Geometria Computacional - TP 1

Bernardo do Nascimento Nunes - 2021031777

Indra Matsiendra Cardoso Dias Ribeiro - 2021031807

1. Introdução

O objetivo do trabalho prático em questão é colocar em prática os assuntos abordados na primeira etapa da matéria de Algoritmos 2. Neste sentido, é solicitado a implementação de diversos algoritmos de geometria computacional que foram vistos em sala. Nesse aspecto, para realizar o que foi pedido utilizamos a linguagem Python e algumas de suas bibliotecas ligadas a datasets e geração de gráficos, que facilitaram a instrumentalização de nosso estudo.

Essa documentação possui seções sobre: análise dos algoritmos utilizados, análise experimental com os gráficos gerados, conclusões e por fim nossas referências.

2. Análise dos algoritmos utilizados

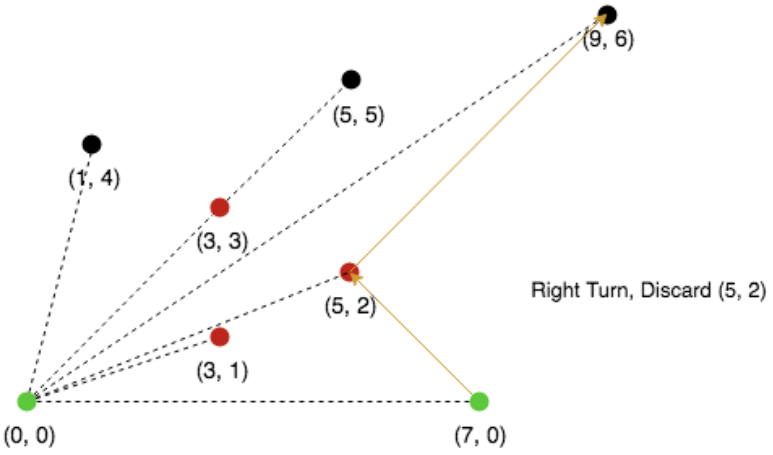
Nessa seção focaremos em descrever brevemente o funcionamento dos algoritmos solicitados e aqueles que julgamos necessários para nosso programa. Ademais, haverá também a análise de complexidade acerca de cada um.

2.1 Envoltória Convexa:

Como primeira tarefa implementamos o algoritmo da envoltória convexa, especificamente o Graham Scan.

A ideia principal por trás do algoritmo Graham Scan é ordenar os pontos com base em seu ângulo polar em relação a um pivô. O ponto de referência é escolhido como o ponto com a menor coordenada Y (e, em caso de empate, a menor coordenada X) entre todos os pontos. A partir desse ponto de referência, os outros pontos são classificados de acordo com seu ângulo polar.

Possui complexidade O(nlogn), onde n é o número de pontos passados. Uma vez que fazemos uma comparação entre todos os pontos não selecionados e o pivô.



2.2 Varredura Linear:

Inicialmente, nossa dupla se propôs a realizar o algoritmo explicado em sala de aula, que utiliza a árvore e possui complexidade ótima. Contudo, tivemos problemas na aplicação do mesmo e então decidimos implementar um com complexidade maior para prosseguirmos o trabalho e assim não atrasar o andamento do mesmo. Com isso, percebemos que os testes realizados por nós estavam sendo executados em um tempo bom. Logo, a dificuldade de implementação da árvore e o bom desempenho geral na execução dos testes optamos por implementar o algoritmo de força-bruta (todos contra todos) para verificar a interseção entre duas envoltórias convexas. Possui complexidade O(nm) em que n e m é a quantidade de pontos de cada envoltória.

2.3 Modelo:

O modelo de classificação desenvolvido neste trabalho é constituído por 5 variáveis globais, que representam as envoltórias convexas de cada classe, os pontos mais próximos das duas envoltórias e qual a classe predominante.

Ao fazer a classificação de um ponto p qualquer o modelo leva em conta a reta perpendicular entre os pontos mais próximos das duas envoltórias, ou seja, se p está mais perto de p1 ele pertence à classe 1, se está mais perto de p2 pertence à classe 2 e se está em cima da reta optamos por classificá-lo como parte da classe predominante(aquela com mais ocorrências).

Ou seja, para cada ponto de entrada, calcula a distância até os dois pontos mais próximos das envoltórias e atribui o rótulo da classe com o ponto mais próximo.

Os algoritmos dessa etapa do trabalho possuem de modelo geral complexidade constante, sendo o único que foge dessa regra o que calcula a menor distância entre dois pontos de envoltórias distintas, que possui complexidade quadrática, uma vez que faz um todos-contra-todos.

2.3 Computação das métricas:

A função de computação das métricas recebe dois argumentos: y_test e y_pred. O argumento y_test é um array que contém os valores reais das classes das amostras de teste. O argumento y_pred é um array que contém os valores das classes previstas pelo modelo para as amostras de teste. Analisamos as seguintes métricas:

- Precisão é a proporção de amostras que foram classificadas como positivas pelo modelo que são de fato positivas. É uma métrica importante para problemas em que é importante evitar classificar amostras negativas como positivas.
- Acurácia é a proporção de amostras que foram classificadas corretamente pelo modelo. É a métrica mais simples de avaliação, mas pode não ser a mais adequada em todos os casos.
- Revocação é a proporção de amostras positivas que foram corretamente classificadas pelo modelo. É uma métrica importante para problemas em que é importante não classificar amostras positivas como negativas.
- F1 é uma média harmônica da precisão e da revocação. É uma métrica que tenta equilibrar as duas métricas.

