Bayes Decision Model: Estratificação de pacientes - COVID

Alice Mangara[1] **Diana Mortágua**[1]

[1] *Departamento de Engenharia Informática, Faculdade de Ciências e Tecnologias, Universidade de Coimbra*
E–mail: *alicemangara@student.dei.uc.pt, dianamortagua@student.dei.uc.pt*

1. INTRODUÇÃO

A estratificação de pacientes com suspeita de **COVID-19** ao serem admitidos nas salas de emergência hospitalares, desempenha um papel crucial na gestão eficaz dos recursos de saúde e na melhoria dos resultados clínicos.

A abordagem aqui apresentada permite aos profissionais de saúde decidir se o paciente deve ser hospitalizado para exames adicionais ou se pode retornar a casa, com base em variáveis clínicas discretas e contínuas. O nosso projeto estuda o uso da rede *Bayesiana* para modelar a decisão crítica, explorando a fusão de informações provenientes de variáveis discretas e contínuas para aprimorar a precisão das decisões clínicas. Este trabalho não apenas compara o desempenho da rede Bayesiana com outras abordagens, mas também analisa a importância relativa das variáveis e o impacto do tratamento discreto versus contínuo das variáveis nos resultados das previsões. Este estudo simula uma possivel contribuição para a otimização dos protocolos de triagem e gestão inicial de pacientes durante a passada pandemia de COVID-19.

Esperamos com este trabalho confirmar que devemos usar apenas as variáveis que seguem distribuição normal como gaussianas uma vez que o bayes decision model assume que as distribuições são gaussianas. Procuramos ver que variáveis são mais importantes para tomar a decisão final, para saber quais podem deixar de ser registadas para esse fim, comparar variáveis discretas e variáveis continuas, e também concluir se a performance do modelo foi aveitável ou não.

2. DADOS

Os dados utilizados neste estudo correspondem a dados de pacientes com suspeita de COVID-19 admitidos em salas de emergência hospitalares. As variáveis adquiridas no momento da admissão são cruciais para a decisão de hospitalização. A seguir, descrevemos cada uma das variáveis consideradas:

- **Gênero (X1)**: variável binária discreta onde 0 indica feminino e 1 indica masculino.
- **Idade (X2)**: variável contínua que a idade do paciente na admissão, variando de 34 a 99 anos.
- **Estado Civil (X3)**: variável binária discreta onde 0 indica solteiro e 1 indica casado.
- **Vacinação (X4)**: Variável binária discreta onde 0 indica não vacinado e 1 indica vacinado.
- **Dificuldade Respiratória (X5)**: Variável categórica que representa o nível de dificuldade respiratória, com valores possíveis de 0 (nenhuma), 1 (alguma), 2 (moderada) e 3 (alta).
- **Frequência Cardíaca (X6)**: Variável contínua que representa a frequência cardíaca do paciente na admissão, variando de 38 a 272 batimentos por minuto.
- **Pressão Arterial (X7)**: Variável contínua que representa a pressão arterial do paciente na admissão, variando de 115 a 164 *mmHg*.
- **Temperatura (X8)**: Variável contínua que representa a temperatura corporal do paciente na admissão, variando de 36.00 a 38.98 graus *Celsius*.
- **Diretrizes Clínicas (X9)**: Regras baseadas na dificuldade respiratória e temperatura para orientar a decisão médica final.

As colunas foram ajustadas para garantir a precisão dos tipos de dados e a consistência nos valores. As variáveis discretas foram convertidas para inteiros, enquanto as variáveis contínuas foram arredondadas para melhorar a legibilidade e a interpretação dos resultados.

3. ANÁLISE EXPLORATÓRIA

Foi feita uma análise exploratória dos dados, para compreender melhor as características dos pacientes com suspeita de COVID-19 e as variáveis que influenciam a decisão.

Primeiramente, obtivemos estatísticas descritivas e informações gerais sobre o conjunto de dados utilizando os métodos *describe()* e *info()* do ***pandas***, para uma visão inicial da distribuição e dos tipos de variáveis presentes no dataset.

