Classificação e agrupamento de dados

Bianca S. Wolfgran

[1]Unieuro
Campus Águas Claras – Brasília, DF – Brazil

biancawolfgram@gmail.com

Abstract. *This article addresses the methods of data classification and clustering using two distinct databases. One database focuses on assessing the possibility of a group of women having diabetes, while the other pertains to differentiating three types of wines. The objective is to analyze the information provided by the databases, compare the performance of classification algorithms, and identify the relevance of attributes for clustering. The results revealed minimal variations in the performance test values of the classification algorithms, whereas clustering emphasized the significance of attribute selection.*

Resumo. *O presente artigo tem como propósito abordar os procedimentos empregados para a classificação e agrupamento de dados, utilizando duas bases de dados distintas. Uma das bases de dados em questão está relacionada à possibilidade de mulheres apresentarem diabetes, enquanto a outra trata da diferenciação de três tipos de vinhos. O objetivo primordial consiste em realizar uma análise das informações contidas nessas bases de dados, além de comparar o desempenho dos algoritmos de classificação e identificar a relevância dos atributos para o processo de agrupamento. Os resultados obtidos demonstraram variações insignificantes nos valores de desempenho dos algoritmos de classificação quando submetidos aos testes. Entretanto, o processo de agrupamento revelou a importância fundamental da seleção adequada dos atributos utilizados.*

1. Introdução

Na mineração de dados existem algumas atividades importantes para a obtenção de dados úteis a depender do problema a ser solucionado, neste artigo serão analisadas informações obtidas de duas base de dados distintas utilizando classificação e agrupamento de dados respectivamente para cada base.

A primeira base de dados será analisada através da classificação de dados, essa base tem como objetivo prever se uma paciente tem diabetes ou não, nesse conjunto de dados, todas as pacientes são mulheres com idade em torno dos 21 anos e todas descendentes dos índios Pima, povo nativo presente nos Estados Unidos.

O segundo conjunto de dados, analisado através do método de agrupamento de dados, contém dados sobre uma análise química de vinhos cultivados por três diferentes vinicultores de uma região em comum da Itália.

O objetivo deste artigo é encontrar e interpretar informações pertinentes retiradas das bases de dados citadas, utilizando os métodos de classificação e agrupamento de dados. Além de comparar diferentes tipos de algoritmos para comparar qual obteve melhor funcionalidade.

2. Classificação de dados

2.1. Base de dados dos Índios Pima

Os Índios Pima são um grupo nativo existente nos EUA (Estados Unidos da América) e México, essa base de dados compreende unicamente a parte que se localiza no EUA, pois segundo a Academia Brasileira de Letras, é o grupo mais obeso e que tem mais diabetes nos EUA. [de Letras and Scliar 2021] Um total de 769 pacientes mulheres desse grupo participaram da base de dados.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	627	50	1
1	85	66	29	0	26.6	351	31	0
8	183	64	0	0	23.3	672	32	1
1	89	66	23	94	28.1	167	21	0
0	137	40	35	168	43.1	2288	33	1
5	116	74	0	0	25.6	201	30	0
3	78	50	32	88	31	248	26	1
10	115	0	0	0	35.3	134	29	0
2	197	70	45	543	30.5	158	53	1
8	125	96	0	0	0	232	54	1
4	110	92	0	0	37.6	191	30	0

Figura 1. Base de dados Índios Pima. Fonte: [Learning 2016]

A base de dados contém 9 colunas, sendo 8 com informações sobre as pacientes e uma com a classificação se tem ou não diabetes e 769 linhas. As informações que o banco usa são respectivamente: quantidade de vezes que a paciente esteve grávida, nível de glicose, nível da pressão sanguínea, grossura da pele que é medida pela dobra cutânea (que segundo o manual de métricas antropométricas [Santos 2012] é a medida da espessura de duas camadas de pele e a gordura subcutânea adjacente), nível de insulina, IMC (índice de massa corporal), FPD (função da pedigree da diabetes: medida à tendência ao desenvolvimento da diabetes com base nas relações genéticas da paciente. [Junior 2016]) e por último a idade da paciente.

