Fundação Getúlio Vargas - FGV
MBA em Business Analytics e Big data

Análise Preditiva Avançada
Trabalho de Conclusão de Disciplina

Análise Preditiva da Qualidade de Vinhos - LVMH

João Vitor de Paiva Marcotti

Rio de Janeiro, 2024

Sumário

1. Desafio Empresarial

O conglomerado de luxo LVMH (Moët Hennessy Louis Vuitton), a fim de relacionar as características químicas de seus vinhos e a qualidade percebida por seus consumidores, solicitou que seu cientista de dados João Vitor de Paiva Marcotti desenvolvesse um modelo preditivo de classificação binomial.

Os vinhos analisados pertencem as marcas Château Cheval Blanc e Château d'Yquem, ambas vinícolas renomadas, com garrafas precificadas acima de 3.000 dólares e foco em clientes exigentes, refinados e de altíssima renda.

Desta forma, o modelo deve classificar vinhos brancos e tintos, de ambas as marcas, em dois grupos:

- Reprovados: rótulos avaliados entre 0 e 5.
- Aprovados: rótulos avaliados entre 6 e 10.

Optou-se por esta divisão pois se entende que garrafas avaliadas em mais de 3.000 dólares não devem ser comercializadas com nada além dos melhores vinhos disponíveis nas vinícolas, atendendo ao paladar dos sofisticados e exigentes clientes do conglomerado.

Portanto, com a aplicação do modelo de análise preditiva, espera-se que os melhores vinhos sejam selecionados e engarrafados sob o nome dessas marcas, garantindo assertividade, satisfação e o maior padrão de qualidade disponível aos clientes.

Os vinhos recusados pelo modelo serão utilizados na destilação de conhaque por meio da marca Hennessy ou para temperar barris de carvalho para uso das destilarias de whisky escocês Glenmorangie e Ardbeg (também pertencentes à LVMH).

2. Coleta de Dados

Os dados de análise química de cada rótulo foram coletados pela equipe de laboratório do grupo LVMH, a partir de amostras extraídas diretamente do barril após o respectivo período de maturação de cada vinho.

As notas avaliativas foram coletadas com o consentimento dos participantes, a partir da opinião mista de especialistas e clientes das vinícolas durante sessões de degustação promovidas a partir dos mesmos lotes analisados em laboratório.

3. Identificação dos Dados

Foram obtidas 11 características químicas a partir da análise de 6497 amostras de vinhos brancos e tintos:

- Acidez fixa (g/dm^3);
- Acidez volátil (g/dm^3);
- Ácido clorídrico (g/dm^3);
- Açúcar residual (g/dm^3);

- Cloretos (g/dm^3);
- Dióxido de enxofre livre (mg/dm^3);
- Dióxido de enxofre total (mg/dm^3);
- Densidade (g/dm^3);
- pH;
- Sulfatos (g/dm^3);
- Concentração alcóolica (%).

4. Pré-processamento dos Dados

A base de dados utilizada não contava com valores faltantes (NaN), portanto não foi necessário nenhum tratamento quanto a isto.

A variável “Tipo”, que se refere ao vinho ser branco ou tinto, foi codificada como 1 e 0, respectivamente. Após isso, escalonamentos doss tipo MinMax e Standard foram utilizados para normalizar as variáveis explicativas.

Quanto à variável resposta, as avaliações entre 0 e 5 foram transformadas em 0 e as avaliações entre 6 e 10 foram transformadas em 1, com o objetivo de classificar os vinhos em aprovados (1) e reprovados (0), tornando um problema multinomial (categorias de 0 a 10) em binomial (0 e 1).