Posteriormente, observámos a distribuição das variáveis Discretas por classe. As variáveis discretas incluídas na análise foram: 'Género' **(G)**, 'Estado Civil' **(MS)**, 'Vacinação' **(V)** e 'Dificuldade Respiratória' **(BD)**.

Aqui [Fig1] estão os resultados, onde podemos observar a influência das variáveis na decisão, comparando as diferenças entre os grupos de decisão:

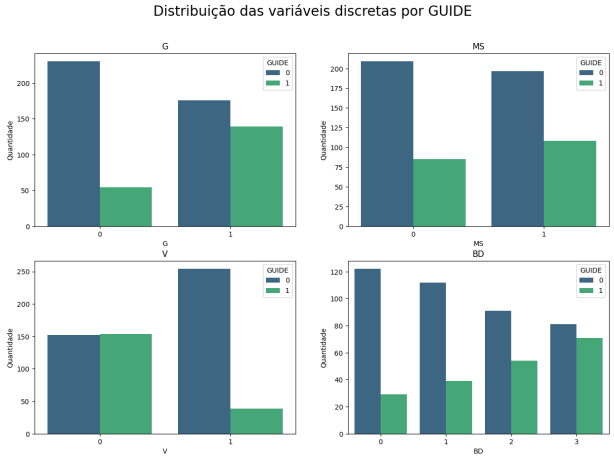

Fig. 1: Distribuição das variáveis discretas por classe.

Podemos aferir que em muitos casos, o valor discreto é um grande indicativo do valor de **GUIDE**, por exemplo, se o valor de **V** for 1, podemos com mais certeza adivinhar que o **GUIDE** será 0 em vez de 1, mas se o valor de **V** for 0, já não podemos ter confiança em nenhuma previsão. Algumas destas distribuições (onde os valores de **GUIDE** diferem bastante para cada valor da variável discreta) indiciam que estas variáveis discretas poderão ser importantes para a tomada de decisão final, principalmente a **V** e o **BD**.

As variáveis contínuas analisadas foram 'Idade' **(AGE)**, 'Frequência Cardíaca' **(HR)**, 'Pressão Arterial' **(BP)** e 'Temperatura' **(T)**. A distribuição dessas variáveis em relação à variável 'GUIDE' foi visualizada usando histogramas com curvas de densidade. Esta análise permitiu visualizarmos [Fig2], numa primeira análise, a gaussianidade das da distribuição das variáveis:

Visualmente podemos ver que a idade apresenta uma forma muito semelhante a uma distribuição normal, enquanto que o **HR** poderia ser semelhante a uma distribuição normal truncada do lado esquerdo, apesar de ainda assim não ser muito semelhante, mas é melhor tratar o truncamento como não sendo gaussiana. Quanto ao **BP** e **T**, podemos claramente ver que os dados não seguem distribuições normais.

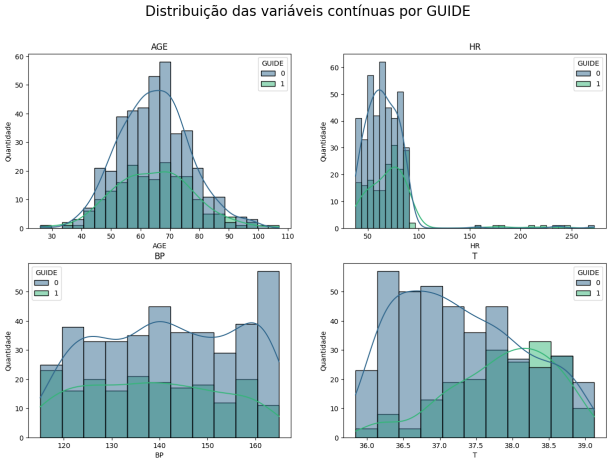

Fig. 2: Distribuição das variáveis contínuas por classe.

