EFC1 - Regressão Linear

Felippe Trigueiro Angelo - RA: 210479

20 de Outubro de 2019

Exercício 1 - Classificação Binária utilizando Regressão Logística

a) Análise dos Atributos de Entrada

No presente exercício será construído um modelo para classificação binária utilizando o algoritmo de regressão logística. Para tal finalidade será utilizada uma base de dados que contém alguns atributos extraídos de sinais de voz. Em cada sinal de voz utilizado foi considerado apenas as frequências entre 0 e 280 Hz. Assim, cada uma das entradas consiste em 19 atributos e um valor de saída contendo 0 ou 1, onde 0 indica que o sinal corresponde a uma voz feminina e 1 a um sinal masculina. Ao todo estão presentes 3168 amostras, das quais 20% são utilizadas para teste. É importante citar que o conjunto de teste foi obtido por meio de uma seleção aleatória onde foi utilizada a semente com valor 42.

Inicialmente foi realizada uma análise de todo o conjunto de dados por meio da representação do histograma de cada um dos atributos e por meio do cálculo da correlação entre eles. Os resultados podem ser vistos nas Figuras 1 e 2.

Observando os histogramas dos atributos é possível ver que a faixa dos seus valores varia significativamente, o que indica que um pré-processamento nos dados pode ser necessário a fim de que todos os atributos tenham contribuições similares no cálculo do erro, evitando assim, um retorno de coeficientes que podem enviezar o nosso modelo. Além disso, é possível ver que alguns atributos possuem distribuição

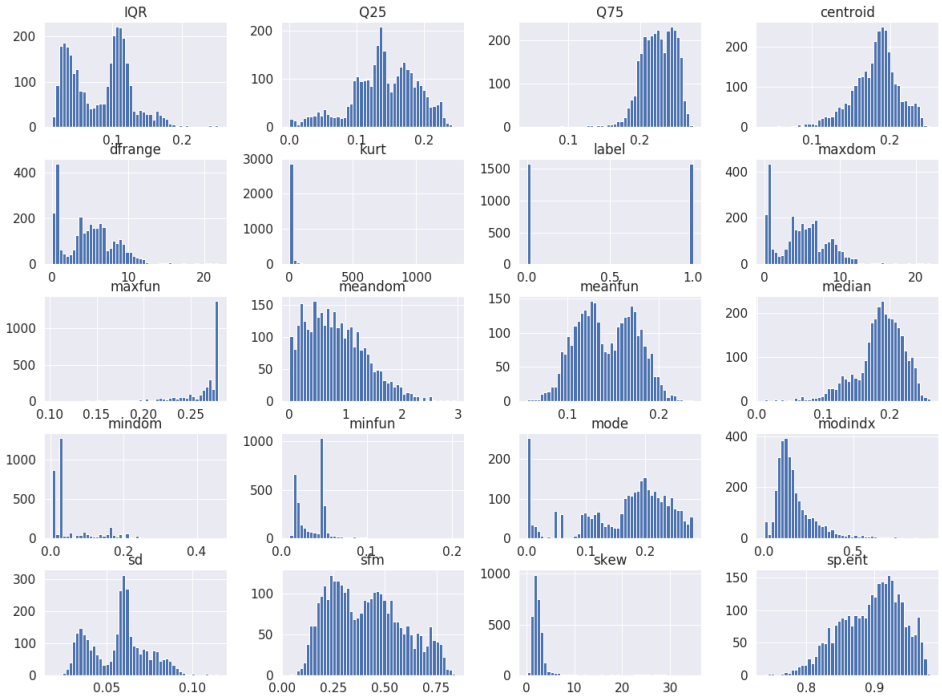

Figura 1: Histograma de cada um dos 19 atributos utilizados para a classificação.

gaussiana.

Observando os valores de correlação entre os atributos merece destaque os valores de correlação entre os atributos e a saída. Dentre estes, IQR, meanFun são os que fornecem a maior contribuição para o valor de saída. Isso indica que os valores de coeficiente relacionados a eles terão maior valor. Além disso alguns atributos, como por exemplo centroid e median possuem alta correlação entre si, indicando que eles podem ser atributos redundantes e podem ser bons candidatos a serem retirados para simplificar a etapa de treinamento.

b) Construção do modelo de Regressão Logística

Antes da realização do treinamento, o conjunto de dados foi normalizado, onde os valores de cada atributo foram convertidos de suas escalas normais para a escala entre 0 e 1, segundo a necessidade demonstrada no item anterior.