2.2. Impressões sobre as informações

Com base na classificação dos dados, o tópico que contém maior importância para o diagnóstico de diabetes da base de dados foi a coluna de glicose como mostra o gráfico na Figura 2. A principal característica de uma pessoa diabética é a taxa de glicose no sangue, uma pessoa já é considerada diabética quando em jejum atinge 126mg/dl [TJDFT 2021].

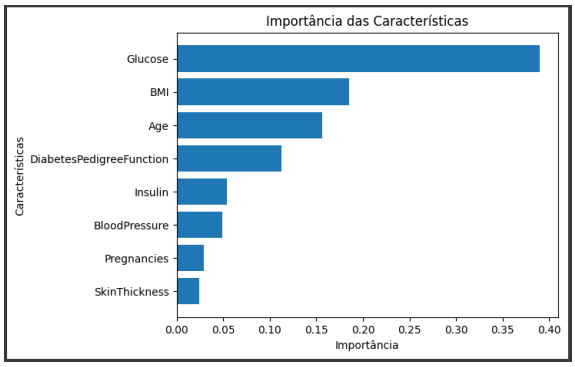

Figura 2. Grau de importância das colunas. Fonte: Autor

Apesar disso, quando analisada a correlação das características de uma pessoa considerada diabética, as principais características que quando relacionadas tendem a ter uma maior taxa de diabetes é a quantidade de vezes que a pessoa teve gravidez e a alta

idade. Seguido pela grossura da pele medida através da dobra cutânea e a quantidade de insulina. Veja Figura 3. Segundo o Clube do Diabetes [do diabetes 2018], quanto maior a concentração de insulina no corpo, ao tentar descartar o excesso do organismo, a pessoa tende a perder muito líquido, causando desidratação da pele e por consequência criando rachaduras, coceira e ambiente propício para infecções.

VALOR	INFORMAÇÕES
0.544341	Gravidez - Idade
0.436783	Grossura da pele - Insulina
0.392573	Grossura da pele - IMC
0.331357	Glicose - Insulina
0.281805	Pressão sanguíne - IMC
0.185071	Insulina - FPD

Figura 3. Tabela correlação entre colunas. Fonte: Autor

A seguir os gráficos com as três principais causas de diabetes e sua relação com as outros tópicos da base de dados:

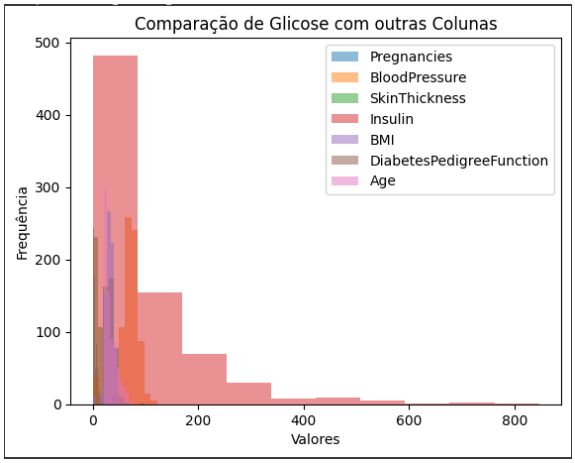

Figura 4. Relação entre glicose e demais colunas. Fonte: Autor

A partir do gráfico na Figura 4 é possível verificar uma frequência muito maior relacionada à glicose do que as outras características.

A grossura da pele corresponde à grande parte também, (Figura 5) a medida é diretamente influenciada pelo IMC, como mostrado anteriormente na tabela na Figura 3, pois quanto maior a quantidade de gordura no corpo, maior o IMC e a medida da dobra cutânea. A seguir o gráfico na Figura 6 comparando a intercorrência do IMC às outras características.

2.3. Algoritmos de classificação

Os algoritmos de classificação utilizados para teste de comparação de desempenho foram respectivamente o GBM (Gradient Boosting Machines), XGBoost (Extreme Gradient Boosting) e Random Forest.