Para verificar a normalidade das distribuições das variáveis contínuas para além da analise visual, foram realizados os testes de ***Shapiro-Wilk***, ***Kolmogorov-Smirnov*** e utilizados os ***Q-Q plots***. Os resultados dos testes indicaram que:

- Pelo teste Shapiro-Wilk e pelos Q-Q plots [Fig3]: Apenas a variável relativa à idade segue uma distribuição normal;
- Pelo teste Kolmogorov-Smirnov: Nenhuma variável segue uma distribuição normal.

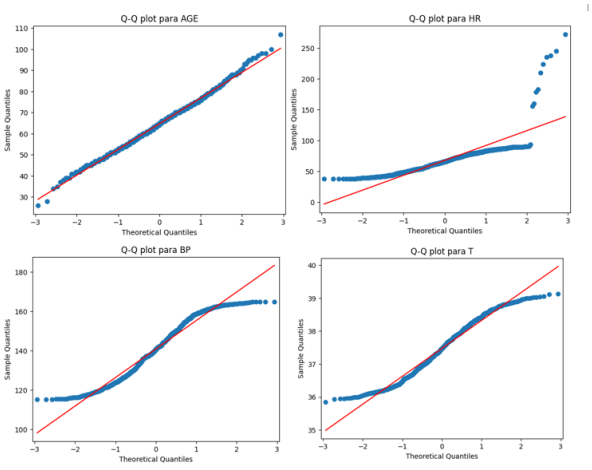

Fig. 3: Q-Q plots.

Usamos esta informação na modelação do nosso

modelo de redes Bayesianas, testando a suposição de normalidade.

Calculámos também uma matriz de correlação [Fig4] para avaliar as relações entre as variáveis. Obtivemos assim *insights* sobre como as variáveis poderão influenciar mais a decisão clínica.

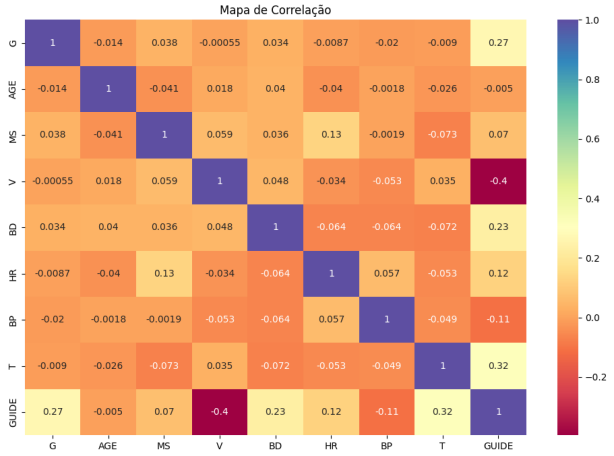

Fig. 4: Correlação entre as variáveis.

Podemos ver que as variáveis relativas a:

- género, vacina, dificuldade em respirar e temperatura são as que melhor se correlacionam com o target. Tirando o género, isto encontra-se dentro do nosso pressuposto subjetivo, o que indica que os dados parecem ser confiáveis.
- idade, estado civil, heart rate e blood pressure são as que menos se correlacionam com o target. Mais uma vez, encontra-se dentro do esperado, tirando a idade que esperarávamos ver como uma das variaveis com maior correlação com o **GUIDE**
- heart rate e estado civil - as mais correlacionadas, apesar de um valor relativamente baixo de 0.4

Finalmente, confirmámos se havia dados em falta, pois a sua existencia podia afetar a precisão dos nossos modelos.

4. METODOLOGIA

Descrição dos Modelos:

- Modelo 1: 'AGE' como gaussiana, todas as outras variáveis discretizadas.
- Modelo 2: Todas as variáveis contínuas discretizadas.
- Modelo 3: Todas as variáveis contínuas como gaussianas.

Para considerar o quão satisfatórios os resultados são teremos em conta a accuracy, a precision, o recall e o f1-score, mas iremos ter em especial atenção à accuracy e ao recall devido ao contexto médico e o facto de ser especialmente importante identificar as pessoas que têm efetivamente covid e devem ficar no hospital (é mais importante identificar corretamente os doentes do que identificar corretamente pessoas saudáveis).

