Relatório de implementação do método KNN com Grid Search e K-Fold Cross-Validation e do método DMC

Angélica Alves Viana

Abril 2019

PPGCC - Aprendizagem de Máquina
Matrícula: 20191018030156

1 Introdução

Este relatório visa descrever a implementação do método KNN (K Near Neiboarhood) e do método DMC(Distância Mínima aos Centroids) os resultados de aplicação dos mesmos as bases da Iris (150 registros de plantas : 50 Iris-setosa, 50 Iris-versicolor e 50 Iris-Virginica), da Coluna Vertebral (310 registros: 100 Normal, 60 Disk Hernia e 150 Spondylolisthesis) e uma base artificial gerada a partir do padrão AND.

2 Desenvolvimento

2.1 Implementação do KNN

A implementação do método KNN foi realizada com base no algoritmo descrito abaixo.

Algorithm 1 Algoritmo K-Vizinhos Mais Próximos (K-NN).

1: Armazene os exemplos em uma tabela;
2: Encontrar na tabela os K vetores mais próximos de X_{new} ;
3: Seja C_k a classe a que pertence a maioria dos K vetores;
4: Avalia os indivíduos;
5: Atribuir a X_{new} a classe da maioria dos K vetores, ou seja: $Classe(X_{new}) = C_k$
6: Se a classificação for correta incluir X_{new} na tabela;

O mesmo foi desenvolvido utilizando a versão 2.7 do Python e cada etapa de implementação do algoritmo é descrita a seguir:

- **Carregamento dos Dados:** Os dados foram carregados e armazenados utilizando as funções *read_csv*() e *read_table*() disponíveis no módulo do pandas. A função *load_data*() implementada recebe o *dataset* e separa em duas matrizes X e Y e converte os valores de Y, que são strings representando as classes, para classes representadas por números inteiros(0, 1 e 2)

- **Normalização dos dados:** Devido as escalas diferentes de grandezas das variáveis independentes presentes nas bases de dados, foi necessário realizar a normalização dos dados de modo que os valores para cada coluna de X contivesse apenas valores pertencentes ao intervalo [0,1]. A normalização por reescala descrita pela fórmula (1) foi aplicada aos dados.

$$x_{ij} = \frac{x_{ij} - min_j}{max_j - min_j} \tag{1}$$

- **Treinamento do Algoritmo:** O treinamento do algoritmo consistiu no simples armazenamento dos vetores de características dos dados de treinamento e suas respectivas classes em uma matriz *train*.

- **Avaliação do Algoritmo:** Foi utilizada metodologia Grid search com k-folds cross validation. Para isso foram feitas 20 realizações de treinamento, sendo 80% dos dados separados para treinamento e os 20% restantes para os testes. Desse modo, 20 modelos foram criados aleatoriamente, para cada conjunto de treinamento, os mesmos foram separados em 5-folds e em um *loop* de tamanho 5, a cada iteração um fold dentre os 4 foi escolhido para realizar a validação do modelo com o k referente aos vizinhos variando de 1 à 24. Para cada k referente aos vizinhos, foi calcula a acurácia do modelo e por fim o k que resultou no melhor valor de acurácia foi utilizado para testar o modelo.

  - **Função** *predict* : Calcula as distâncias euclidianas de um vetor x_i para todos os vetores linhas pertencentes ao X de treinamento, isso resulta em um vetor de distâncias euclidianas do tamanho da quantidade de linha de X. Dentre as k menores distâncias encontradas, a classe do vetor do X de treinamento associada a distância mais frequente é atribuída ao x_i
  - **Função** *accuracy_metric* : Calcula a taxa de acerto do algoritmo, que consiste na quantidade de acertos das predições pela quantidade total de dados de teste.

2.2 Resultados - KNN - Base da Iris

Para a base da iris, dos 150 padrões, foram separados 120(80%) para treinamento e 30(20%) para teste. Em cada realização, foram utilizados 24 valores para o parâmetro k do KNN referente a quantidade de vizinhos, com o objetivo de encontrar àquele que obtivesse o maior valor de acurácia no treinamento

para então utilizá-lo no teste. Desse modo, 5 combinações de treino e teste na validação foram feitas, com testes contendo 24 (1/5) padrões e treinamentos com 96 (4/5). Algumas medidas para avaliar a performance do modelo foram retiradas após as 20 realizações e podem ser vistas na tabela abaixo.

