consumidor humano, neste caso, profissionais área da saúde.

Os métodos de explicação podem ser divididos em oito categorias, e subdividas por três dicotomias. A primeira diz respeito se a explicação é para um modelo inteiro (global) ou para um ponto de dados de entrada específico (local). A segunda é se a explicação é uma representação exata do modelo (exata) ou se contém alguma aproximação (aproximada). A terceira diz respeito se a linguagem usada na criação da explicação é baseada nas features (baseada em features) ou em pontos de dados completos (baseado em amostras) [8].

Portanto, foi feito o ajuste do modelo Decision Tree que é um modelo interpretável para verificar o seu funcionamento. Além disso, foi aplicado dois métodos de explicabilidade para modelos não interpretáveis ao modelo de Regressão Logística, o LIME que é um modelo local e aproximado e o Partial Dependence Plot que é global e aproximado.

De mesmo modo, foi explorado o modelo DiCE para criar explicações (amostras) que atinjam a saída esperada pelo modelo.

A. Decision Trees

Conforme indica [9], as árvores de decisão são usadas para classificação e regressão, através da aprendizagem de regras de decisão simples que são inferidas através dos dados. Quanto mais profunda a árvore, mais complexas são as regras de decisão e mais adequado o modelo.

As árvores são muito utilizadas pela sua facilidade de interpretação, além disso é possível validar um modelo usando testes estatísticos, isso torna possível explicar a confiabilidade do modelo. Entretanto, elas podem ser instáveis pois pequenas variações nos dados podem resultar na geração de uma árvore completamente diferente, sendo necessário o uso de um ensemble para mitigar o problema. Além disso, elas não são boas em extrapolação, visto que as predições são aproximações constantes por partes.

Segundo [10], a árvore de decisão particiona recursivamente o espaço de features de forma que as amostras com os mesmos rótulos sejam agrupadas.

A qualidade do candidato é computada usando a função de impureza (função de loss) H, a escolha depende do tipo de task (classificação ou regressão). A medida mais comum utilizada para medir a impureza é a Gini, mas também podemos utilizar a função de entropia.

B. Explicabilidade de modelo não interpretável

1) LIME: Pensando num modelo não interpretável local, [11] introduzem o que chamam de *Local Interpretable Model-Agnostic Explanations (LIME)*. O fato de ser agnósticos em termos de de modelo, o que é feito para aprender o comportamento do modelo subjacente é perturbar a entrada e ver como as predições mudam. Isso é um benefício em termos de interpretabilidade, porque é possível perturbar a entrada alterando componentes que fazem sentido para pessoas mesmo que o modelo esteja usando componentes muito mais complicados como recursos.

No LIME é gerada uma explicação aproximando o modelo subjacente por um interpretável (como um modelo linear com apenas alguns coeficientes diferentes de zero), aprendido em perturbações da instância original (por exemplo, remover palavras ou ocultar partes da imagem). A principal intuição por trás do LIME é que é mais fácil aproximar um modelo de caixa-preta por um modelo simples localmente (na vizinhança da previsão que se quer explicar), em vez de tentar aproximar um modelo globalmente. Isso é feito ponderando as imagens perturbadas por sua similaridade com a instância que se quer explicar.

2) Partial Dependence Plot: Os gráficos de dependência parcial podem ser usados para visualizar e analisar a interação entre a resposta alvo ('`icu_IS DEAD_1`') e um conjunto de features de entrada de interesse. Intuitivamente, podemos interpretar a dependência parcial como a resposta esperada do alvo em função das características de entrada de interesse [12].

Os gráficos mostram o efeito médio nas previsões conforme o valor das alterações das features. Internamente, para o cálculo dos valores de entrada, é encontrado valores únicos de cada feature, e verificado como as previsões individuais mudam (expectativa condicional individual, ICE). Calculando a média dessas previsões (dependência parcial, PD), obtemos as entradas dos gráficos.