Para o treinamento do modelo de Regressão Logística foi utilizado o algoritmo de gradiente descendente com coeficiente de aprendizado fixado em 1. Além disso foi imposto um limitador de 3000 iterações ao algoritmo. Não foi utilizada uma etapa

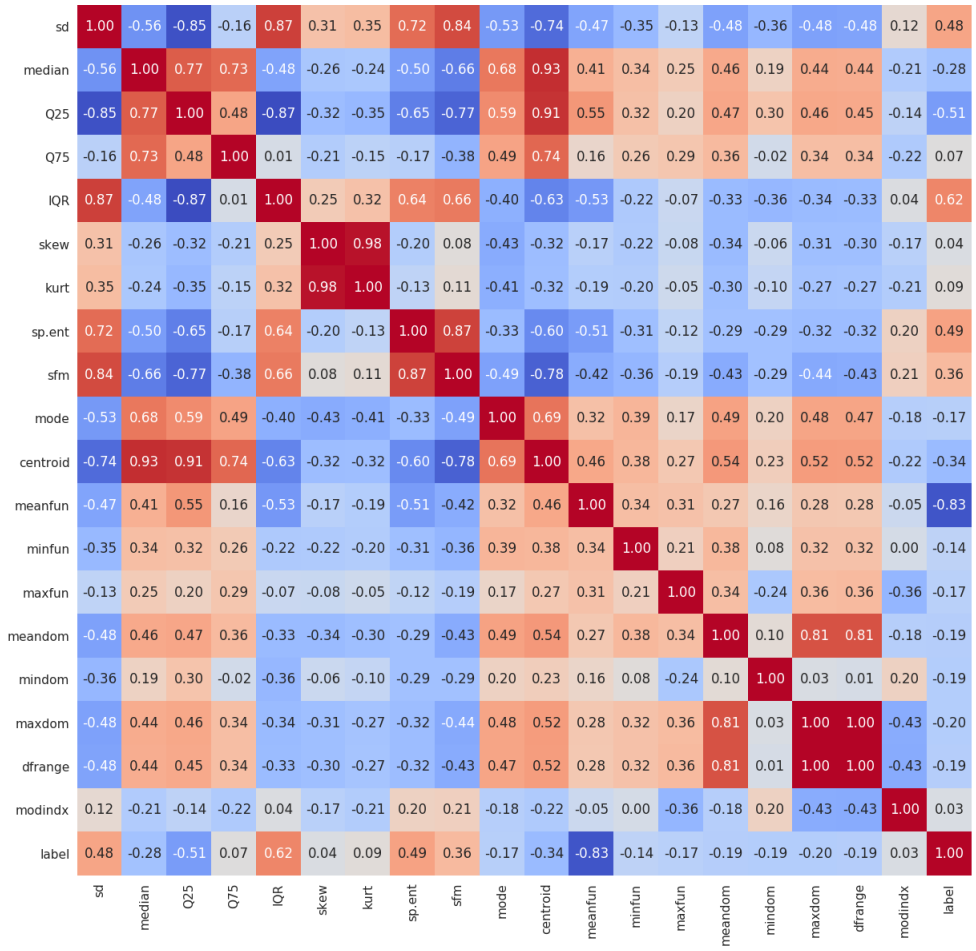

Figura 2: Cálculo da correlação entre os atributos utilizados na classificação.

de validação cruzada.

Em seguida foram calculadas algumas métricas para avaliar o desempenho do modelo no conjunto de teste. As métricas utilizadas foram a curva ROC (Receiver Operating Characteristic) e a medida F1. A curva ROC consiste basicamente em uma curva que plota a taxa de verdadeiro positivo contra a taxa de falso positivo. Portanto a sua construção se baseia no cálculo dessas taxas para um conjunto de valores de limiar. Ela pode ser vista como uma métrica que mensura a capacidade do classificador de distinguir entre as classes. Assim, quanto maior a área sobre a curva, ou consequentemente, quanto mais para a esquerda e pra cima, melhor é a capacidade do classificador em discriminar entre as classes. A medida F1 consiste em uma combinação das medidas Recall e Precisão. Quanto mais próximo de 1, melhores os resultados de Precisão e Recall e consequentemente, melhor o modelo.

Devido a sua definição, a medida F1 para o caso de uma classificação binária é dependente do valor de limiar. Assim, foi calculada a medida F1 em função de um conjunto de valores de limiar. Ambas as curvas podem ser visualizadas nas Figuras 3 e 4.

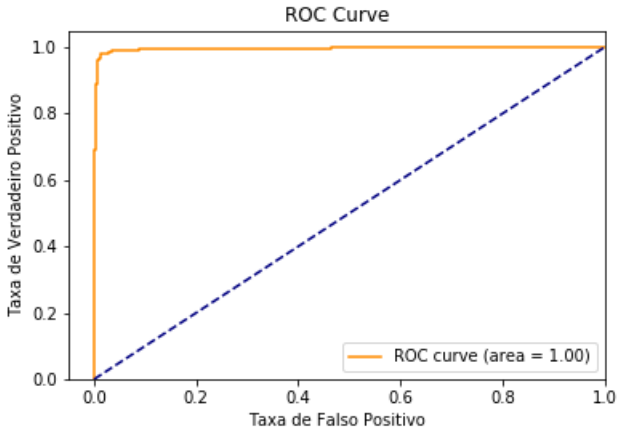

Figura 3: Curva ROC para o caso da classificação binária.