	Treinamento	Teste
Acurácia	$97,33$	$97,33$
Desvio Padrão	$0,77$	$3,05$
Tempo Médio (s)	$0,0000$	$0,0974$

O gráfico de dispersão abaixo mostra a relação entre os valores de k e as acurácias de treinamento obtidas nas 20 realizações para a base de dados da coluna vertebral. Podemos observar que os valores mais recorrentes se encontram no intervalo de [3,7], ou seja, para uma quantidade de vizinhos pequena, o algoritmo obtém melhores valores de acurácia para a base de dados da iris.

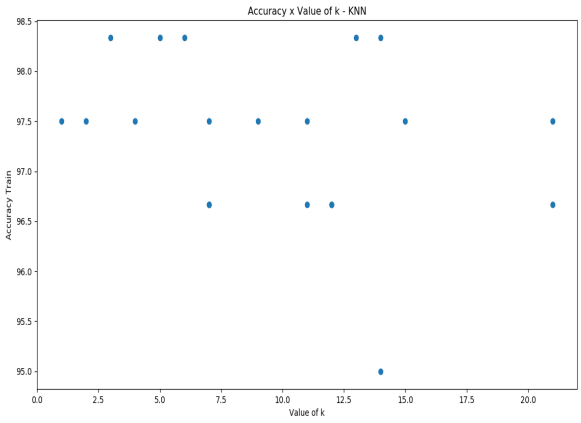

Figure 1: Acurácia de Treinamento x Valor de K

As superfícies de decisão do algoritmo, podem ser observadas a partir da representação gráfica dos atributos plotados dois a dois, A Figura 2 mostra as superfícies de decisão referentes a uma das 20 realizações que obteve o valor de acurácia mais próximo da média das acurácias nas 20 realizações utilizando-se os atributos de comprimento e largura das sépalas nos eixos x e y respectivamente, em que os dados distribuídos na forma de x, + e * são os dados de teste e representam respectivamente dados das classes Iris-setosa (Vermelho), Iris-versicolor

(Azul) e Iris-virgínica (Verde), já as bolinhas pequenas representam os dados de treinamento para cada classe.

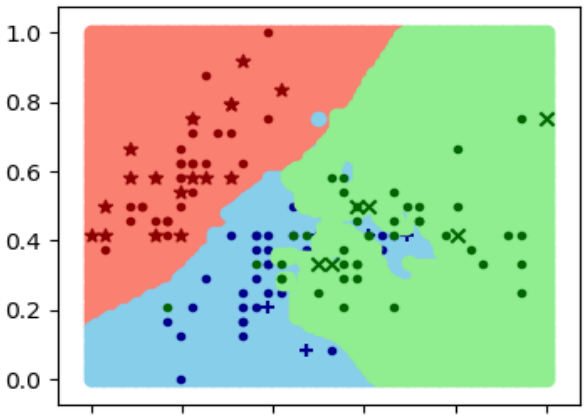

Figure 2: Superfície de Decisão - Base da Iris (KNN)

A matriz de confusão de teste de uma das realizações que obteve acurácia muito próxima a média das acurácias pode ser vista abaixo, cuja acurácia é de $\approx 96,67\%$.

	0	1	2
0	11	0	0
1	0	10	1
2	0	0	8

A matriz de confusão acima representa que nessa realização haviam 11 padrões da classe 0 (Iris-setosa) e todos foram classificados corretamente. além disso haviam 11 padrões da classe 1 (Iris-versicolor), dos quais 10 foram classificados corretamente e apenas 1 foi classificado incorretamente como sendo da classe 2. Já para os padrões da classe 2 (Iris-virgínica), haviam 8 padrões dessa classe e todos foram classificados corretamente. Desse modo a taxa de acerto nessa realização para o teste foi de $(11+10+8)/(11+11+8) = 29/30 \approx 96,67\%$.

2.3 Resultados - KNN - Base da Coluna Vertebral

Para a base da coluna vertebral, dos 310 padrões, foram separados 248(80%) para treinamento e 62(20%) para teste. Em cada realização, novamente foram

utilizados 24 valores para o parâmetro k do KNN referente a quantidade de vizinhos. Desse modo, 5 combinações de treino e teste na validação foram feitas, com testes contendo 61 (1/5) padrões e treinamentos com 244 (4/5). É importante destacar que algumas amostras no procedimento de validação foram desconsideradas para facilitar o procedimento de divisão dos folds na validação cruzada. Algumas medidas para avaliar a performance do modelo foram retiradas após as 20 realizações e podem ser vistas na tabela abaixo.