C. Contrafactual

1) DiCE: O DiCE fornece explicações contrafactuais mostrando versões perturbadas das features da mesma entrada

que teriam um resultado diferente, por exemplo, “Você teria recebido o empréstimo se sua renda fosse maior em R$ 10.000”. Em outras palavras, uma explicação contrafactual ajuda um sujeito de decisão a decidir o que deve fazer a seguir para obter um resultado desejado, em vez de fornecer apenas recursos importantes que contribuíram para a previsão [13].

IV. RESULTADOS E DISCUSSÕES

Neste trabalhos buscamos aplicar metodologias que facilitem a compreensão de algoritmos de aprendizado de máquina, no qual exploramos o Decision Tree para verificar seu comportamento por ser um modelo interpretável. Além disso, foi aplicado dois métodos de explicabilidade para modelos não interpretáveis ao modelo de Regressão Logística, o LIME que é um modelo local e aproximado e o Partial Dependence Plot que é global e aproximado.

De mesmo modo, foi explorado o modelo DiCE para criar explicações (amostras) que atinjam a saída esperada pelo modelo.

A. Decision Tree

Para treinamento do modelo utilizamos 75% dos dados para treino e 25% dos dados para teste, além de balancear os dados de treino usando SMOTE. O modelo obteve 76% de recall e 89% de acurácia. A estrutura escolhida através de GridSearch foi utilizando “`max_leaf_nodes`": 99, “`min_samples_split`": 3, o que é considerado uma árvore razoavelmente profunda.

As cinco features consideradas mais importantes para o modelo estão descritas na Tabela I.

Feature	Importância
`patients_DESTINATION UNIT_cti_adulto`	0.490011
`patients_STAY DURATION`	0.129358
`icu_AGE`	0.100792
`icu_HOSPITALIZATION TYPE_EMPTY`	0.082964
`patients_DESTINATION UNIT_cti_semi`	0.038974

Tabela I
FEATURES MAIS IMPORTANTES - DECISION TREE

Na Figura 1, podemos analisar o primeiro ramo em caso das primeiras condições serem falsas. A árvore completa pode ser encontrada no apêndice, Figura 5.

Podemos analisar o ramo contido na Figura 1 por profundidade. O nó raiz está em profundidade zero, já ao longo da árvore de decisão há um nó em cada ponto onde uma pergunta é feita, essa ação divide os dados em subconjuntos menores.

Portanto, no nó raiz é verificado se o valor da feature “`patients_DESTINATION UNIT_cti_adulto`” é ≤ 0.5, com base no resultado negativo seguimos para o ramo explorado. Como temos o valor do gini maior do que 0, isso implica que as amostras contidas nesse nó pertencem a classes distintas. O samples mostra a quantidade de amostras no nó. e o value informa quantas amostras se enquadram em cada categoria. E por fim, o class representa a classe que mais ocorre dentro do nó.

Para cada profundidade será explicado os nós proeminentes:

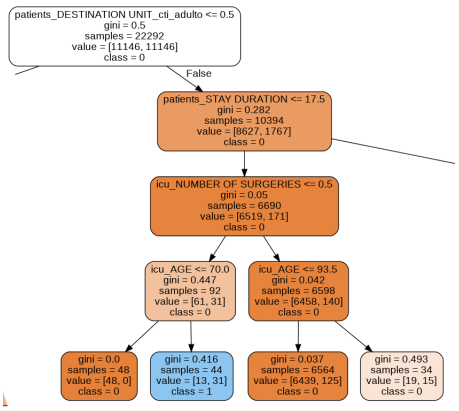

Figura 1. Ramo secundário de decisão do modelo Decision Tree (modelo completo [5])