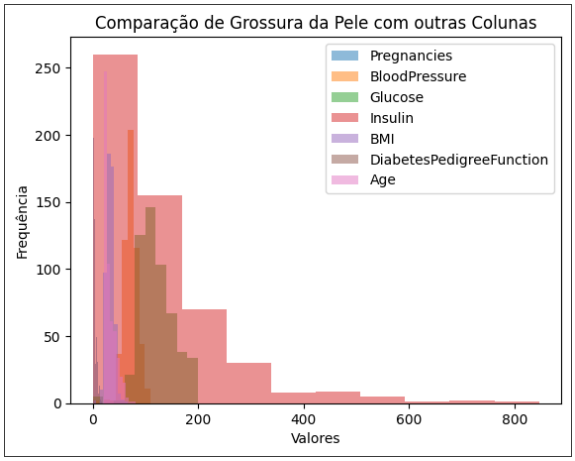

Figura 5. Relação entre a pele e demais colunas. Fonte: Autor

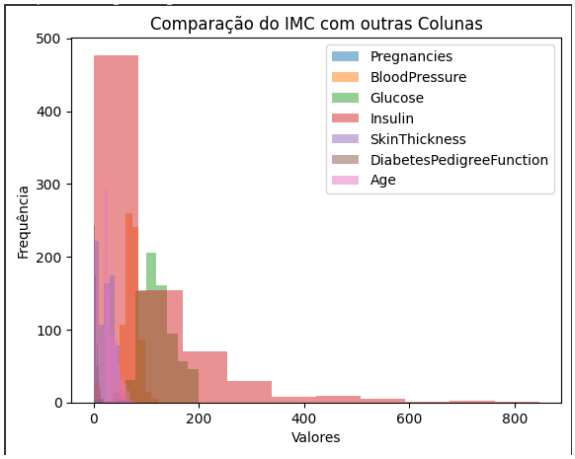

Figura 6. Relação entre IMC e demais colunas. Fonte: Autor

Tanto o XGBoost quanto o GBM são algoritmos de aprendizado de máquina baseados em boosting, que são usados para problemas de regressão e classificação. A principal diferença entre o XGBoost e o GBM está na forma como eles otimizam a função de perda durante o processo de boosting. [AnalythicsVidhia 2017]

Já o algoritmo Random Forest é baseado em decisão em árvores, combinando a saída em várias árvores para alcançar um resultado único. [IBM 2018]

Ao executar os testes por métricas em python, os seguintes resultados foram obtidos. (Veja :Figura 7)

Como é possível notar pelo gráfico acima na Figura 7, os valores das métricas dos algoritmos ficaram muito próximos.

Da mesmo forma como aconteceu com as outras métricas, a curva ROC não foi diferente, os algoritmos de Boosting tiveram o mesmo resultado enquanto o Random Forest teve uma leve diferença para melhor, como pode ser observado a seguir na Figura 8:

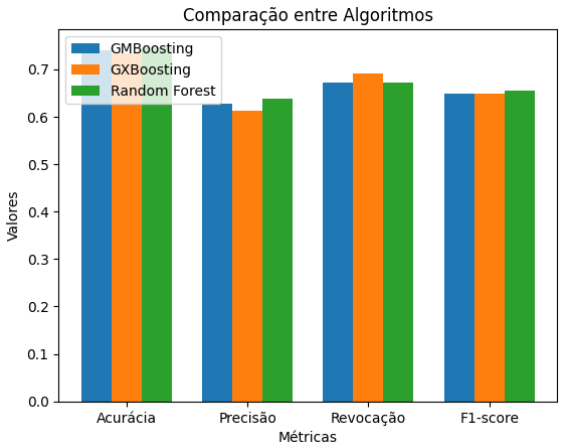

Figura 7. Comparação de desempenho. Fonte: Autor

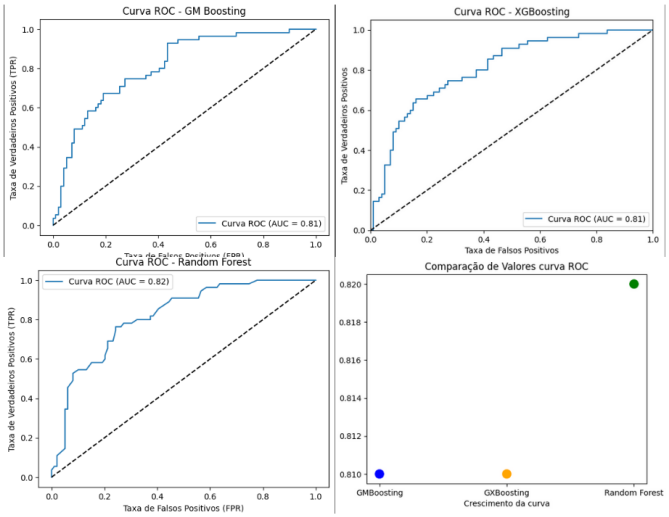

Figura 8. Curva ROC de cada algoritmo. Fonte: Autor

3. Agrupamento de dados

3.1. Banco de dados de vinhos

A base de dados escolhida para o teste de agrupamento foi sobre vinhos de três vinícolas diferentes na mesma região da Itália, o objetivo da base é diferenciar os vinhos a partir da quantidade de alguns elementos químicos e características de cor em cada um.

