Vinícius de Paula Pilan vinicius.pilan@unesp.br

Modelo para classificação de dados no formato de tabela

Notebook com todo o processo de preparação dos dados e treinamento dos modelos realizado:

https://colab.research.google.com/drive/1kapTAsU30CZXB5wWnTy3cgRy2ibogU_N?usp=sharing

Modelos treinados:

Regressão Logística e Classificador com Floresta Aleatória

Fatores levados em conta na classificação:

1. Tabagismo (nunca fumou, ex-fumante, fumante de baixa frequência, fumante de alta frequência)
2. Situação empregatícia (funcionário público, privado, desempregado ou trabalhador informal)
3. Gênero
4. Tipo de residência (rural ou urbana)
5. Estado civil casado
6. Idade
7. Índice de massa corporal
8. Presença de hipertensão
9. Nível de glicose sanguíneo
10. Presença de problema cardíaco

Resultados obtidos sobre o desempenho dos modelos criados:

Regressão Logística

Porcentagem de acertos: 73.86% +- 2.94%

Porcentagem de acertos especificamente para casos positivos: 76.65% +- 5.36%

Porcentagem de acertos especificamente para casos negativos: 71.95% +- 6.76%

Média da área da curva ROC (max. 1): 0.8032 +- 0.0378

Random Forest

Porcentagem de acertos: 70.24% +- 4.08%

Porcentagem de acertos especificamente para casos positivos: 69.81% +- 6.10%

Porcentagem de acertos especificamente para casos negativos: 71.32% +- 7.39%

Média da área da curva ROC (max. 1): 0.7891 +- 0.0418

Considerações importantes:

1. O conjunto de dados a priori estava extremamente desbalanceado e por isso foi realizado o balanceamento. Os casos de **não avc** foram reduzidos aleatoriamente para que a proporção **avc** e **não avc** se aproximassem da seguinte forma:
    a. 52.5% para não avc (275 casos)
    b. 47.5% para avc (249 casos)

2. Como o conjunto de dados (após o balanceamento) não ficou muito extenso, a avaliação foi feita através de validação cruzada, a partir de 05 dobras.

Conclusão

O algoritmo, dentre os testados, que apresentou melhor desempenho com relação ao problema abordado foi o de Regressão Logística. No geral, o modelo construído com este algoritmo acertou o equivalente ao intervalo de 71 a 74 para cada 100 pessoas testadas.

Com relação à falsas previsões, o modelo teve melhor desempenho para julgar corretamente casos em que o quadro se aproximou de alguém vítima da doença. Para esse tipo de caso (caso positivo), o modelo acertou o que se equivale ao intervalo de 71 a 82 pessoas de 100 avaliadas.

Ainda com relação à falsas previsões, o número de acertos do modelo ao julgar casos de NÃO AVC foi equivalente ao intervalo de 65 a 79 a cada 100 pessoas avaliadas. Esses valores de falsos indicaram, juntamente com a média da curva ROC de 0.765 a 0.841, que o modelo teve dificuldades parecidas para julgar casos de positivos e negativos.

Porém, os números indicam que este modelo teria melhor desempenho no julgamento de casos positivos, o que é interessante para o problema abordado, já que errar o julgamento de um caso positivo resulta em um falso negativo e, na situação atual, isso seria pior do que gerar um falso positivo.