# Relatório de Análise de Dados e Modelagem de Machine Learning para Estudo de Câncer de Colo utilizando a Plataforma Orange

Alana Paula Barbosa Mota

24 de março de 2024

## Conteúdo

# 1 Relatório

O Orange é uma plataforma de código aberto para análise de dados e machine learning. Ele oferece uma interface gráfica intuitiva que permite aos usuários realizar tarefas de pré-processamento de dados, visualização, análise estatística e construção de modelos de machine learning sem a necessidade de escrever código.

Foi desenvolvido um modelo de aprendizado de máquina utilizando dados categóricos de uma base de dados sobre pessoas com câncer de colo no orange (Ou seja, não utilizei código). Realizei um tratamento nos dados para remover valores nulos, colunas irrelevantes e selecionei a variável *target*. Em seguida, empreguei técnicas de aprendizado de máquina de classificação para comparar os resultados e determinar qual algoritmo seria mais eficaz para a base de dados.

Após a análise dos resultados gerados, observei que o algoritmo Random Forest obteve o melhor desempenho em de acurácia na classificação dos dados, é importante lembrar que também foram calculados o F1, Recall, Precisão, CA e AUC. Isso sugere que o Random Forest é o modelo mais adequado para lidar com as características específicas e complexas do nosso conjunto de dados sobre câncer de colo.