5. Análise Exploratória dos Dados

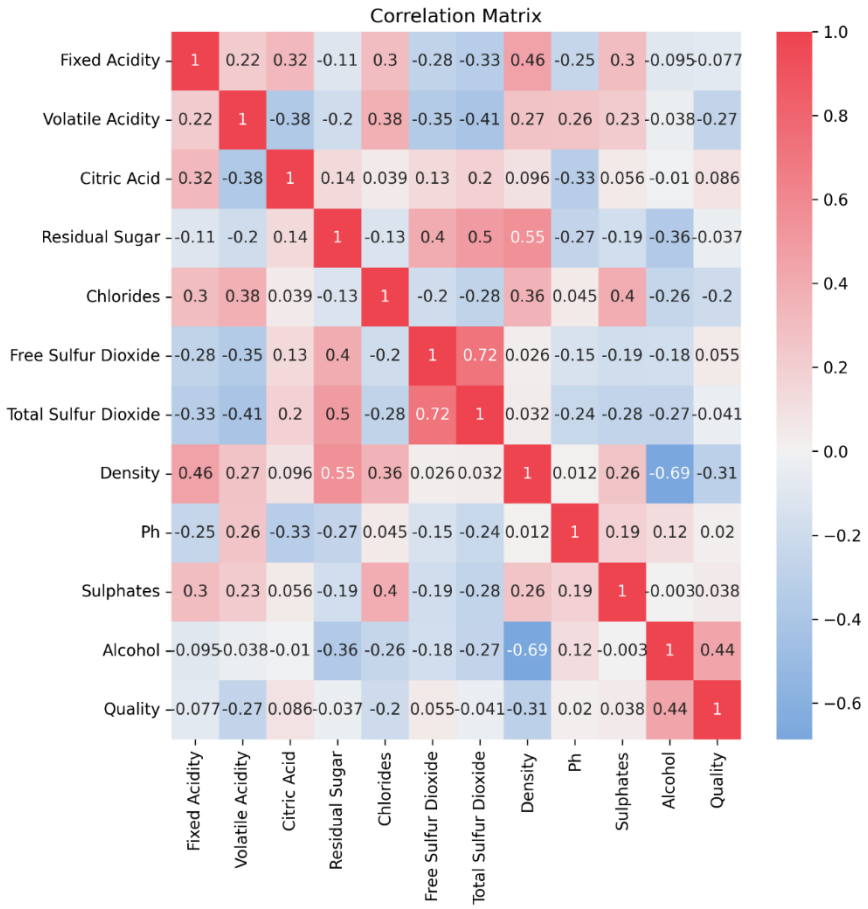
Avaliando o mapa de correlações (Figura 1), nota-se correlação positiva entre as variáveis:

- Dióxido de enxofre livre e dióxido de enxofre total;
- Densidade e açúcar residual.
- Densidade e acidez fixa.

Nota-se também a correlação, negativa ou positiva, entre a variável explicativa qualidade e as variáveis:

- Concentração alcóolica;
- Densidade;
- Cloretos;
- Acidez volátil.

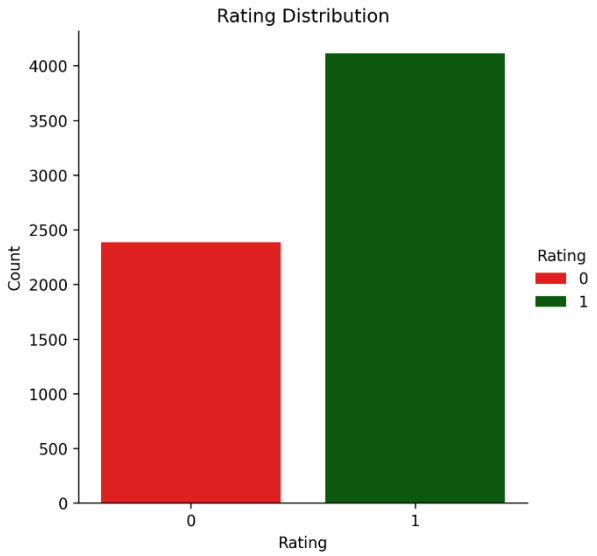
Figura 1. Matriz de correlação.



Fonte: Autoria própria.

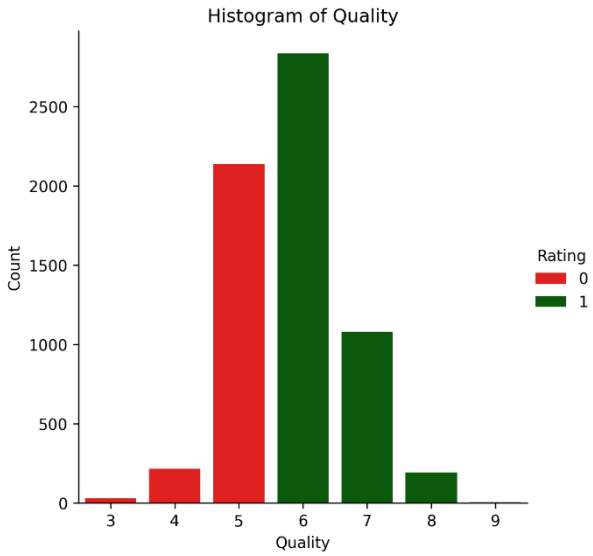
Deve-se destacar a diferença entre as categorias das variáveis explicativas, com 4113 vinhos classificados como aprovados e 2384 como reprovados (Figuras 2 e 3).

Figura 2. Variável explicativa binomial.



Fonte: Autoria própria.