1) é verificado se o “`patients_STAY DURATION`” é $\leq$ 17.5 dias, se verdadeiro, seguimos para a próxima profundidade, caso contrário, seguimos para o próximo ramo.
2) é verificado se o paciente teve “`icu_NUMBER OF SURGERIES`” $\leq$ 0.5, neste caso, temos uma situação para cada resultado.
3) Neste caso, se verdadeiro (T) e se falso (F):
   (T) Verificamos se “`icu_AGE`” $\leq$ 70 anos, neste caso, também teremos uma situação para cada resultado
   (F) Veificamos se “`icu_AGE`” $\leq$ 93.5 anos, neste caso, também teremos uma situação para cada resultado
4) Para obter a classificação final, seguindo o ramo, teremos duas opções vindas do ramo positivo (T) e duas opções vindos do ramo negativo (F), portanto, o ramo também será incluido na nomenclatura:

(T - T) Se o ramo anterior for positivo, classificamos a amostra como classe 0;
(T - F) Se o ramo anterior for negativo, classificamos a amostra como classe 1;
(F - T) Se o ramo anterior for positivo, classificamos a amostra como classe 0;
(F - F) Se o ramo anterior for negativo, classificamos a amostra como classe 0;

B. Explicabilidade de modelo não interpretável

Por outro lado, para aplicarmos os métodos LIME, Partial Dependence plot e DiCE, utilizaremos o modelo de Logistic Regression como base. O qual treinando com a mesma divisão de dados utilizada no modelo de Decision Tree obteve 90% de acurácia e 57% de recall. E as cinco features mais importantes para o modelo estão descritas na Tabela II.

1) LIME: O LIME pode nos dar uma intuição muito profunda por trás de um processo de decisão de um determinado modelo de caixa preta, ao mesmo tempo que fornece insights sólidos sobre o conjunto de dados, descrevendo as features relevantes para cada decisão.

Feature	Importância
icu_CID_S77	0.198420
icu_CID_M47	0.186263
icu_CID_C49	0.185064
icu_CID_T81	0.183158
icu_CID_L89	0.182907

Tabela II
FEATURES MAIS IMPORTÂNTES - LOGISTIC REGRESSION

A primeira amostra a ser analisada é pertencente à classe 0, e foi predito pelo modelo de regressão logística corretamente, ou seja, é um verdadeiro negativo. Neste caso, o modelo tinha 100% de certeza sobre a classificação, e levou em conta as características descritas pela Figura 2, que no caso representa as features que são levadas em conta para a escolha da classe 0 como vermelhas e as pertencentes a classe 1 como verdes. É interessante observar que a feature "`icu_HOSPITALIZATION TYPE_EMPTY`" que foi considerada como uma das features mais importantes no caso da classificação utilizando Decision Tree, também foi a features que mais influenciou na decisão, considerando que o valor esperado seja > 0. De mesmo modo, as features "`patients_ORIGIN UNIT centro cirúrgico`" > 0 e "`patients_DESTINATION UNIT_cti_adulto`" entre 0 e 1 também foram grandes responsáveis pela classificação.

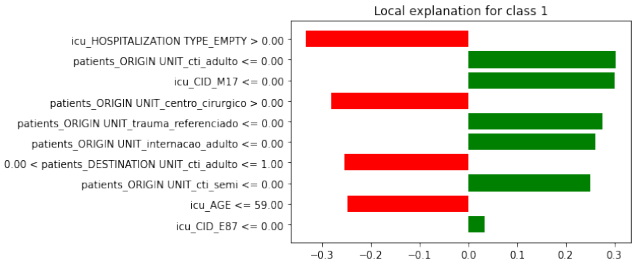

Figura 2. Verdadeiro Negativo (TN)

Por outro lado, a segunda amostra a ser analisada é pertencente à classe 1, e foi predita pelo modelo de regressão logística corretamente, ou seja, é um verdadeiro positivo. Neste caso, o modelo tinha 59% de certeza sobre a classificação, e levou em conta as características descritas pela Figura 3. Neste caso, a feature "`patients_ORIGIN UNIT centro cirúrgico`" comentada anteriormente, nesse caso assume o papel contrário, visto que é uma das features que mais influenciaram na decisão, entretanto, dessa vez considerando que o valor esperado seja < 0. É importante citar que as features negativas nesse caso são de grande influência pois apenas duas features tiveram o efeito de levar o modelo a ter 41% de certeza sobre a classe 0, apenas considerando a idade e se houve trauma referenciado ("`patients_ORIGIN UNIT_trauma referenciado`" e "`icu_AGE`", respectivamente.