Para um uso mais simplificado do conjunto de dados, foram realizados alguns tratamentos, a base original continha apenas uma coluna com todos os valores separados por vírgula, como mostra a Figura 9.

1,14.23,1.71,2.43,15.6,127,2.8,3.06,.28,2.29,5.64,1.04,3.92,1065
1,13.2,1.78,2.14,11.2,100,2.65,2.76,.26,1.28,4.38,1.05,3.4,1050
1,13.16,2.36,2.67,18.6,101,2.8,3.24,.3,2.81,5.68,1.03,3.17,1185
1,14.37,1.95,2.5,16.8,113,3.85,3.49,.24,2.18,7.8,.86,3.45,1480
1,13.24,2.59,2.87,21,118,2.8,2.69,.39,1.82,4.32,1.04,2.93,735
1,14.2,1.76,2.45,15.2,112,3.27,3.39,.34,1.97,6.75,1.05,2.85,1450

Figura 9. Banco de dados de vinhos. Fonte: [Aeberhard and Forina 1991]

Primeiramente, as informações foram separadas em colunas, sendo 1 a classe e as outras 13 os valores encontrados de cada elemento dos vinhos. Na página web de onde o banco de dados foi retirado, foi possível verificar a descrição de cada tópico para cada coluna, sendo eles respectivamente: classe, teor alcoólico, melicácido, cinzas, alcalinidade das cinzas, magnésio, total de fenóis, flavonóides, não flavonóides fenóis, proantocianidinas, intensidade da cor, matiz, OD280 OD315 vinhos diluídos e prolina. Resultando em 14 colunas e 179 linhas, que pode ser conferido na Figura 10.

Class	Alcool	Melicacido	Cinzas	Alcalinidade_das_cinzas	Magnesio	Total_fenois	Flavonoides	Naoflavanoides_fenois	Proantocianinas	Intensidade_da_cor	Matiz	OD280_OD315_vinhos_diluídos	Prolina
1.0	14.23	1.71	2.43	15.6	127.0	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065.0
1.0	13.2	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050.0
1.0	13.16	2.36	2.67	18.6	101.0	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185.0

Figura 10. Banco de dados de vinhos com cabeçalho. Fonte: [Aeberhard and Forina 1991]

3.2. Impressões sobre as informações no banco de dados

Utilizando o método de agrupamento com o algoritmo K-means, foi possível separar as 3 classes de vinhos com base no teor alcoólico e intensidade de cor, segundo a [Miwa 2022], é possível observar a diferença de um vinho para outro a partir do teor alcoólico, intensidade da cor e cheiro. A base utilizada oferece uma coluna para o teor alcoólico e intensidade da cor, sendo esses os parâmetros escolhidos para o agrupamento das classes. Veja a Figura 11 seguir.

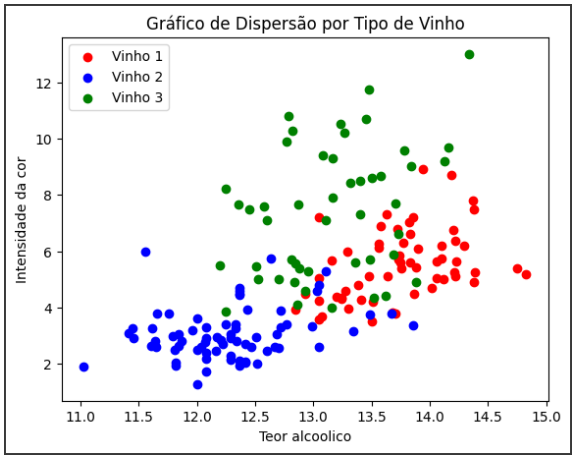

Figura 11. Gráfico de dispersão . Fonte: Autor

Com o gráfico de dispersão é possível observar que as características do Vinho 1 são alto teor alcoólico e uma intensidade média de cor, o Vinho 2 tem o menor teor alcoólico e intensidade de cor, por fim o Vinho 3 é o que tem a maior intensidade de cor e teor alcoólico mediano.