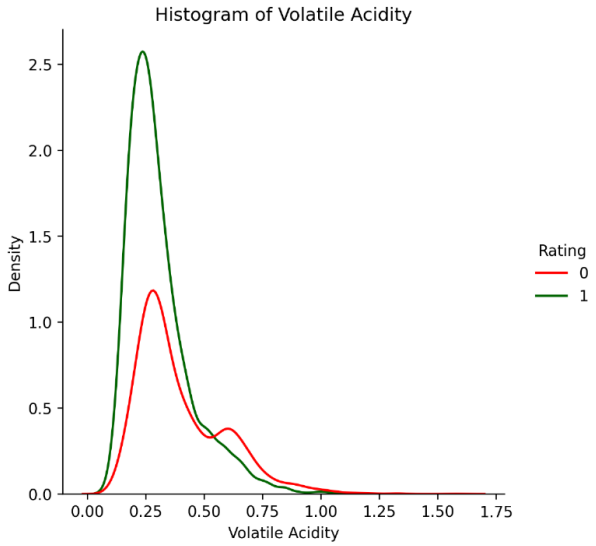
Figura 3. Variável explicativa multinomial.



Fonte: Autoria própria.

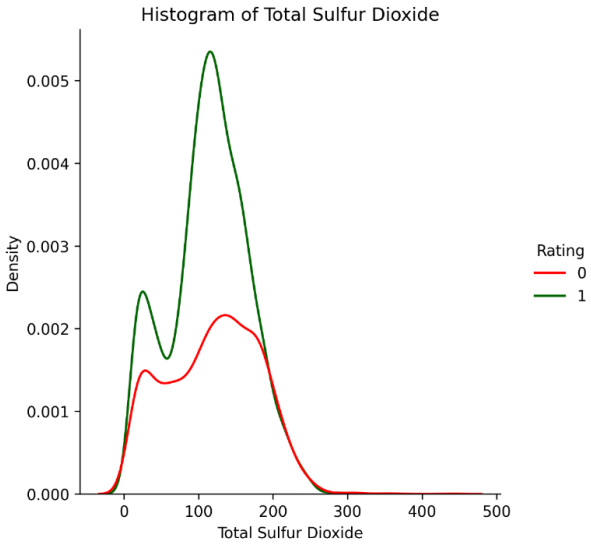
A seguir, tem-se o histograma das variáveis explicativas (Figuras 4 a 14), com os vinhos aprovados em verde e os reprovados em vermelho.

Figura 4. Acidez volátil.



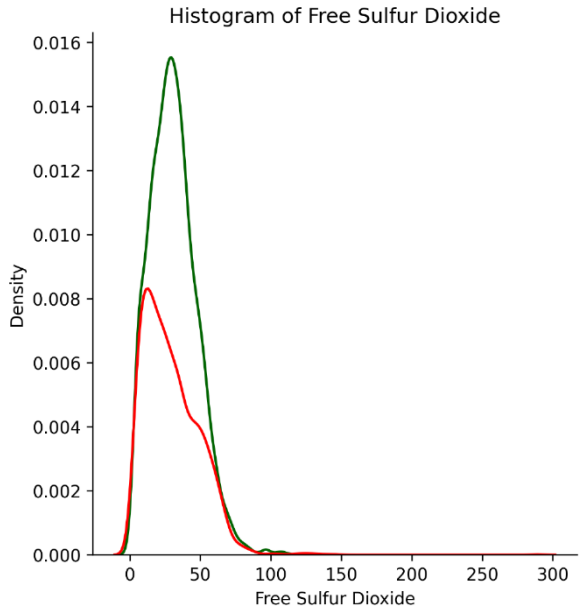
Fonte: Autoria própria.

Figura 5. Dióxido de enxofre total.



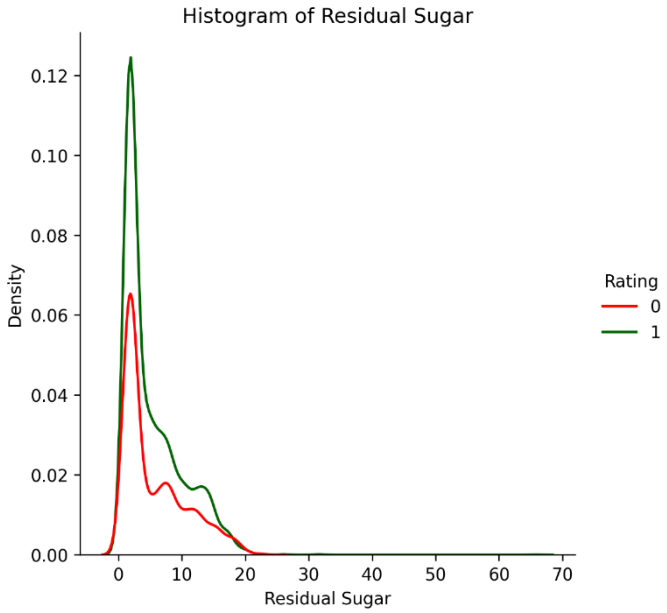
Fonte: Autoria própria.