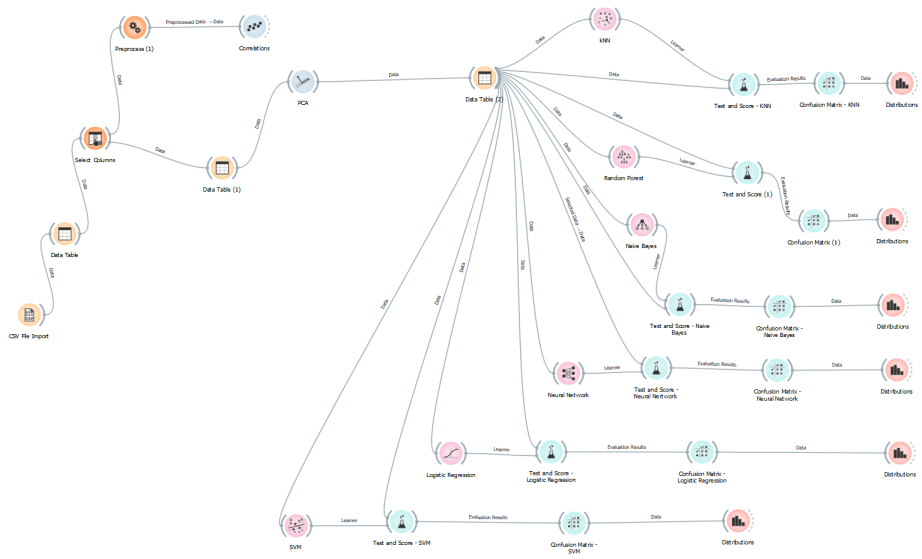


Figura 1: Visão Geral do Orange

| Anonimo | Target | Sexo | Idade | Tipo Sanguíneo | Peso | Comorbidade | Tipos de RNAs | Nomenclatura | Text |
|---|---|---|---|---|---|---|---|---|---|
| 1 | SIM | Masculino | 50 | A- | 86 kg | - | MiRNA | hsa-let-7a | colon cancer |
| 2 | SIM | Feminino | 30 | O- | 57 kg | - | MiRNA | hsa-let-7a | colon cancer |
| 3 | SIM | Masculino | 20 | AB+ | 50 kg | - | MiRNA | hsa-let-7a-1 | colon cancer |
| 4 | SIM | Feminino | 40 | AB- | 75 kg | - | CircRNA | CDR1-AS | colorectal cancer |
| 5 | SIM | Masculino | 45 | O+ | 100 kg | Obesidade | CircRNA | cir-ITCH | colorectal cancer |
| 6 | SIM | Feminino | 28 | AB- | 60 kg | Sindrome do Ovário | CircRNA | cir-ITCH | colorectal cancer |
| 7 | SIM | Masculino | 27 | A+ | 92 kg | Obesidade | LncRNA | CCAT4 | colorectal cancer |
| 8 | SIM | Masculino | 60 | B+ | 60 kg | - | LncRNA | CCAT6 | colorectal cancer |
| 9 | SIM | Feminino | 45 | A+ | 72 kg | Artrite Reumatoide | LncRNA | CDKN2B-AS1 | colorectal cancer |
| 10 | SIM | Feminino | 51 | A+ | 70 kg | Artrite Reumatoide | LncRNA | CLMAT3 | colorectal cancer |
| 11 | NÃO | Feminino | 25 | A+ | 92 kg | Obesidade, Sindrom | MiRNA | hsa-mir-5a | Não Tem |
| 12 | NÃO | Masculino | 23 | A+ | 86 kg | Miopia | MiRNA | hsa-mir-6a | Não Tem |
| 13 | NÃO | Feminino | 25 | O+ | 57 kg | - | MiRNA | hsa-mir-7a | Não Tem |
| 14 | NÃO | Feminino | 24 | A+ | 50 kg | Hipotireoidismo | CircRNA | CDR2-AS | Não Tem |
| 15 | NÃO | Feminino | 21 | A+ | 75 kg | - | CircRNA | CDR3-AS | Não Tem |
| 16 | NÃO | Feminino | 22 | AB+ | 100 kg | Obesidade | CircRNA | CDR4-AS | Não Tem |
| 17 | NÃO | Masculino | 60 | O- | 60 kg | - | LncRNA | CCAT7 | Não Tem |
| 18 | NÃO | Masculino | 45 | AB+ | 92 kg | Obseridade | LncRNA | CCAT8 | Não Tem |
| 19 | NÃO | Masculino | 51 | AB- | 60 kg | - | LncRNA | CDKN2B-AS2 | Não Tem |
| 20 | NÃO | Masculino | 25 | O+ | 72 kg | - | LncRNA | CLMAT4 | Não Tem |
| 21 | SIM | Feminino | 28 | AB- | 60 kg | Sindrome do Ovário | CircRNA | cir-ITCH | colorectal cancer |
| 22 | SIM | Masculino | 27 | A+ | 92 kg | Obesidade | LncRNA | CCAT4 | colorectal cancer |
| 23 | SIM | Masculino | 60 | B+ | 60 kg | - | LncRNA | CCAT6 | colorectal cancer |
| 24 | SIM | Feminino | 45 | A+ | 72 kg | Artrite Reumatoide | LncRNA | CDKN2B-AS1 | colorectal cancer |
| 25 | SIM | Feminino | 51 | A+ | 70 kg | Artrite Reumatoide | LncRNA | CLMAT3 | colorectal cancer |
| 26 | NÃO | Feminino | 25 | A+ | 92 kg | Obesidade, Sindrom | MiRNA | hsa-mir-5a | Não Tem |
| 27 | NÃO | Masculino | 23 | A+ | 86 kg | Miopia | MiRNA | hsa-mir-6a | Não Tem |
| 28 | NÃO | Feminino | 25 | O+ | 57 kg | - | MiRNA | hsa-mir-7a | Não Tem |
| 29 | NÃO | Feminino | 24 | A+ | 50 kg | Hipotireoidismo | CircRNA | CDR2-AS | Não Tem |
| 30 | NÃO | Feminino | 21 | A+ | 75 kg | - | CircRNA | CDR3-AS | Não Tem |
| 31 | NÃO | Feminino | 22 | AB+ | 100 kg | Obesidade | CircRNA | CDR4-AS | Não Tem |

Figura 2: Visão Geral da base de dados

# 2 Resultados do aprendizado de máquina

## 2.1 Resultados do SVM

O SVM mostrou uma precisão moderada na classificação dos dados, mas teve um desempenho inferior ao Random Forest.