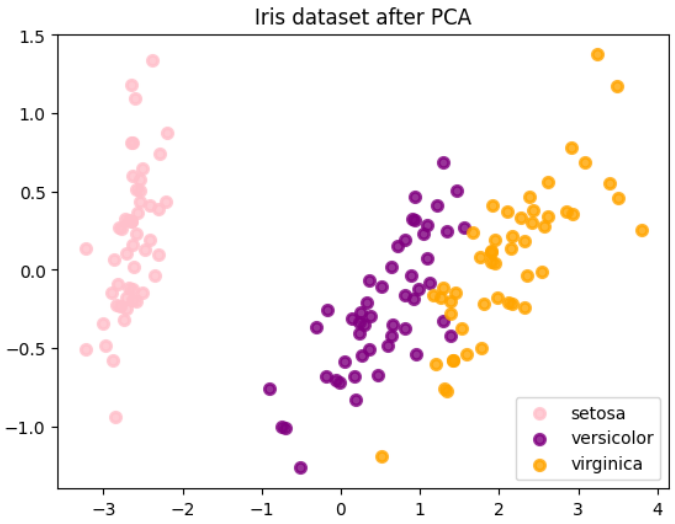
Para achar as métricas citadas usamos o sklearn.metrics.

3. Experimentos

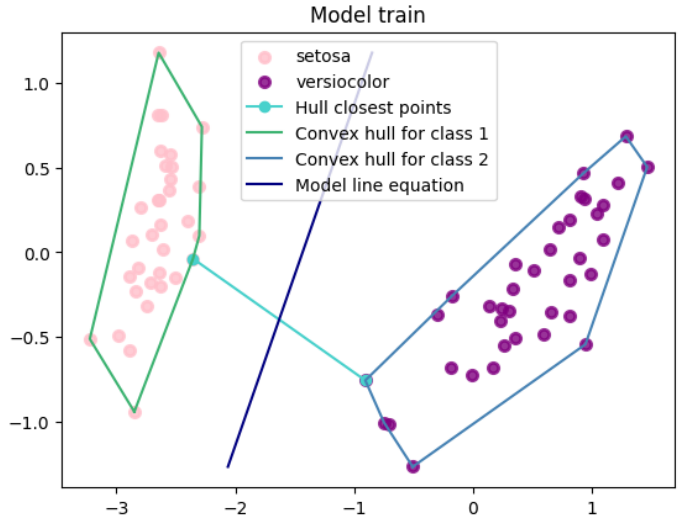
Nesta seção iremos apresentar os resultados obtidos com os dados escolhidos, discorrendo e pontuando aquilo que julgamos necessário acerca de cada experimento. Ademais, faremos uma breve descrição de cada um.

3.1 Iris:

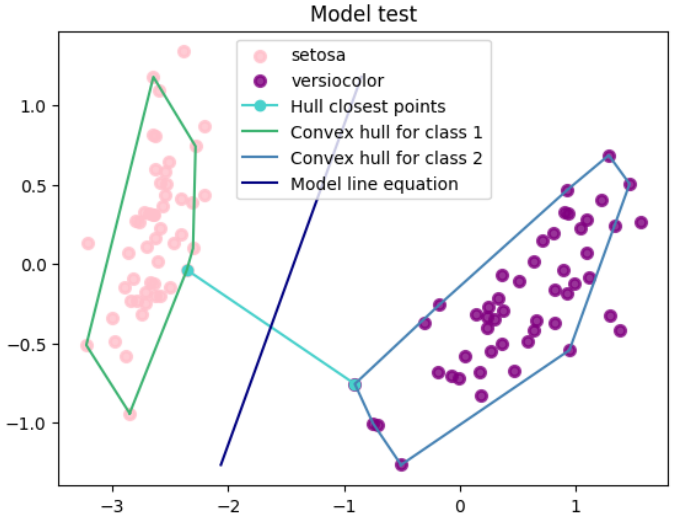
Começamos os testes por um dos datasets mais famosos ao se tratar de testes de algoritmos de classificação. O dataset contém medições de quatro características de flores de íris: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala. As flores são classificadas em três espécies: Iris setosa, Iris versicolor e Iris virginica.



Utilizamos o mesmo padrão em todos nossos datasets. Inicialmente nós plotamos as classes após a realização do PCA. Reduzindo assim a dimensionalidade do dataset e sendo possível plotar em um gráfico 2d.



Após isso, fazemos o plot do gráfico, com as envoltórias convexas de cada classe selecionada, a reta da menor distância entre pontos de cada envoltória e a reta perpendicular. Em alguns gráficos a reta perpendicular pode parecer torta, mas isso só ocorre pela diferença de proporção entre o eixo y e x em alguns datasets. Ademais, perceba que a envoltória é já gerada apenas com os pontos de treino, por isso pode haver a ausência de alguns pontos do gráfico do pca para esse.