4. Conclusão

Após a execução de alguns testes de desempenho, obtendo resultados com pouca diferença, tendo alcançado valores próximos a 1, que é o valor ideal, conclui -se que para a base de dados utilizada os três algoritmos de classificação foram satisfatórios. Isso sugere que esses algoritmos são eficazes na tarefa de identificar e classificar corretamente as mulheres em relação à presença de diabetes. Além disso, foi possível observar a importância

de cada coluna presente na base, no momento do diagnóstico de diabetes. Atributos como idade, índice de massa corporal (IMC), medida cutânea e níveis de glicose mostraram-se relevantes para a detecção precisa da condição.

Ao utilizar a técnica de agrupamento para diferenciar os vinhos, a compreensão de cada coluna se mostrou crucial. Através da análise dos atributos relacionados ao teor alcoólico, intensidade de cor e outras características dos vinhos, foi possível identificar padrões e agrupar os vinhos de forma mais precisa. Essa abordagem permitiu uma melhor compreensão das nuances e diferenças entre os diferentes tipos de vinho, contribuindo para a seleção de um plano mais eficaz de diferenciação.

Em suma, os resultados obtidos com os algoritmos de classificação destacaram sua eficácia na tarefa de diagnóstico de diabetes, enquanto o agrupamento revelou a importância da análise criteriosa dos atributos na diferenciação dos vinhos. Esses achados reforçam a importância da utilização de técnicas de análise de dados para obter percepções valiosas e embasar decisões mais informadas em diferentes domínios de aplicação.

Referências

Aeberhard, S. and Forina, M. (1991). Wine. Disponível em: https://archive.ics.uci.edu/dataset/109/wine. Acesso em: 21 de junho de 2023.

AnalythicsVidhia (2017). Light GBM vc XGBOOST: Which algotithm takes the crown. Disponível em: https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/. Acesso em: 21 de junho de 2023.

de Letras, A. B. and Scliar, M. (2021). Diabetes: Aprendendo com os índios. Disponível em: https://www.academia.org.br/artigos/diabetes-aprendendo-com-os-indios. Acessado em: 20-06-2023.

do diabetes, C. (2018). Diabetes: sinais de alerta dados pe pele. Disponível em: https://clubedodiabetes.com/2018/08/diabetes-sinais-de-alerta-dados-pela-pele/. Acessado em: 21-06-2023.

IBM (2018). What is a Random Forest? Disponível em: https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems. Acesso em: 22 de junho de 2023.

Junior, E. E. R. (2016). Machine Learning - Métodos de classificação supervisionados. Disponível em: http://www.leg.ufpr.br/~eferreira/CE064/work5.html. Acessado em: 20-06-2023.

Learning, U. M. (2016). Pima Indians Diabetes Database. Disponível em: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download. Acesso em: 20 de junho de 2023.

Miwa, M. (2022). Estilos de vinho: quais são, o momento ideal de cada um e dicas para conhecer todos. Disponível em: https://revistaadega.uol.com.br/artigo/

diferentes-tipos-de-vinhos-e-suas-caracteristicas-qual-a-sua-preferic
2593.html. Acesso em: 21 de junho de 2023.

Santos, I. K. S. d. (2012). Manual de técnicas antropométricas. UCI Machine Learning Repository. Disponível em: http://www.fsp.usp.br/lanpop/wp-content/uploads/2019/01/suprailiaca_DC.pdf. Acessado em: 20-06-2023.

TJDFT (2021). Você sabe o que é a pré-diabetes? Entenda tudo sobre o assunto! Disponível em:https://www.tjdft.jus.br/informacoes/programas-projetos-e-acoes/pro-vida/dicas-de-saude/pilulas-de-saude/voce-sabe-o-que-e-a-pre-diabetes-entenda-tudo-sobre-o-assunto#:~:text=O%20estado%20de%20normalidade%20da,mg%2Fdl%20s%C3%A3o%20considerados%20diab%C3%A9ticos. Acessado em: 20-06-2023.