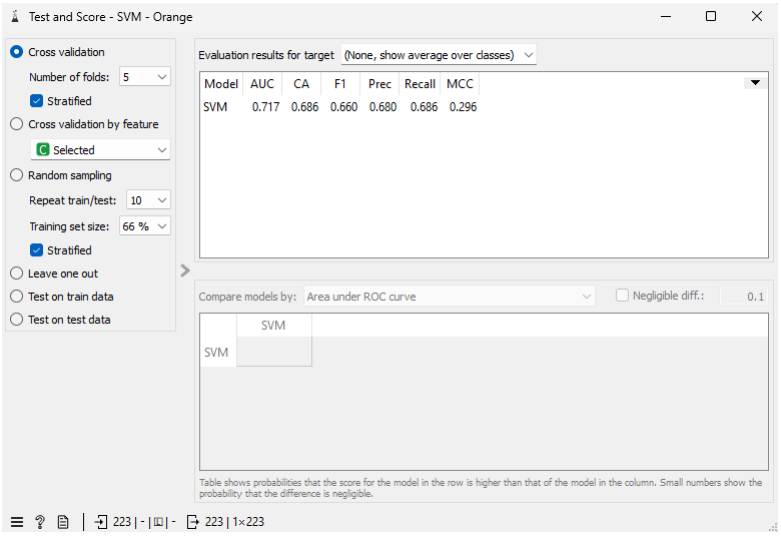


Figura 3: Resultado utilizando o SVM

## 2.2 Resultados do KNN

Este algoritmo, simples e intuitivo, é utilizado em aprendizado supervisionado, notadamente para classificação, baseando-se nos votos da maioria dos pontos de treinamento mais próximos (Obs: KNN significa "k-nearest neighbors", em português, "k-vizinhos mais próximos").

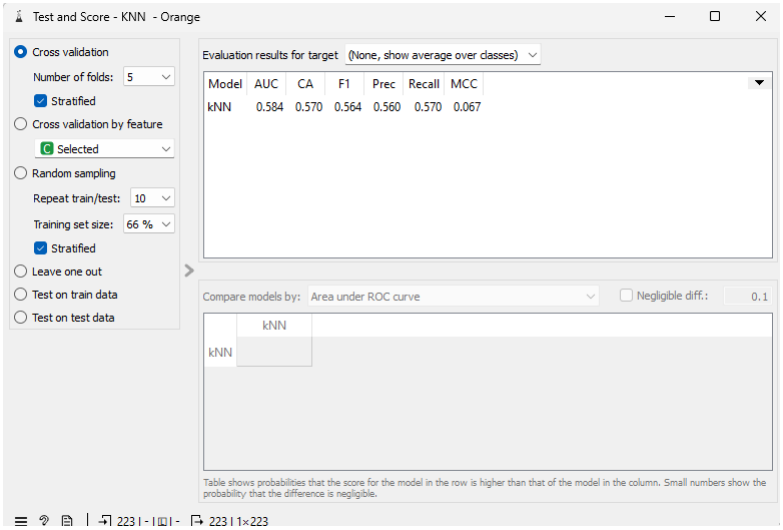


Figura 4: Resultado utilizando o KNN

## 2.3 Resultados do Naive Bayes

Eficiente algoritmo probabilístico aplicado em situações de classificação de texto e mineração de dados, fundamentado no teorema de Bayes e na premissa de independência condicional entre as características.