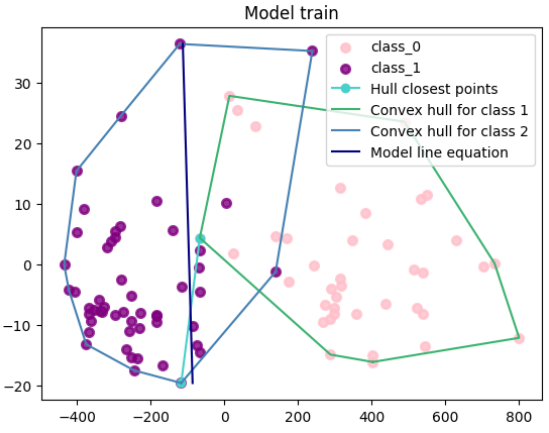
No model test realizamos a avaliação dos pontos e exibição dos mesmos após classificação,

Resultados:
Acurácia: 1.0 Precisão: 1.0
Revocação: 1.0 F1: 1.0

Os resultados eram esperados, uma vez que as as classes escolhidas não possuem interseção entre suas envoltórias. Ademais, os pontos estão distribuídos relativamente distantes e ficam bem separados pela reta do modelo.

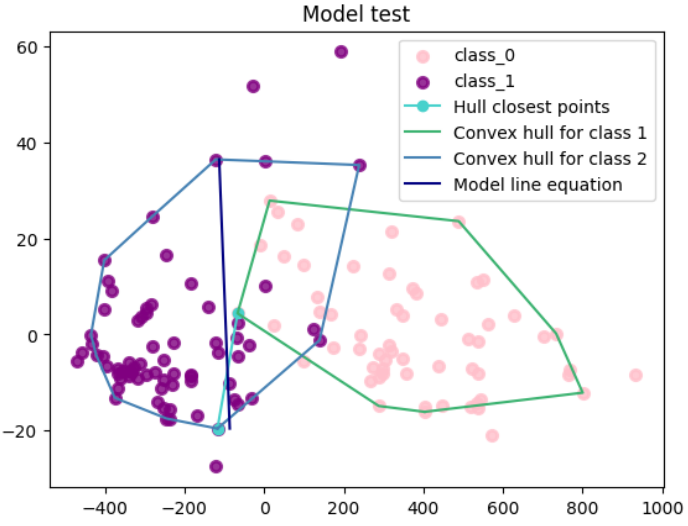
3.2 Wine:

O dataset contém medições de 13 características químicas de vinhos produzidos na Itália. Os vinhos são classificados em três classes: Class 1, Class 2 e Class 3.



Esse dataset exemplifica e demonstra uma decisão de projeto nosso, uma vez que há a interseção entre as envoltórias contudo o modelo ainda é aplicado. Nesse sentido, é lançado um erro com a seguinte mensagem: Watch out! There is an intersection between classes, the model won't work as expected!

Nesse aspecto, decidimos fazer dessa forma para poder elaborar mais análises acerca dos datasets, uma vez que é relativamente complicado achar classes que possuem suas envoltórias sem nenhuma interseção.



Resultado:

Acurácia: 0.8461538461538461

Precisão: 0.8816568047337279

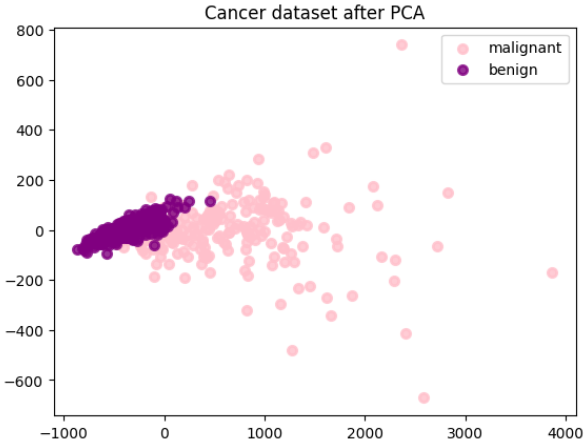
Revocação: 0.8461538461538461

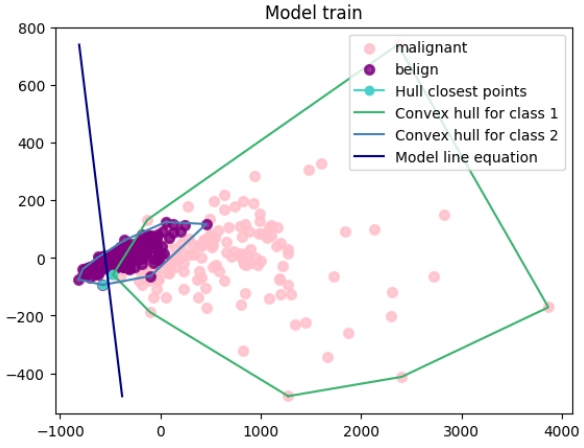
F1: 0.8417642140468229

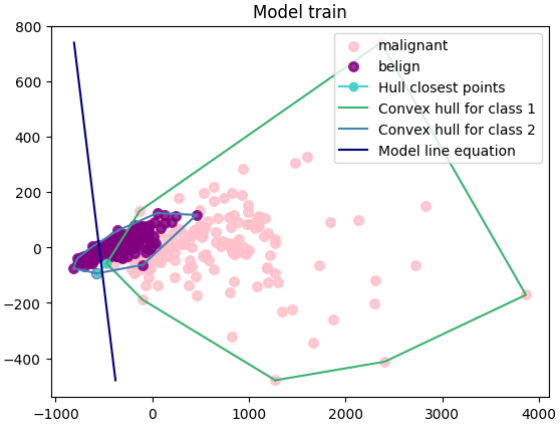
Podemos observar que ocorre uma diminuição do valor das métricas (esperado com a interseção), mas se mantém ainda um bom resultado uma vez que há uma considerável separação.