Figura 6. Dióxido de enxofre livre.



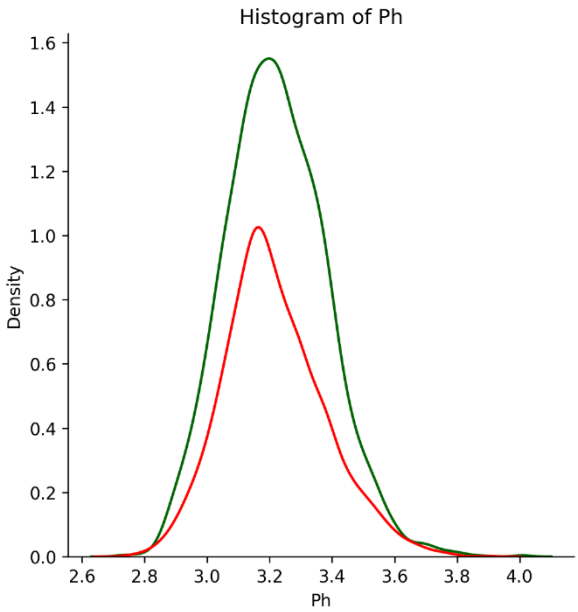
Fonte: Autoria própria.

Figura 7. Açúcar residual.



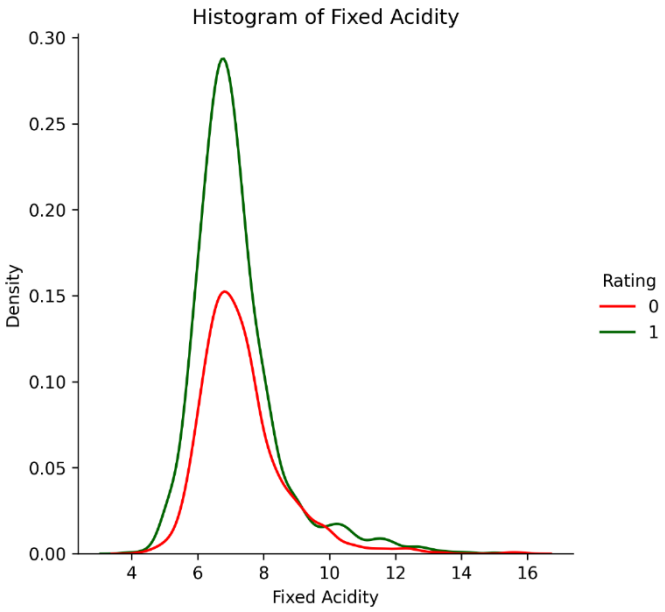
Fonte: Autoria própria

Figura 8. pH



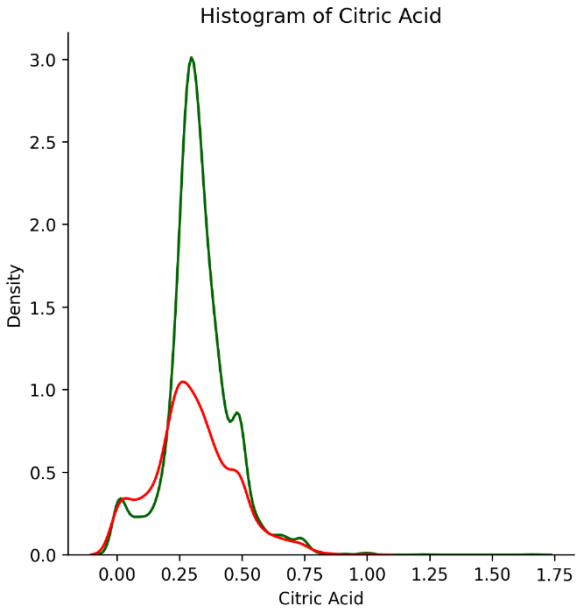
Fonte: Autoria própria.

Figura 9. Acidez fixa.