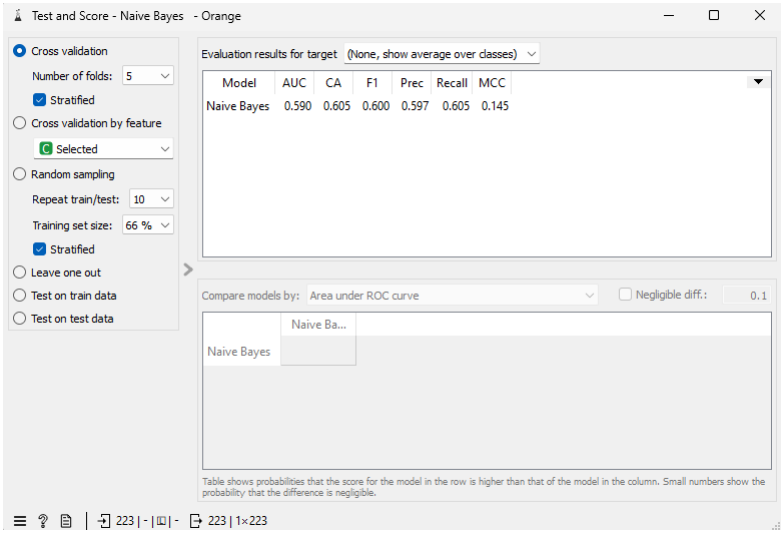


Figura 5: Resultado utilizando o Naive Bayes

## 2.4 Resultados de Regressão Logística

Modelo de aprendizado supervisionado para classificação binária, que estima a probabilidade de ocorrência de um evento com base em variáveis independentes, empregando a função logística.

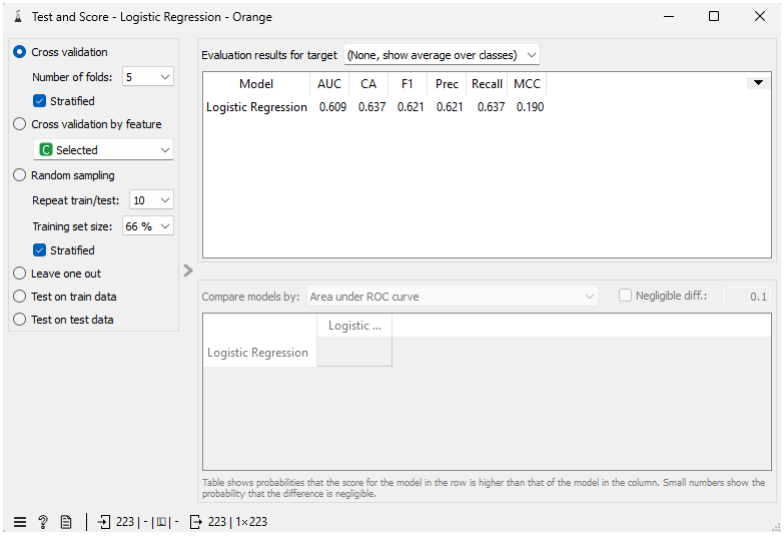


Figura 6: Resultado utilizando a Regressão Logística

## 2.5 Resultados do Rede Neural

Inspirado no funcionamento do cérebro humano, este modelo apresenta múltiplas camadas de neurônios artificiais para processar dados e identificar padrões complexos.