5. RESULTADOS

Analisando os valores relativos à performance, o modelo 1 apresenta accuracy de 0.83, precision de 0.8, recall de 0.67 e f1-score de 0.72. Esta performance a nível de accuracy é boa (acima de 0.7) e a nível de recall, apesar de desejávelmente ser acima de 0.7, está bastante perto, pelo que consideramos aceitável.

O modelo 2 apresenta accuracy de 0.76, precision de 0.67, recall de 0.55 e f1-score de 0.606. A accuracy ainda é aceitável mas tanto a precision como o recall deixam a desejar na sua performance.

O modelo 3 apresenta accuracy de 0.816, precision de 0.8, recall de 0.6 e f1-score de 0.68. A accuracy tem uma boa performance mas o recall poderia ser melhor.

Por estes valores concluímos que o modelo 1 é o preferível, uma vez que todos os valores de avaliação de performance são melhores, e o modelo 2 é o pior entre os três.

Vamos agora analisar a importância das variáveis, vendo como a accuracy e recall dos modelos se alteram sem cada uma das variáveis.

A nível da accuracy, no modelo 1 [Fig5] temos uma descida maior quando são retiradas as colunas **G**, **V** e **T**, baixando 6-9 pontos percentuais. Com o retiro da coluna **BP** a accuracy aumenta 2 pontos percentuais. Em todas as outras diminui ligeiramente ou mantém-se.

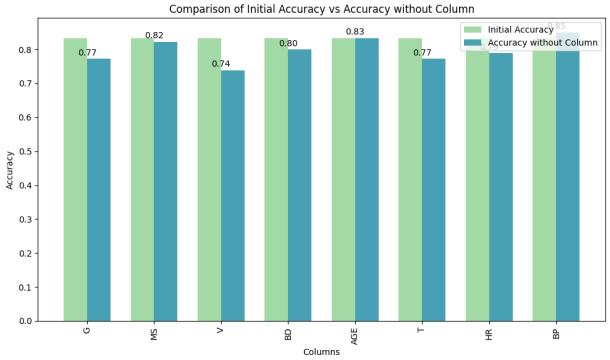

Fig. 5: Accuracy.

No modelo 2 [Fig6] há uma descida de accuracy de 8 pontos percentuais retirando a coluna **V** mas em todos os outros casos desce só ligeiramente ou aumenta ligeiramente. A maior subida é retirando

AGE e **BP**.

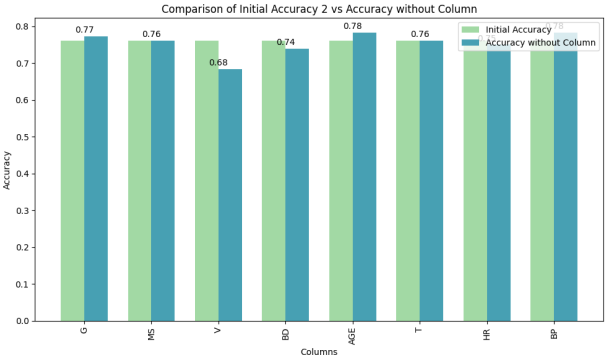

Fig. 6: Accuracy - modelo2.

No modelo 3 [Fig7] a maior descida de accuracy é de novo retirando a coluna **V**, seguindo-se com uma descida de 4 pontos percentuais **T** e **G**. A maior subida é retirando a coluna **BP**

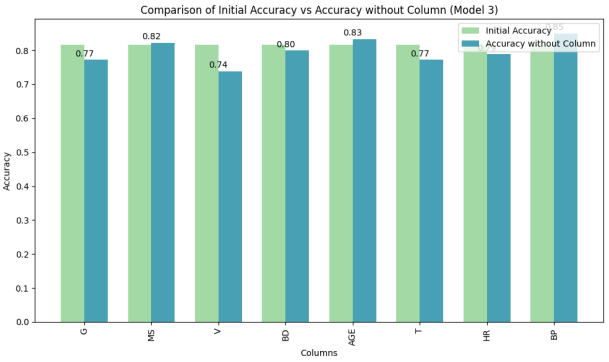

Fig. 7: Accuracy - modelo3.