3.3 Cancer:

O dataset contém dados de pacientes com câncer de mama, incluindo características como tamanho do tumor, forma do núcleo e textura da borda. As pacientes são classificadas como benignas ou malignas.







Acurácia: 0.52046783625731

Precisão: 0.7916515426497278

Revocação: 0.52046783625731

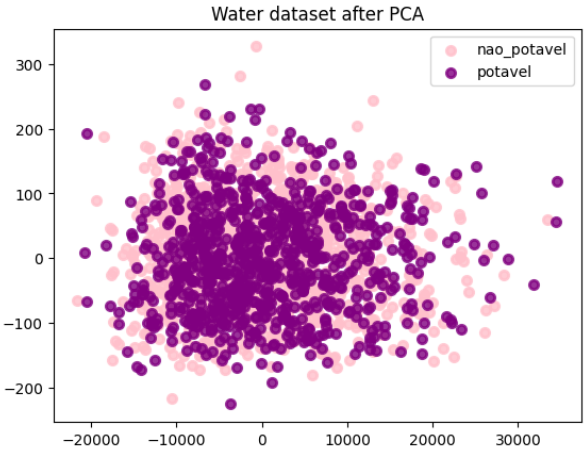
F1: 0.4682684754365822

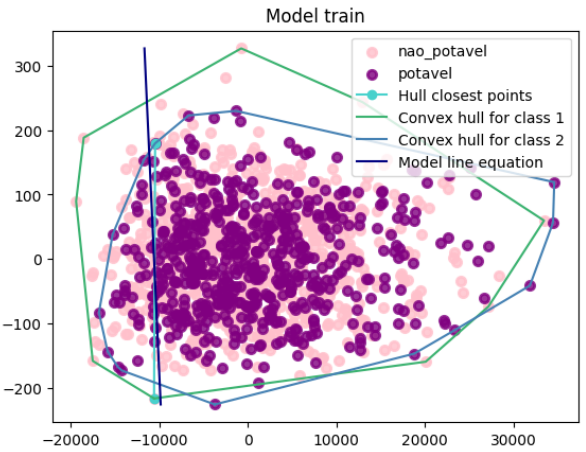
Neste casos a envoltórias da classe belign fica boa parte dentro da malignant. Isso Ademais é possível perceber uma concentração maior da classe belign na esquerda do gráfico, sendo perceptível um

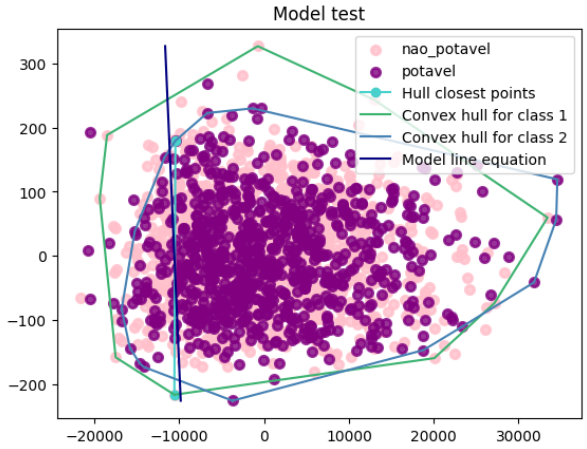
padrão, que não é captado pelo modelo. Nesse aspecto, acredito que ele falha bastante e outros modelos como o de fazer a classificação pelos k próximos pontos (k sendo um número positivo ímpar) funcionaria bem melhor.

3.4 Water Potability:

O dataset contém dados de qualidade da água de várias fontes, incluindo poços, rios e lagos. As amostras são classificadas como potável ou não potável.







Durante a procura dos datasets pensamos que nesse haveria uma maior separabilidade dos dados, contudo fomos surpreendidos inclusive com envoltórias quase iguais. O que obviamente fez ter os piores resultados até então.