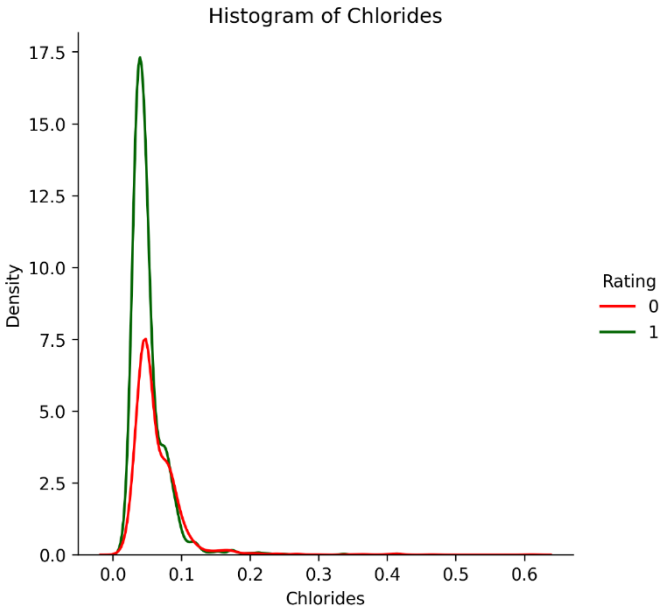
Fonte: Autoria própria.

Figura 10. Ácido cítrico.



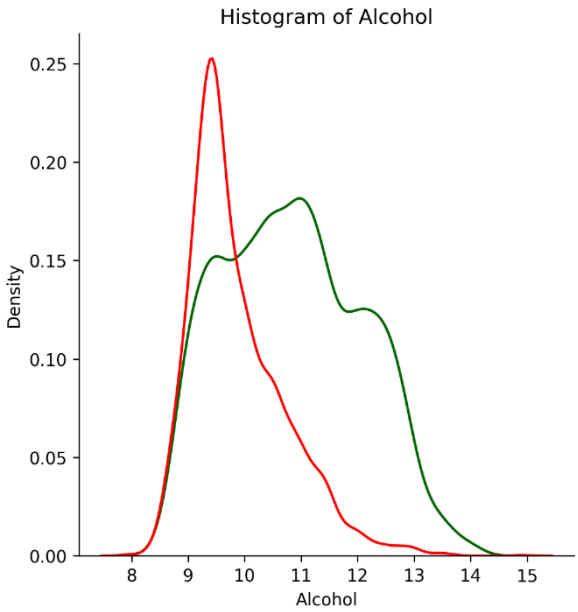
Fonte: Autoria própria.

Figura 11. Cloretos.



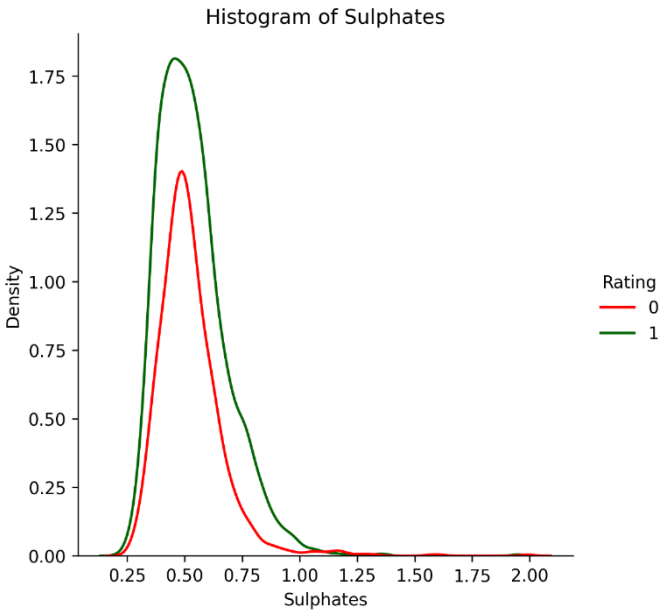
Fonte: Autoria própria.

Figura 12. Álcool.



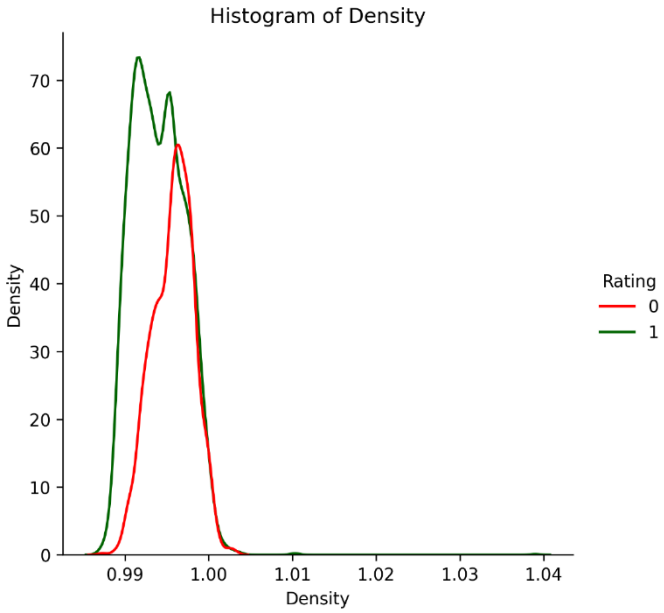
Fonte: Autoria própria.