	Treinamento	Teste
Acurácia	$78,90$	$78,90$
Desvio Padrão	$1,10$	$4,37$
Tempo Médio (s)	$0,0000$	0.4690

O gráfico de dispersão abaixo mostra a relação entre os valores de k e as acurácias de treinamento obtidas nas 20 realizações. Podemos observar a partir de uma breve análise que os valores mais recorrentes se encontram no intervalo de [12,22] e com acurácias mais altas.

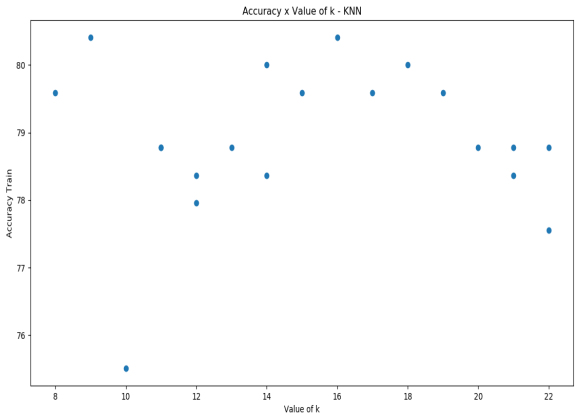

Figure 3: Acurácia de Treinamento x Valor de K

As superfícies de decisão do algoritmo podem ser observadas a partir da representação gráfica dos atributos plotados dois a dois, A Figura 4 mostra a superfície de decisão referente aos dois primeiros atributos da base de uma das 20 realizações que obteve o valor de acurácia mais próximo da média das acurácias, em que os dados distribuídos em formato de x, + e * novamente

são os dados de teste e representam respectivamente dados das classes Hérnia de Disco (Vermelho), Spondylolisthesis (Azul) e Normal (Verde) e os pontos pequenos são dados do treinamento dessa realização.

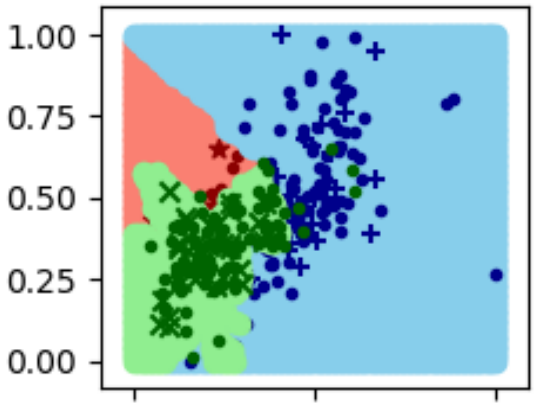

Figure 4: Superfície de Decisão - Base Coluna Vertebral (KNN)

A matriz de confusão de teste para uma das realizações que obteve acurácia muito próxima a média das acurácias das 20 realizações pode ser vista abaixo, cuja acurácia é de $\approx 78,70\%$.

	0	1	2
0	8	0	3
1	0	27	4
2	3	3	13

Essa matriz de confusão representa que de um conjunto de teste de 61 padrões, haviam 11 da classe 0 (Hernia de Disco), dos quais 8 foram preditos corretamente e 3 foram erroneamente classificados como sendo da classe 2. Haviam também 31 padrões da classe 1 (Spondylolisthesis), dos quais foram preditos corretamente 27 e 4 foram preditos como sendo da classe 2. Já para os padrões da classe 2 (Normal) haviam 19 registros, dos quais 13 foram classificados corretamente, 3 foram classificados como sendo da classe 0 e outros 3 classificados como sendo da classe 1. Obtendo-se, assim, uma taxa de acerto de $(8+27+13)/(11+31+19) = 48/61 \approx 78,70\%$.

2.4 Resultados - KNN - Base Artificial

Para validar se o algoritmo foi implementado corretamente, foi gerada uma base de dados artificial com pontos pontos aleatórios na vizinhança dos pontos (0;0), (0,1) , (1;0) e (1;1). De modo que todos os pontos ao redor de (1;1) pertencessem a classe 1 e os demais pontos a classe 0. Foram gerados 150 pontos, dos quais 30 eram da classe 0 divididos proporcionalmente entre os pontos de referência da classe e 10 da classe 1. os dados foram gerados de modo que estivessem a no máximo uma distância de 0.1 para mais ou para menos tanto para o primeiro atributo, como para o segundo. A representação gráfica desses dados no plano 2D pode ser vista no gráfico a seguir.