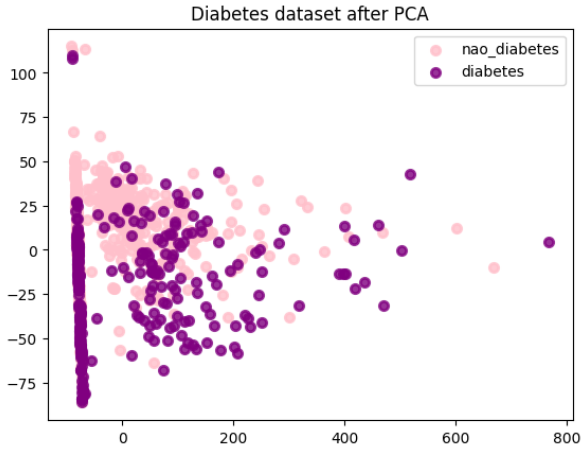
Acurácia: 0.43211920529801323

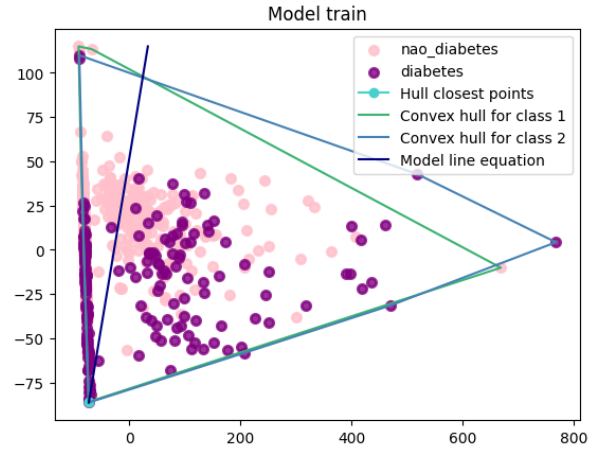
Precisão: 0.519220244665195

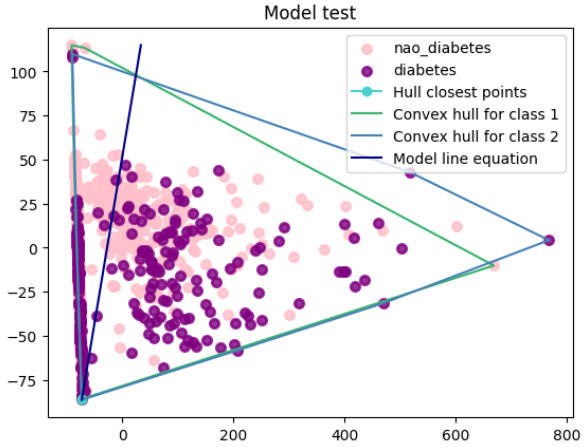
Revocação: 0.43211920529801323

F1: 0.33964281428077003

3.5 Diabetes:







O dataset em questão avalia duas classes: pessoas com diabetes ou não. Ela apresenta basicamente o mesmo formato das envoltórias, similarmente ao último dataset, o que faz ambos terem resultados parecidos.

Acurácia: 0.5930735930735931

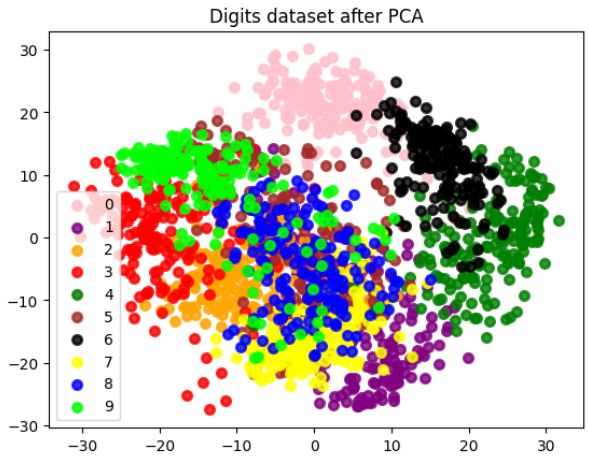
Precisão: 0.6076773815278663

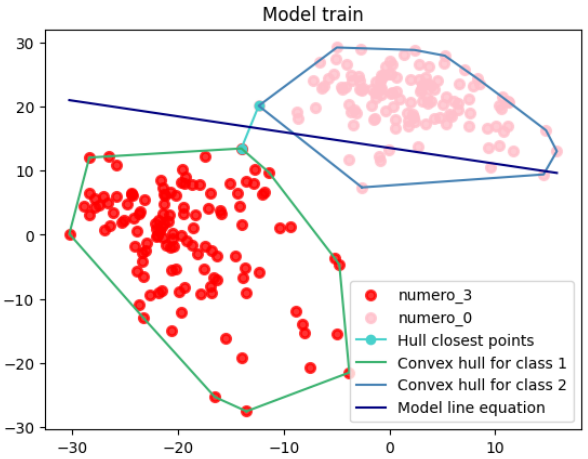
Revocação: 0.5930735930735931

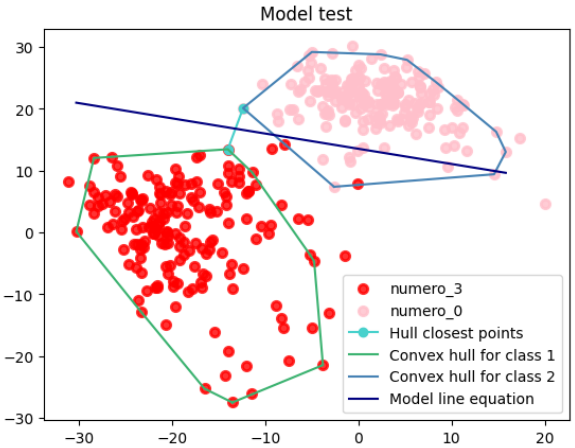
F1: 0.5988495332520994

3.6 Digits:

O dataset contém imagens de dígitos manuscritos, de 0 a 9. As imagens são representadas como matrizes de 8x8 pixels, com cada pixel contendo um valor de intensidade de 0 a 16. Trabalhar com esse dataset foi bem interessante, porque é intuitivo para a gente números que possuem semelhança ou não, possuindo envoltórias convexas que se interceptam ou não, respectivamente.