2) Partial Dependence Plot: O Partial Dependence Plot mostra o efeito marginal que uma ou mais features têm no resultado previsto de um modelo de aprendizado de máquina.

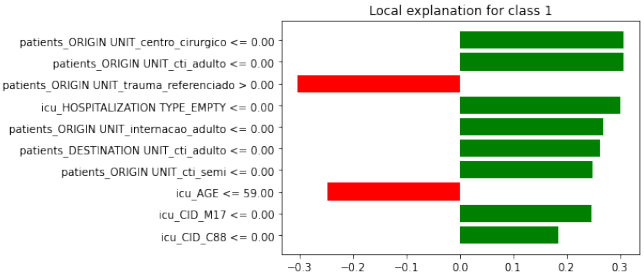

Figura 3. Verdadeiro Positivo (TP)

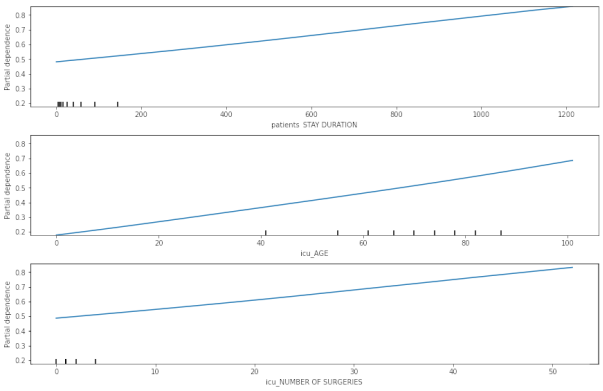

Figura 4. Partial Dependence Plot de 3 features ("`patients_STAY DURATION`", "`icu AGE`" e "`icu_NUMBER OF SURGERIES`"). Os gráficos mostram a inclinação positiva. As marcações no eixo x mostram a distribuição de dados.

Para os resultados, utilizando o conjunto de dados original, avaliamos que apenas as features de idade, número de cirurgias e tempo de estadia tem associação positiva (conforme mostra a Figura 4. Considerando as outras features presentes no modelo para análise, ou não tinham relevância para o efeito marginal ou tinham pouca importância para o resultado final. A feature que mais influencia é a de idade, seguida do tempo de estadia e por último, temos o número de cirurgias, com menor importância para o desfecho.

3) Contrafactual - DiCE: Foram criados 10 exemplos contrafactuais para cada classe utilizando o DiCE, e de forma aleatória, ou seja, faz aleatoriamente a amostra de pontos próximos de um ponto de consulta e retorna contrafactuais como pontos cujos rótulos previstos são a classe desejada.

Para gerar os exemplos da classe negativa, o modelo perturbou as features presentes na tabela III:

Feature	Quantidade de vezes
icu_AGE	6
patients_STAY DURATION	3
icu_HOSPITALIZATION TYPE_EMPTY, icu_HOSPITALIZATION TYPE_INACTIVE - MUNICIPALITY REGULATION - NIR, icu_HOSPITALIZATION TYPE_INFECTION / POSTOPERATIVE COMPLICATION I, icu_HOSPITALIZATION TYPE_STATE REGULATION - NIR, icu_CID_L97, icu_CID_M46, icu_CID_R55, icu_CID_S12, icu_CID_S32, icu_CID_S72, icu_CID_T88, patients_ORIGIN UNIT_cti_adulto, patients_ORIGIN UNIT_internacao_pediatria, patients_DESTINATION UNIT_cti_semi	2

Tabela III
FEATURES QUE SOFRERAM PERTURBAÇÃO

Por outro lado, para gerar os exemplos da classe positiva o modelo perturbou duas vezes as features: `patients_ICU DURATION`, `icu_CID_C79`, `icu_CID_D21`, `icu_CID_E10`, `icu_CID_E87`, `icu_CID_G82`, `icu_CID_I95`, `icu_CID_J12`, `icu_CID_J18`, `icu_CID_L97`, `icu_CID_M87`, `icu_CID_Q72`, `icu_CID_S42`, `icu_CID_S73`.