A nível do recall, temos resultados mais evidentes e substanciais. No modelo 1 [Fig8] vemos uma descida de 22 pontos percentuais retirando a coluna **V** ou a coluna **T**, tendo também uma descida de 9 pontos retirando **BD**. Em nenhum dos casos o recall fica melhor retirando uma variável.

No modelo 2 [Fig9] temos uma descida de 20 pontos percentuais retirando a coluna **V** e temos uma subida de 8 pontos quando retirada **BP**.

No modelo 3 [Fig10] as mudanças não são tão substanciais, mas vemos ainda uma descida de 10,5 pontos percentuais retirando a coluna **V** e uma descida de 8,5 retirando **T**. Temos também uma subida de 3 pontos percentuais returando **BP**.

Podemos ver na figura 11 a accuracy dos modelos nas diferentes circunstâncias tidas em conta, e na figura 12 o mesmo relativamente ao recall.

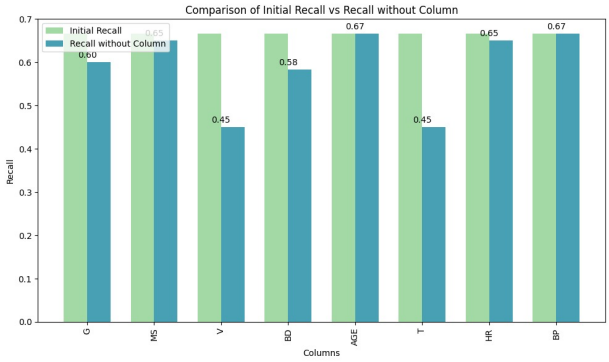

Fig. 8: Recall.

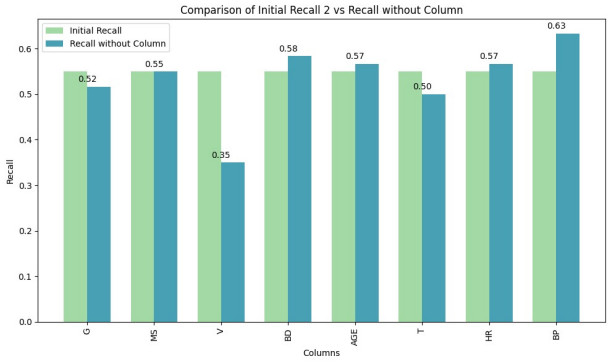

Fig. 9: Recall modelo2.

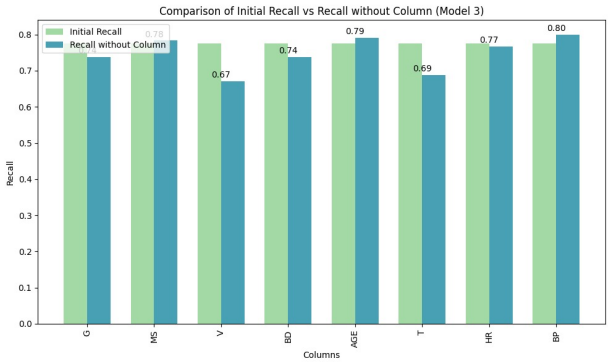

Fig. 10: Recall modelo3.

6. DISCUSSÃO E CONCLUSÃO

Pelos resultados, é visível que, a nível de accuracy, a coluna **V** é a mais importante em todos os modelos (seguindo-se a **T** e **G**) e que a menos importante é o **BP**, pois a accuracy melhora em todos os casos quando esta coluna é retirada.

Segundo o recall, concluímos que as colunas **V** e **T** são também as mais essenciais. Vemos também que, novamente, a coluna **BP** parece ser a que mais prejudica os resultados.