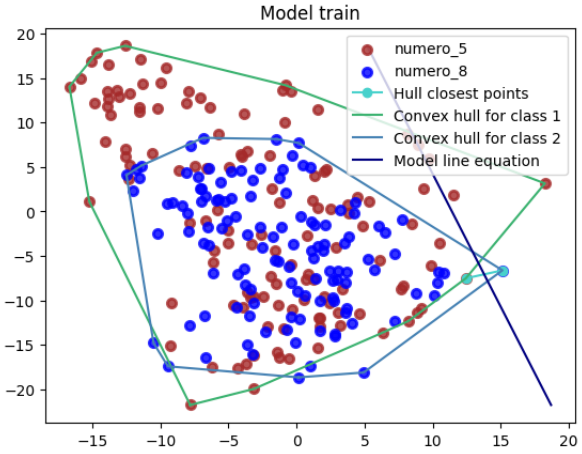
Inicialmente começamos pelos dígitos 3 e 0 que possuem poucas semelhanças, o que fica perceptível na construção do modelo. O que gera resultados melhores.

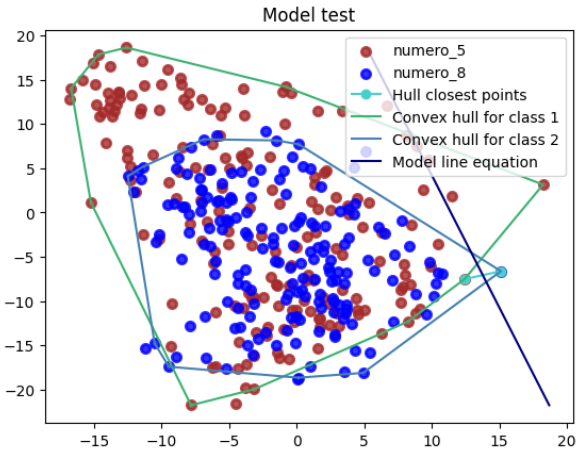
Acurácia: 0.9908256880733946

Precisão: 0.9909866409142122

Revocação: 0.9908256880733946

F1: 0.9908225951743042





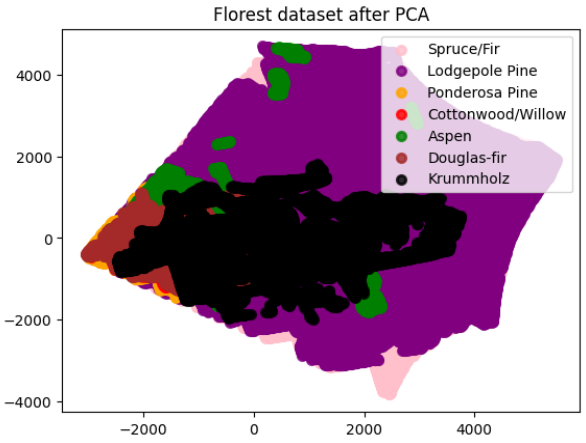
Os outros dois números escolhidos foram 5 e 8, uma vez que se assemelha muito, e o resultado não poderia ser outro, as envoltórias se interceptam.

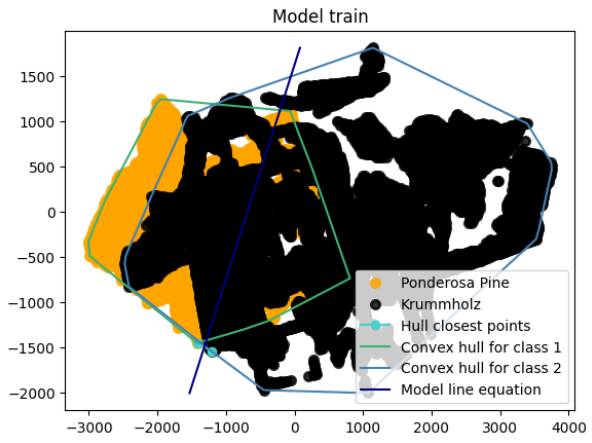
Acurácia: 0.51401869158878 Precisão: 0.2642152153026465

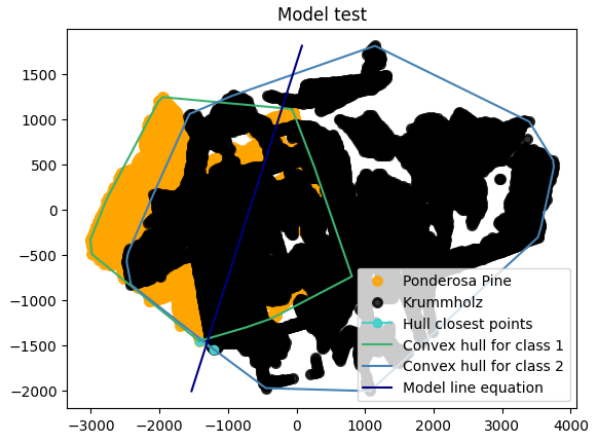
Revocação: 0.514018691588785 F1: 0.34902503749855773

3.7 Cover Type:

O dataset contém dados de florestas nos Estados Unidos, incluindo informações sobre a composição das árvores, a elevação e o solo. A variável alvo é o tipo de cobertura florestal, que pode ser uma das sete classes: Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, ou Krummholz.