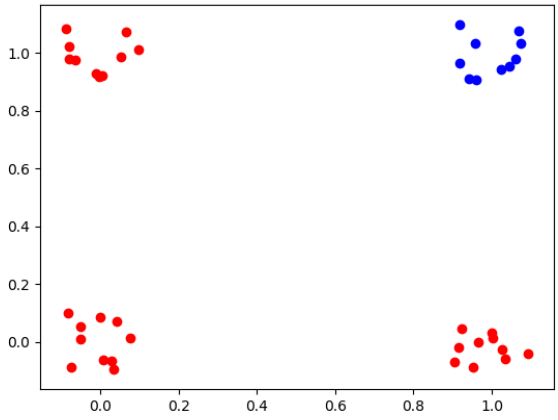

Figure 5: Dados artificiais gerados para teste.

Ao final das 20 realizações aplicando-se o KNN à essa base artificial, foram obtidas as seguintes medidas de avaliação.

	Treinamento	Teste
Acurácia	$100,00$	$100,00$
Desvio Padrão	$0,00$	$0,00$
Tempo Médio (s)	$0,0000$	$0,0118$

Para valores de k variando de [1,7] nas 20 realizações, o algoritmo obteve 100% de acurácia de teste. O melhor valor de k para todas as realizações foi fixado como sendo o primeiro dentre os 22, portanto k = 1.

A superfície de decisão do algoritmo pode ser observada a partir da representação gráfica dos atributos plotados no eixo x e no eixo y, A Figura 4 mostra a superfície de decisão do classificador referente a uma das 20 realizações que obteve o valor de acurácia mais próximo da média das acurácias, em que os dados distribuídos em formato de x e + representam os dados de teste das classes 0 e 1 respectivamente e os pontos pequenos representamm os dados de treinamento.

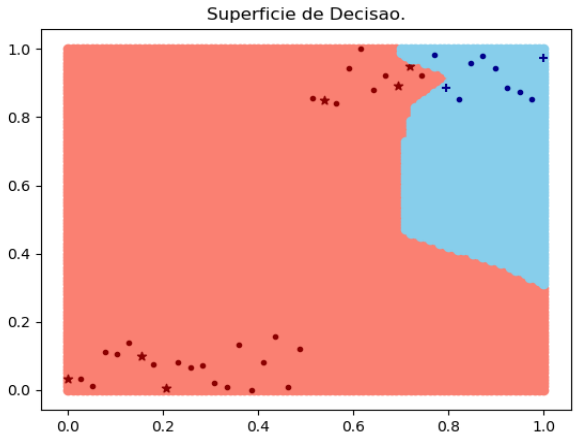

Figure 6: Superfície de Decisão - Base Artificial (KNN)

Um exemplo de uma matriz de confusão de uma das 20 realizações de teste é mostrada abaixo

	0	1
0	6	0
1	0	2

Essa matriz de confusão representa que de um conjunto de teste de 40 padrões, haviam 6 da classe 0 e 2 da classe 1 e todos foram classificados corretamente, obtendo-se então uma acurácia de 100%.

2.5 Implementação do DMC

A implementação do DMC foi baseada no seguinte algoritmo

Algorithm 2 Algoritmo de Distância Mínima aos Centroids (DMC).

1: Separe o conjunto de dados em treinamento e teste;
2: Calcule as médias(centroids) por classe nos dados de treinamento;
3: Calcule as distâncias dos padrões de teste para os centroids;
4: Atribua a classe representada pela centroid que gerou a menor distância para o dado de teste;
5: Calcule a acurácia do modelo;

Foram feitas 20 realizações de treinamento e teste com o DMC, na qual para cada realização foram gerados conjuntos embaralhados de dados de treino (80%) e teste (20%). Os procedimentos de carregamento dos dados e normalização foram idênticos aos utilizados para o método KNN, já o treinamento e o teste do algoritmo nas 20 realizações são descritos abaixo

- **Treinamento:** Para cada realização, um conjunto com 80% dos dados foi utilizado para realizar o treinamento. O treinamento consistiu em separar esses dados por classe e calcular as médias de cada classe, correspondendo aos centroids que representam as classes. Em ambas as bases de dados, foram obtidas 3 centroids, cada uma representando uma classe do conjunto de dados.