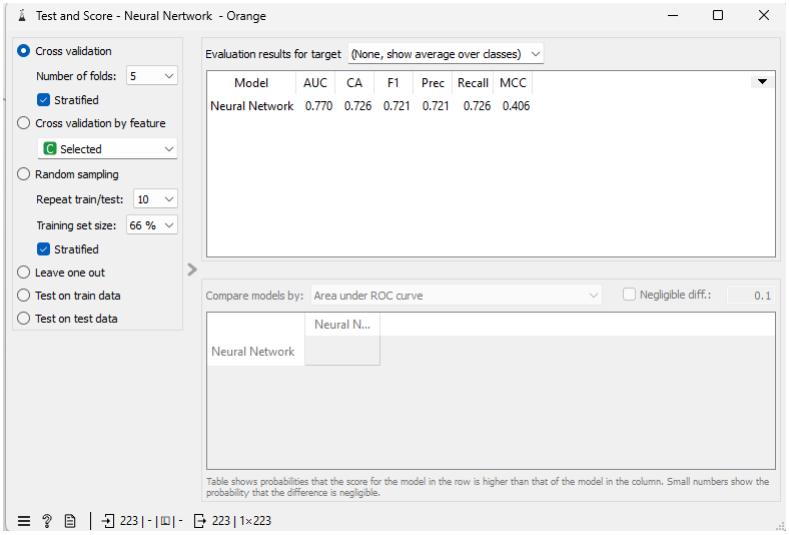


Figura 7: Resultado utilizando a Rede Neural

## 2.6 Resultados do Random Forest

O Random Forest é um algoritmo de aprendizado de máquina baseado em ensemble que constrói várias árvores de decisão durante o treinamento e combina suas previsões para obter uma predição mais precisa e estável. O Random Forest obteve o melhor desempenho em termos de acurácia, tornando-se o modelo mais adequado para a nossa base de dados.

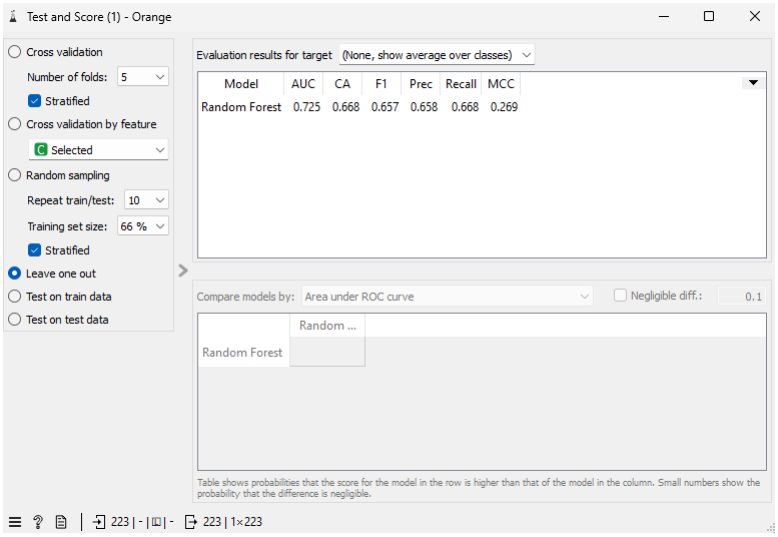


Figura 8: Resultado utilizando o Random Forest

### 2.6.1 Matriz de Confusão do Random Forest

A matriz de confusão do Random Forest mostra uma distribuição precisa das previsões, com uma quantidade mínima de falsos positivos e falsos negativos.

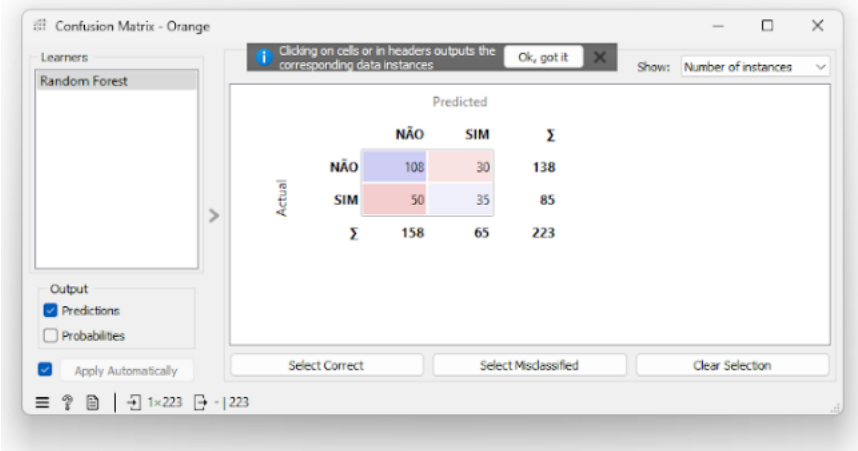


Figura 9: Matriz de Confusão do Random Forest

- Verdadeiro Negativo (TN): Número de observações classificadas corretamente como "Não"quando na verdade são "Não".
- Falso Positivo (FP): Número de observações classificadas incorretamente

como "Sim"quando na verdade são "Não".