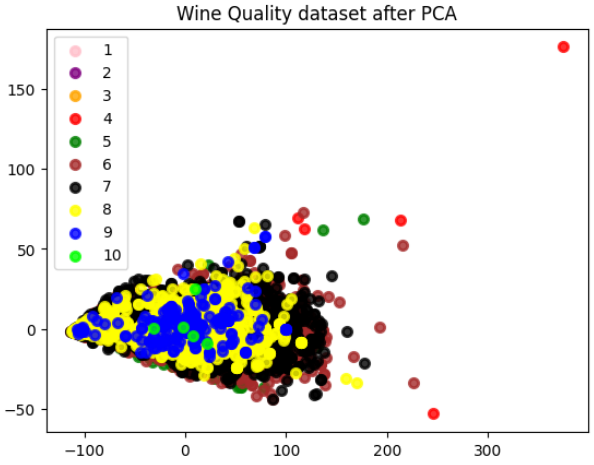
Esse foi o maior dataset usado nos nossos experimentos, e logo o que demorou mais demorou a ser processado. Escolhemos os dois tipos que menos possuem interseção, para gerar resultados melhores, e foi bem considerando a interseção das envoltórias convexas.

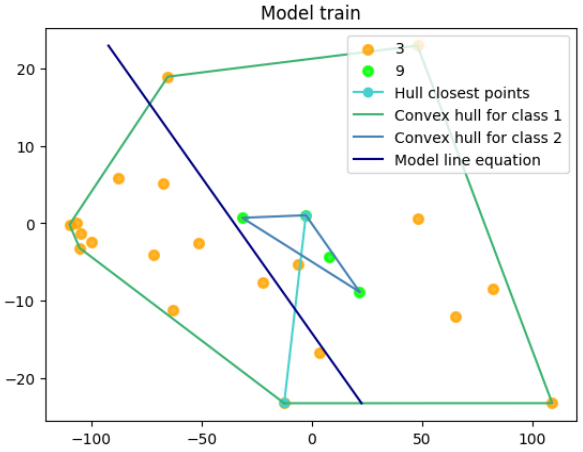
Acurácia: 0.8473431668740004 Precisão: 0.8458668234011368

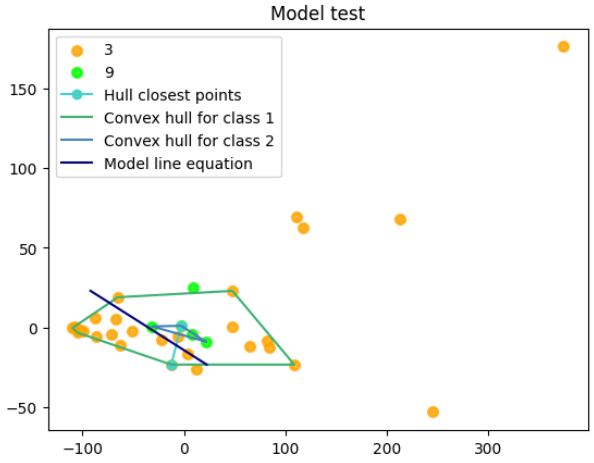
Revocação: 0.8473431668740004 F1: 0.8456079392656397

3.8 Wine Quality:

O dataset contém dados de vinhos brancos e tintos produzidos em Portugal. As características incluem atributos físicos e químicos, como teor de álcool, acidez volátil, ácido cítrico, açúcar residual, cloretos, dióxido de enxofre livre e total, densidade, pH, sulfatos e potássio. A variável alvo é uma medida da qualidade do vinho, avaliada por especialistas em uma escala de 0 a 10.







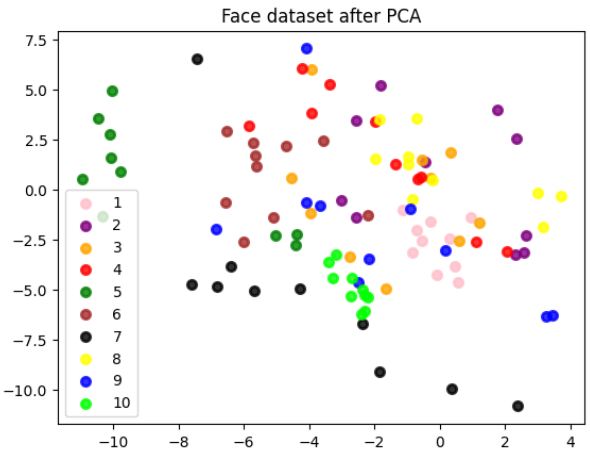
Nesse método identificamos uma particularidade, envoltórias contidas não possuem interseção entre seus segmentos e com isso não será gerada a mensagem de erro. Contudo, é igualmente negativo.

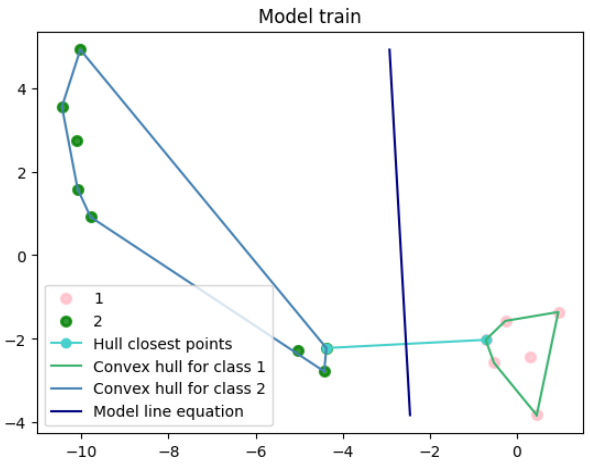
Acurácia: 0.4545454545454545
Precisão: 0.922077922077922