- **Teste:** Para cada realização, um conjunto contendo 20% dos dados foi utilizado para testar a capacidade de generalização do modelo treinado. Para cada padrão de teste, foram computadas 3 distâncias, referentes as distâncias do dado para os centroids. A classe representada pela centroid que apresentou a menor distância para o padrão de teste foi atribuída ao mesmo.

- **Acurácia:** Foram calculadas as taxas de acerto referentes a cada teste e por fim calculou-se a média dessas taxas de acerto com o objetivo de calcular a acurácia de teste do modelo.

2.6 Resultados - DMC - Base da Iris

Para a base de dados da Iris, para cada realização o conjunto de dados foi dividido em treinamento (120) e teste (30), foram encotradas as centroids para cada classe do conjunto de treinamento e por fim, o teste foi realizado ccalculando-se as distâncias de cada padrão de teste ao valor dos centroids, atribuindo a classe representada pela centroid que obteve o menor valor de distância para o padrão.

Algumas medidas de avaliação do modelo obtidas no teste, podem ser visualizadas a partir da tabela abaixo

	Teste
Acurácia	$92,59$
Desvio Padrão	$3,97$
Tempo Médio (s)	$0,0064$

É importante ressaltar que o tempo médio de treinamento do DMC se equiparou ao tempo médio de treinamento, atingindo o valor médio de $\approx 0,006s$.

As superfícies de decisão podem ser mostradas a partir da combinação dos atributos par a par, a figura abaixo representa a superfície de decisão gerada com os dois primeiros atributos da base da iris, referente ao comprimento e a largura das sépalas, os padrões apresentados e as cores têm o mesmo significado do apresentado na seção sobre o KNN e as bolas pretas representam os centroids.

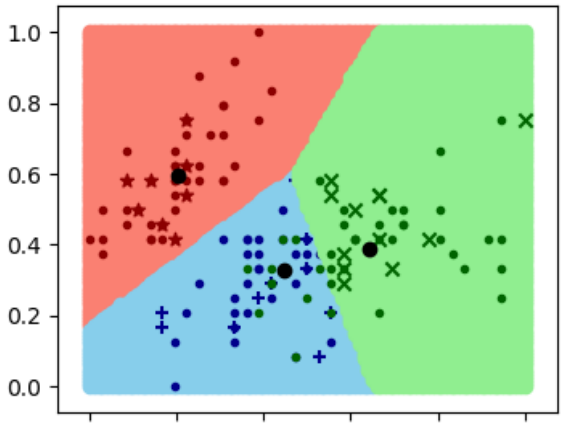

Figure 7: Superfície de Decisão - Base da Iris (DMC)

Abaixo, tem-se a representação da matriz de confusão de um dos testes, cuja taxa de acerto mais se aproxima dos valores encontrados para a acurácia.

	0	1	2
0	10	0	0
1	0	12	1
2	0	1	6

Os resultados nessa matriz de confusão representam que em um conjunto de dados de treino contendo 61 padrões, dos quais 10 eram da classe 0, 13 da classe 1 e 7 da classe 2, todos os padrões da classe 0 foram preditos corretamente, 1 padrão da classe 1 foi predito como sendo da classe 2 e um padrão da classe 2 foi predito como sendo da classe 1. Obtem-se, assim, a taxa de acerto no teste dessa realização, definida por $(10+12+6)/(10+13+7) = 28/30 \approx 93,33\%$.

2.7 Resultados - DMC - Base da Coluna Vertebral

Por fim, foi realizado o experimento de aplicação da base de dados da coluna vertebral no algoritmo de classificação DMC. Dos 310 padrões da base original, 62 foram utilizados para teste e o restante para treinamento. Assim como na base de dados da iris, foram encontradas 3 centroids, pois haviam 3 classes também para essa base. As mesmas medidas de avaliação de performance utilizadas nos demais experimentos, foi utilizada aqui. A tabela que representa essas medidas para essa base pode ser vista abaixo

	Teste
Acurácia	$74,18$
Desvio Padrão	$4,91$
Tempo Médio (s)	$0,0057$

Já para essa base, o tempo médio de treinamento foi bem inferior ao de teste, atingindo o valor de 0,0004s.