Para ambos os casos, o modelo classificou corretamente as classes esperadas.

REFERÊNCIAS

[1] C. M. Patino and J. C. Ferreira, "Prognostic studies for health care decision making," *Jornal Brasileiro de Pneumologia*, vol. 43, pp. 252–252, Aug. 2017.

[2] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[3] Z. Jiang, L. Bo, Z. Xu, Y. Song, J. Wang, P. Wen, X. Wan, T. Yang, X. Deng, and J. Bian, "An explainable machine learning algorithm for risk factor analysis of in-hospital mortality in sepsis survivors with icu readmission," *Computer Methods and Programs in Biomedicine*, vol. 204, p. 106040, 2021.

[4] J. W. Hughes, J. E. Olgin, R. Avram, S. A. Abreau, T. Sittler, K. Radia, H. Hsia, T. Walters, B. Lee, J. E. Gonzalez, *et al.*, "Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation," *JAMA cardiology*, vol. 6, no. 11, pp. 1285–1295, 2021.

[5] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *WIREs Data Mining and Knowledge Discovery*, vol. 10, June 2020.

[6] E. M. Polce, K. N. Kunze, M. C. Fu, G. E. Garrigues, B. Forsythe, G. P. Nicholson, B. J. Cole, and N. N. Verma, "Development of supervised machine learning algorithms for prediction of satisfaction at 2 years following total shoulder arthroplasty," *Journal of Shoulder and Elbow Surgery*, vol. 30, no. 6, pp. e290–e299, 2021.

[7] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Explainable ai in healthcare," in *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pp. 1–2, IEEE, 2020.

[8] K. R. Varshney, *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.

[9] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[12] "4.1. partial dependence and individual conditional expectation plots."

[13] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," *CoRR*, vol. abs/1905.07697, 2019.

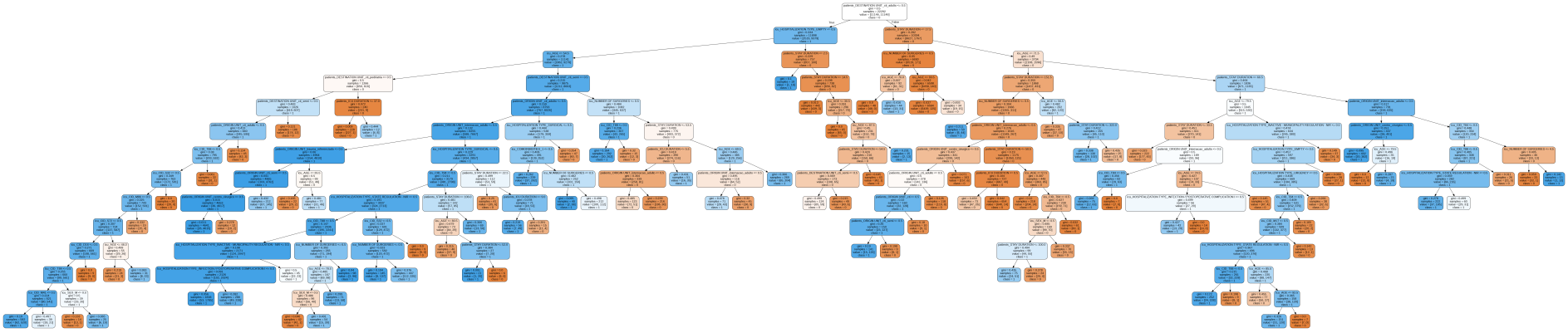

Figura 5. Árvore de Decisão completa