Figura 13. Sulfatos.



Fonte: Autoria própria.

Figura 14. Densidade.



Fonte: Autoria própria.

5. Desenvolvimento dos Modelos

Desenvolveu-se modelos dos tipos Support Vector Machines (SVM), Random Forest, Naive Bayes, Regressão Logística e Redes Neurais Artificiais. Variou-se seus parâmetros por meio da função “GridSearchCV”, do pacote Scikit Learn:

- Regressão Logística:
  - 'C': [0.1, 1, 10, 100].

    - ‘penalty': ['l1', 'l2'].
    - 'solver': ['saga', 'liblinear'].
- SVM:
    - 'C': [0.1, 1, 10, 100].
- Naive Bayes:
    - 'var_smoothing': [1e-9, 1e-8, 1e-7, 1e-6, 1e-5].
- Random Forest:
    - 'n_estimators': [100, 200, 300].
    - 'max_depth': [10, 20, 30].
    - 'min_samples_split': [2, 5, 10].
    - 'min_samples_leaf': [1, 2, 4].
- Neural Network:
    - 'hidden_layer_sizes': [(150,), (150,50), (100,30)].
    - 'activation': ['relu', 'tanh'].

Os modelos foram treinados e testados com variáveis não escalonadas e escalonadas por meio dos métodos Standard e MinMax, alternando as suas variáveis explicativas entre 4 grupos:

- X1: Contém todas as variáveis explicativas.
- X2: Apenas 'Alcohol', 'Density', 'Chlorides'e 'Volatile Acidity', pois possuem alta correlação com a variável explicativa.
- X3: Todas as variáveis exceto 'Total Sulfur Dioxide','Residual Sugar' e ‘Fixed Acidity'], pois possuem alta correlação entre si.
- X4: Apenas as variáveis mais significativas segundo os melhores modelos de regressão logística.

Desta forma, obteve-se 60 modelos com os melhores hiperparâmetros em relação a um determinado grupo de variáveis explicativas e tipo de escalonamento (Tabela 1).

6. Comparação dos Modelos

Entende-se que classificar um lote de vinhos como aprovado de forma equivocada, entregando aos clientes produtos de baixa ou média qualidade, afeta negativamente tanto a vinícola em questão, quanto a imagem do grupo LVMH, podendo ocasionar danos superiores ao prejuízo causado pela reprovação e descarte equivocado de um lote de vinho de excelente qualidade.

Desta forma, tem-se a sensibilidade, ou recall, como a principal métrica a ser considerada na escolha e ranqueamento dos modelos desenvolvidos. Além disso, também se considerou o erro médio quadrático (MSE) e a área sob a curva ROC (AUC).

7. Avaliação dos Modelos

Tabela 1. Métricas dos modelos selecionados.

Modelo	Var	Esc	MSE	Acurácia	F1	Recall	AUC
RL	X1	Not	0.3477	0.6523	0.7708	0.9291	0.5558