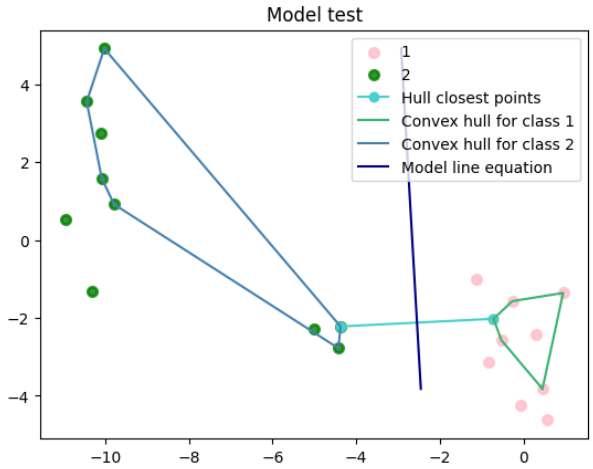
Revocação:0.45454545454545453
F1: 0.5422077922077922

3.9 Olivetti Faces:

O dataset contém imagens de 40 pessoas, com 10 imagens por pessoa. As imagens são representadas como matrizes de 64x64 pixels, com cada pixel contendo um valor de intensidade de 0 a 255. No nosso experimento usamos apenas 10 pessoas.







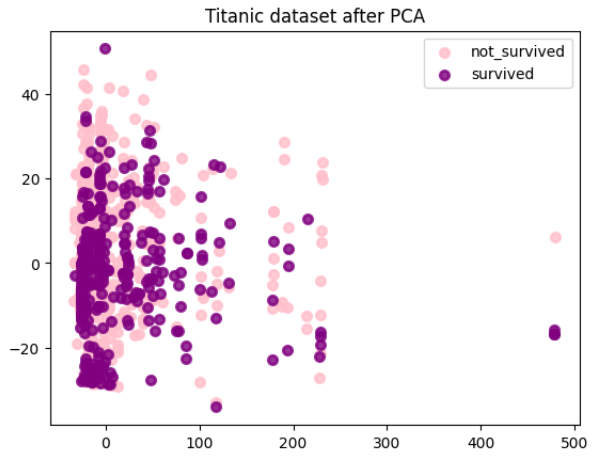
Mais um dataset em que a separabilidade completa das envoltórias convexas escolhidas foi revertida em ótimos resultados.

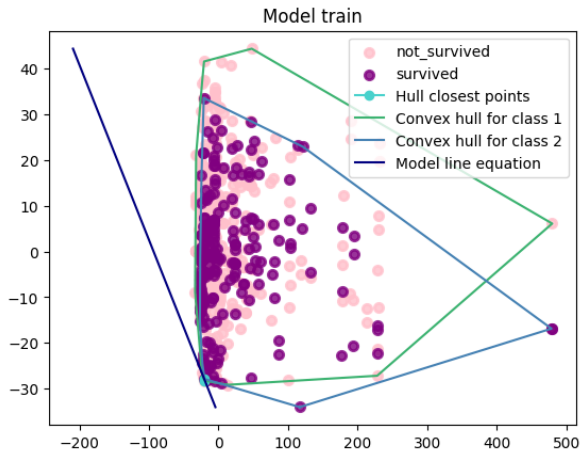
Acurácia: 1.0 Precisão: 1.0

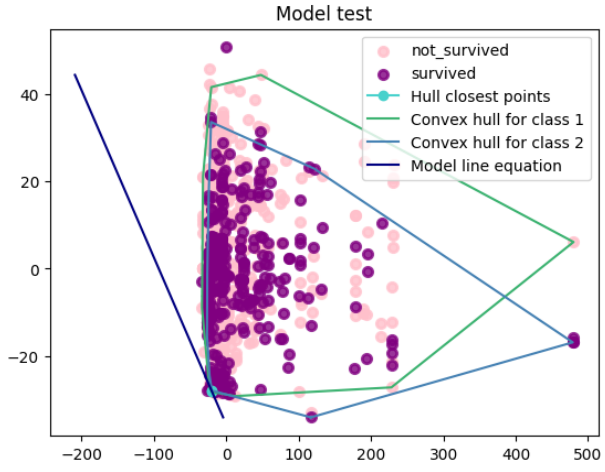
Revocação: 1.0 F1: 1.0

3.10 Titanic:

O dataset contém dados de passageiros a bordo do Titanic. As características incluem informações como nome, sexo, idade e classe social. A variável alvo é a sobrevivência, que pode ser "sobreviveu" ou "não sobreviveu".







Esse dataset foi de longe o com pior resultado apresentado, o motivo fica evidente no gráfico, a reta do modelo foi traçada basicamente deixando um lado com todos os pontos e o outro sem nada. Isso é refletido nos resultados.

Acurácia: 0.2926208651399491

Precisão: 0.32122398381940365

Revocação: 0.2926208651399491

F1: 0.13803395962373724

4. Conclusão

O trabalho prático de geometria computacional foi um grande desafio e uma grande oportunidade de aprendizado. Pudemos aplicar os conceitos vistos em sala de aula e desenvolver nossos próprios algoritmos.

Em relação aos resultados, podemos observar que o modelo apresentou um bom desempenho em datasets com envoltórias convexas bem separadas. No entanto, o desempenho diminuiu em datasets com envoltórias convexas com interseções ou com envoltórias convexas muito semelhantes, o que já era de se esperar.