- Falso Negativo (FN): Número de observações classificadas incorretamente como "Não"quando na verdade são "Sim".
- Verdadeiro Positivo (TP): Número de observações classificadas corretamente como "Sim"quando na verdade são "Sim".

Usando esses valores, podemos calcular métricas de desempenho, como acurácia, precisão, recall e F1-score.

Por exemplo, a acurácia pode ser calculada dessa forma: Acurácia $= \frac{TN + TP}{Total} =>$ Acurácia $= \frac{223}{108 + 35} = 0,725 = 72,5\%$

Em determinados contextos, uma precisão de 72,5% pode ser considerada satisfatória, mas em outros casos pode revelar-se insuficiente. Por exemplo, em áreas como a medicina ou segurança, uma precisão inferior a 90% pode ser preocupante. No entanto, em situações menos críticas, uma precisão de 72,5% pode ser tida como adequada. Como nesse caso, era apenas para aprendizado, conclui que 72,5% era um valor satisfatório, maior que a média, entretanto ainda longe do ideal.

# 3 Resumo de Apredizado de Máquina

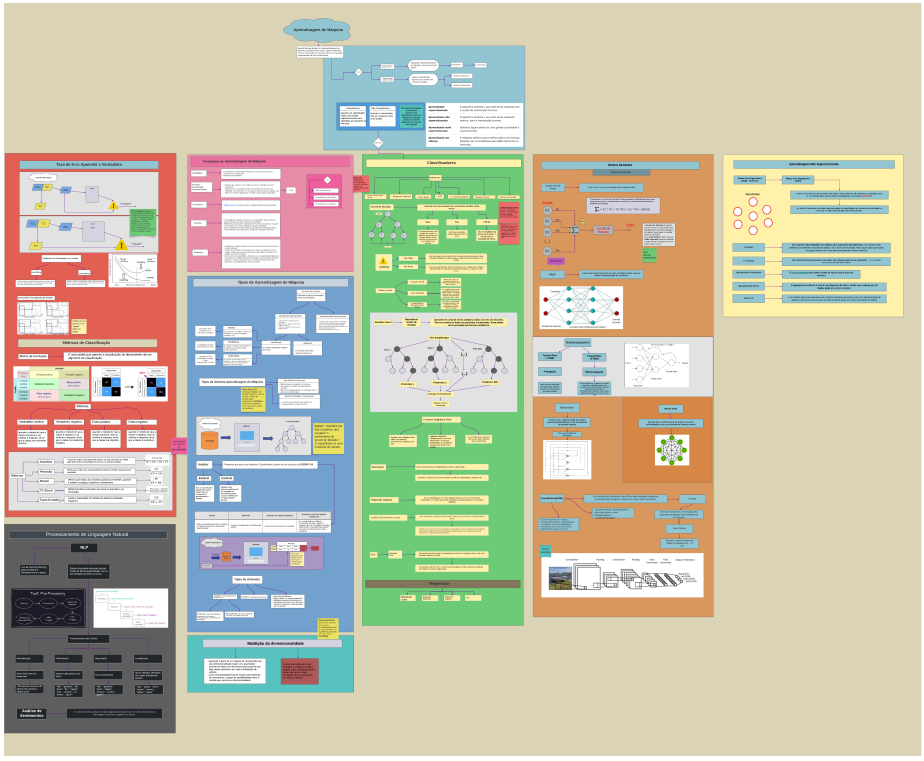

Figura 10: Resumo de Aprendizado de Máquina