RL	X2	Not	0.2672	0.7328	0.8074	0.8900	0.6780
RL	X3	Not	0.3713	0.6287	0.7720	0.9992	0.4996
RL	X4	Not	0.2728	0.7272	0.7973	0.8525	0.6835
RL	X1	MinMax	0.2810	0.7190	0.7970	0.8769	0.6639
RL	X2	MinMax	0.2846	0.7154	0.7931	0.8672	0.6625
RL	X3	MinMax	0.2831	0.7169	0.7951	0.8729	0.6626
RL	X4	MinMax	0.2810	0.7190	0.7970	0.8769	0.6639
RL	X1	Standard	0.2554	0.7446	0.8059	0.8427	0.7104
RL	X2	Standard	0.2605	0.7395	0.8026	0.8419	0.7038
RL	X3	Standard	0.2615	0.7385	0.8003	0.8329	0.7055
RL	X4	Standard	0.2544	0.7456	0.8069	0.8443	0.7112
SVM	X1	Not	0.3708	0.6292	0.7724	1.0000	0.5000
SVM	X2	Not	0.3508	0.6492	0.7782	0.9780	0.5346
SVM	X3	Not	0.3708	0.6292	0.7724	1.0000	0.5000
SVM	X4	Not	0.3226	0.6774	0.7858	0.9405	0.5857
SVM	X1	MinMax	0.2708	0.7292	0.8066	0.8973	0.6706
SVM	X2	MinMax	0.2621	0.7379	0.8071	0.8712	0.6915
SVM	X3	MinMax	0.2682	0.7318	0.8072	0.8924	0.6758
SVM	X4	MinMax	0.2708	0.7292	0.8066	0.8973	0.6706
SVM	X1	Standard	0.2497	0.7503	0.8157	0.8786	0.7055
SVM	X2	Standard	0.2544	0.7456	0.8114	0.8696	0.7024
SVM	X3	Standard	0.2544	0.7456	0.8120	0.8729	0.7013
SVM	X4	Standard	0.2472	0.7528	0.8177	0.8810	0.7081
NB	X1	Not	0.3226	0.6774	0.7640	0.8297	0.6244
NB	X2	Not	0.3159	0.6841	0.7732	0.8557	0.6243
NB	X3	Not	0.3092	0.6908	0.7742	0.8427	0.6378
NB	X4	Not	0.3138	0.6862	0.7743	0.8557	0.6270
NB	X1	MinMax	0.3185	0.6815	0.7583	0.7938	0.6424
NB	X2	MinMax	0.3138	0.6862	0.7673	0.8223	0.6387
NB	X3	MinMax	0.3190	0.6810	0.7630	0.8158	0.6340
NB	X4	MinMax	0.3185	0.6815	0.7583	0.7938	0.6424
NB	X1	Standard	0.3185	0.6815	0.7583	0.7938	0.6424
NB	X2	Standard	0.3138	0.6862	0.7673	0.8223	0.6387
NB	X3	Standard	0.3190	0.6810	0.7630	0.8158	0.6340
NB	X4	Standard	0.3154	0.6846	0.7604	0.7954	0.6460
RF	X1	Not	0.1759	0.8241	0.8647	0.8932	0.8000
RF	X2	Not	0.2046	0.7954	0.8405	0.8566	0.7741
RF	X3	Not	0.1846	0.8154	0.8579	0.8859	0.7908
RF	X4	Not	0.2000	0.8000	0.8467	0.8778	0.7729
RF	X1	MinMax	0.1708	0.8292	0.8682	0.8941	0.8066
RF	X2	MinMax	0.2082	0.7918	0.8376	0.8533	0.7704
RF	X3	MinMax	0.1862	0.8138	0.8566	0.8835	0.7896
RF	X4	MinMax	0.1708	0.8292	0.8682	0.8941	0.8066
RF	X1	Standard	0.1749	0.8251	0.8654	0.8932	0.8014
RF	X2	Standard	0.2077	0.7923	0.8382	0.8549	0.7705
RF	X3	Standard	0.1856	0.8144	0.8571	0.8851	0.7897
RF	X4	Standard	0.1795	0.8205	0.8611	0.8843	0.7983
NN	X1	Not	0.2508	0.7492	0.8089	0.8435	0.7164

NN	X2	Not	0.2805	0.7195	0.7882	0.8297	0.6811
NN	X3	Not	0.2626	0.7374	0.7957	0.8126	0.7113
NN	X4	Not	0.2764	0.7236	0.7779	0.7694	0.7076
NN	X1	MinMax	0.2523	0.7477	0.8136	0.8753	0.7032
NN	X2	MinMax	0.2605	0.7395	0.8034	0.8460	0.7024
NN	X3	MinMax	0.2333	0.7667	0.8153	0.8183	0.7487
NN	X4	MinMax	0.2523	0.7477	0.8136	0.8753	0.7032
NN	X1	Standard	0.2318	0.7682	0.8223	0.8525	0.7388
NN	X2	Standard	0.2482	0.7518	0.8094	0.8378	0.7218
NN	X3	Standard	0.2349	0.7651	0.8214	0.8582	0.7327
NN	X4	Standard	0.2323	0.7677	0.8241	0.8647	0.7339

Fonte: Autoria própria.

8. Escolha de um Modelo

Como mencionado na seção 6, a principal métrica a ser levada em consideração na escolha do modelo foi a sua sensibilidade. Assim, primeiramente se ordenou os modelos em ordem decrescente de acordo com esta métrica:

Tabela 1. Modelos selecionados ordenados pela sensibilidade/recall.