As superfícies de decisão desse algoritmo podem ser observadas abaixo a partir da combinação dos 6 atributos da base combinados 2 a 2, a imagem abaixo representa a superfície de decisão dos dois primeiros atributos: pelvic incidence e pelvic tilt.

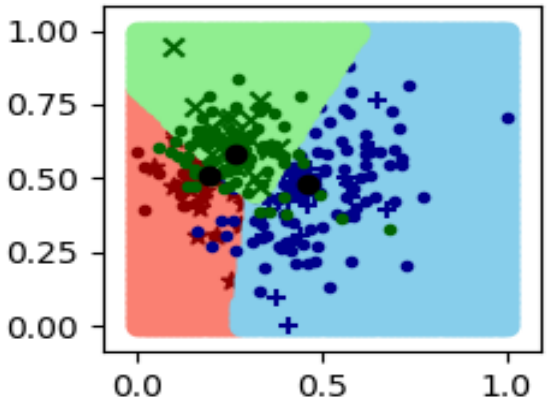

Figure 8: Superfície de Decisão - Base da Coluna Vertebral (DMC)

Uma matriz de confusão referente a uma realização de teste dentre as 20, pode ser analisada abaixo

	0	1	2
0	9	1	2
1	1	18	3
2	6	3	18

Essa matriz de confusão de uma das realizações representa que de um conjunto contendo 61 padrões de teste, haviam 12 padrões da classe 0, dos quais 9 foram preditos corretamente, 1 foi predito como sendo da classe 1 e 2 como sendo da classe 2. Haviam, além disso, 22 padrões da classe 1, dos quais 18 foram preditos corretamente, 1 foi predito como sendo da classe 0 e 3 como sendo da classe 2. Já para os padrões da classe 2, de 27 padrões, 18 foram preditos corretamente, 6 foram preditos como sendo da classe 0 e 3 como sendo da classe 1. Obtendo-se, assim, a taxa de acerto de $(9+18+18)/(12+22+27) = 45/61 \approx 73,77\%$.

2.8 Resultados - DMC - Base Artificial

Os resultados obtidos após as 20 realizações aplicando-se o DMC a base artificial podem ser vistos a partir da tabela abaixo

	Teste
Acurácia	$100,00$
Desvio Padrão	$0,0$
Tempo Médio (s)	$0,0002$

O algoritmo obteve 100% de acurácia em todos os testes, com tempo de treinamento médio superior ao tempo de teste, atingindo $0,0002$s.

A Figura 9 mostra a superfície de decisão do classificador, em que os dados distribuídos em formato de x e + representam os dados de teste das classes 0 e 1 respectivamente, os pontos pequenos representam os dados de treinamento e os pontos pretos representam os centroids das três classes.

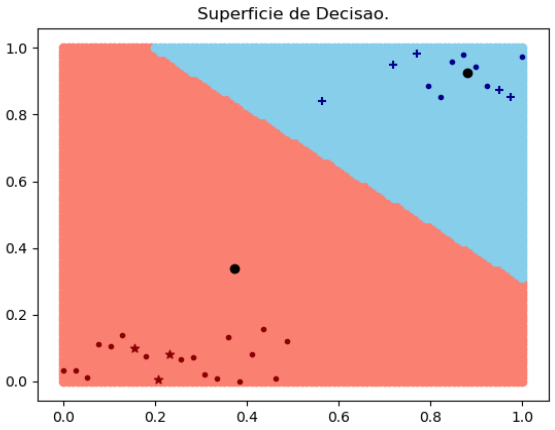

Figure 9: Superfície de Decisão - Base Artificial (DMC)

Um exemplo de uma matriz de confusão de uma das 20 realizações de teste é mostrada abaixo

	0	1
0	5	0
1	0	3

Essa matriz de confusão representa que de um conjunto de teste de 40 padrões, haviam 5 da classe 0 e 3 da classe 1 e todos foram classificados corretamente, obtendo-se então uma acurácia de 100%.

3 Conclusão

Os resultados obtidos para as três bases aplicadas aos algoritmos KNN e DMC, nos permite concluir que o DMC classifica os dados de todas as bases com uma acurácia bem próxima a do KNN, no entanto, o mesmo consome uma quantidade de tempo muito inferior ao tempo gasto no procedimento de teste do KNN, pois a classificação no mesmo consiste em apenas calcular distâncias aos centroids, já no KNN o compto de distâncias é muito maior, muito embora o treinamento nesse algoritmo seja tão simples, que apresente um tempo computacional desprezível.