Podemos ver que estas conclusões sobre variáveis

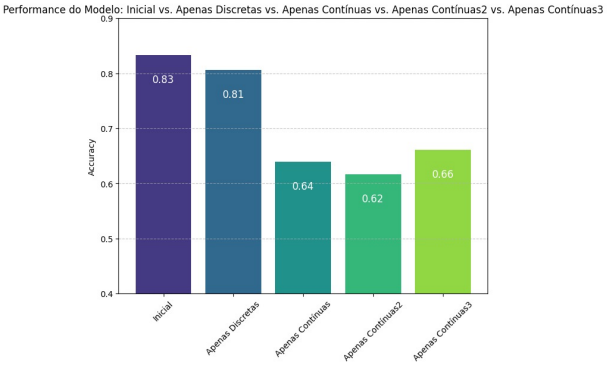

Fig. 11: Performance- accuracy.

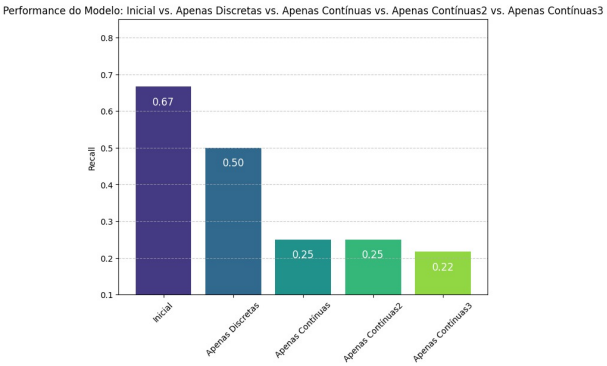

Fig. 12: Performance- recall.

mais importantes ou menos importantes são bastante congruentes às conclusões tiradas analisando as figuras 1 e 2 na fase exploratória.

Pela figura 11 e 12, podemos ver que a tanto a accuracy como o recall têm os melhores valores quando considerado o dataset inicial inteiro, mas que utilizar só variáveis discretas apresentam 97 e 75 por cento respectivamente da accuracy e do recall do dataset. Concluímos então que as variáveis discretas do nosso conjunto de dados são que mais afetam o resultado.

Os resultados tanto da accuracy, como do recall são semelhantes entre os 3 modelos usádos. Os seus resultados com a retirada de apenas uma variável, apesar de diferentes, também eram semelhantes, pelo que concluímos que as diferenças entre os 3 modelos não são significativas para o resultado final usando o bayes decision model.

Concluímos que as variáveis discretas são mais importantes para o resultado do que as contínuas (em geral), e que a coluna **BP** é dispensável sem perda de resultados (pelo contrário, trazendo uma melhoria nos valores de performance do bayes decision model). A performance do classificador foi aceitável na maioria dos casos, como referido na apresentação de resultados dos modelos. No caso de usar os dados todos ou só os dados discretos tivemos performance boas em ambos os casos, ainda melhores do que nos modelos.

Quanto a avaliar a questão das variáveis contínuas serem gaussianas ou não, podemos analisar isso nos modelos. O primeiro modelo onde considerámos a única variável que os testes indicaram como gaussiana (**AGE**) como sendo efetivamente gaussiana e todas as outras como discretizadas, teve o melhor resultado tanto na accuracy como no recall, pelo que consideramos este tratamento como o mais correto neste caso onde apenas uma das variáveis contínuas seguia uma ditribuição gaussiana.

Concluímos então que as colunas mais importantes foram **V** e **T**, que variáveis discretas são mais importantes que as contínuas neste dataset para este fim, e que, tendo variáveis contínuas onde algumas seguem distribuição normal e outras não, a melhor opção é utilizar como gaussianas as que seguem a distribuição, enquanto se discretizam as outras.

REFERENCES

[1] Código disponibilizado pelos professores nas aulas práticas

[2] Chatgpt, https://openai.com/index/chatgpt/