#	Modelo	Var	Esc	MSE	Acurácia	F1	Recall	AUC
1	SVM	X1	Not	0.3708	0.6292	0.7724	10.000	0.5000
2	SVM	X3	Not	0.3708	0.6292	0.7724	10.000	0.5000
3	RL	X3	Not	0.3713	0.6287	0.7720	0.9992	0.4996
4	SVM	X2	Not	0.3508	0.6492	0.7782	0.9780	0.5346
5	SVM	X4	Not	0.3226	0.6774	0.7858	0.9405	0.5857
6	RL	X1	Not	0.3477	0.6523	0.7708	0.9291	0.5558
7	SVM	X1	MinMax	0.2708	0.7292	0.8066	0.8973	0.6706
8	SVM	X4	MinMax	0.2708	0.7292	0.8066	0.8973	0.6706
9	RF	X1	MinMax	0.1708	0.8292	0.8682	0.8941	0.8066
10	RF	X4	MinMax	0.1708	0.8292	0.8682	0.8941	0.8066

Fonte: Autoria própria.

No entanto, analisando-se os 10 modelos de maior sensibilidade/recall, notou-se que os 6 primeiros apresentam área sob a curva (AUC) próximas às de um modelo aleatório (0.50). Desta forma, descartou-os e se obteve um novo ranking:

Tabela 3. Modelos selecionados ordenados pela sensibilidade/recall.

#	Modelo	Var	Esc	MSE	Acurácia	F1	Recall	AUC
1	SVM	X1	MinMax	0.2708	0.7292	0.8066	0.8973	0.6706
2	SVM	X4	MinMax	0.2708	0.7292	0.8066	0.8973	0.6706
3	RF	X1	MinMax	0.1708	0.8292	0.8682	0.8941	0.8066
4	RF	X4	MinMax	0.1708	0.8292	0.8682	0.8941	0.8066
5	RF	X1	Standard	0.1749	0.8251	0.8654	0.8932	0.8014
6	RF	X1	Not	0.1759	0.8241	0.8647	0.8932	0.8000
7	SVM	X3	MinMax	0.2682	0.7318	0.8072	0.8924	0.6758
8	RL	X2	Not	0.2672	0.7328	0.8074	0.8900	0.6780

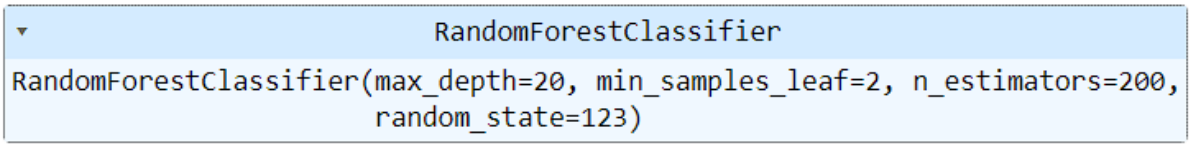
9	RF	X3	Not	0.1846	0.8154	0.8579	0.8859	0.7908
10	RF	X3	Standard	0.1856	0.8144	0.8571	0.8851	0.7897

Fonte: Autoria própria.

Notou-se que os modelos 3 e 4, do tipo Random Forest, apresentam a melhor combinação de métricas MSE, acurácia, F1, recall e AUC, empatados em todas elas.

Desta forma, em razão de trabalhar com menos variáveis, selecionou-se o modelo 4: Random Forest, com seleção de variáveis do grupo X4, escalonamento MinMax, com os seguintes atributos:

Figura 15. Atributos do modelo selecionado.



Fonte: Autoria própria.

Por fim, em razão da diferença entre as categorias das variáveis explicativas (Figuras 2 e 3), decidiu-se aplicar a técnica de tratamento de dados desbalanceados SMOTE (Synthetic Minority Over-sampling TEchnique) às variáveis do grupo X4 e retreinar o melhor modelo, apresentando um comparativo entre suas duas performances (Tabela 4).

Tabela 4. Comparativo de performances do melhor modelo.

Modelo	Var	Esc	MSE	Acurácia	F1	Recall	AUC
RF SMOTE	X4	MinMax	0.1754	0.8246	0.859	0.8492	0.816
RF	X4	MinMax	0.1708	0.8292	0.8682	0.8941	0.8066

Fonte: Autoria própria.

Nota-se que o uso da técnica SMOTE para rebalanceamento da base de dados de treino não apresentou melhora ao modelo. Portanto, manteve-se a escolha do melhor modelo anterior.

Referências Bibiográficas

[1] - GridSearchCV. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. Acesso em: 10/07/2024.

[2] - SMOTE. Disponível em: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html. Acesso em: 11/10/2024.