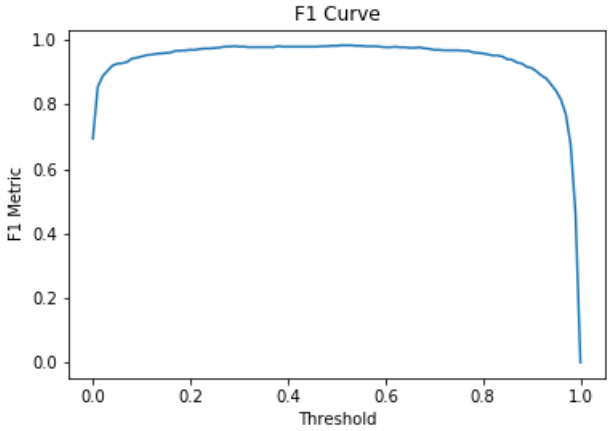

Figura 4: Curva da medida F1 em função dos valores de limiar utilizados na classificação.

c) Análise dos Resultados

Por meio da análise da curva ROC é possível ver que a área sob a curva é de aproximadamente 1, (Neste ponto é necessário citar que a área da curva não é exatamente

1 como pode ser visto pelo seu formato. Entretanto devido o seu valor ser muito próximo, tal número foi arredondado) o que indica que o classificador possui uma distinção entre as distribuições de probabilidade das classes que é praticamente perfeita.

Na análise da curva F1 é possível ver que não houve diferença significativa entre os valores de limiar 0.2 e 0.7. Entretanto a curva tem seu valor máximo no limiar igual a 0.51 e que corresponde ao valor de F1 igual a 0.984, o que também indica que o classificador obteve tambem um bom desempenho nas medidas de Precisão e Recall.

Assim, como a medida F1 foi máxima no valor de limiar igual a 0.51 ele pode ser colocado como o valor mais adequado para a separação entre as classes. Com esse valor foi calculada a matriz de confusão (Figura 5) e a acurácia do classificador. A acurácia obtida foi de 0.983.

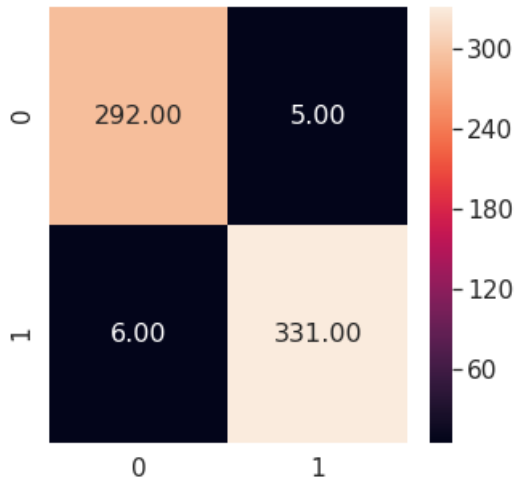

Figura 5: Matriz de confusão para o caso da classificação binária.

A acurácia corresponde ao percentual de acertos (valores classificados corretamente, ou seja, verdadeiros positivos e negativos) do classificador. Como ela obteve um valor bem próximo de 1 é possível dizer que o classificador realiza uma separação quase que perfeitamente. Tal comportamento pode ser observado pela matriz de confusão, onde podemos ver que a quantidade de valores classificados corretamente

é muito superior aos demais (Falsos positivos e negativos).

Exercício 2 - Classificação multi-classe

a) Regressão Logística com a Abordagem Softmax

Na segunda parte do exercício será utilizado o algoritmo de regressão logística para classificação multi-classe. Nessa etapa será utilizada a abordagem softmax. A abordagem softmax consiste em se obter uma saída, onde cada um dos seus Q valores corresponde a probabilidade da saída pertencer a determinada classe Q_i. Assim, o vetor de saídas é descrito como:

$$\hat{y}(\mathbf{x}(i)) = \frac{e^{\phi(\mathbf{x}(i))^T \mathbf{w}_k}}{\sum_j e^{\phi(\mathbf{x}(i)^T)\mathbf{w}_k}} \tag{1}$$

A função de custo é descrita como:

$$J_{CE}(\mathbf{W}) = -\sum_{i=0}^{N-1}\sum_{k=1}^{Q} y_{i,k} log\hat{y}(\mathbf{x}(i)) \tag{2}$$

É importante citar que o vetor contendo os rótulos dos conjuntos de teste e de treinamento deve ser convertido para a representação *one-hot encoding*.

O conjunto de treinamento utilizado nessa atividade contém atributos nos domínios do tempo e da frequência extraídos de sinais de acelerômetro e giroscópio de um smartphone. Os rótulos correspondentes aos dados indicam qual a atividade realizada por um voluntário humano durante a aquisição dos sinais: 0 – caminhada; 1 – subindo escadas; 2 – descendo escadas; 3 – sentado; 4 – em pé; 5 – deitado. O conjunto de dados já está separado em uma parte para treinamento e outra para teste. Ao todo, temos 7352 amostras de treinamento e 2947 amostras de teste; cada amostra é descrita por 561 atributos temporais ou espectrais.

Durante a etapa de treinamento os dados de treinamento e de teste foram normalizados para que os seus valores de atributo estejam compreendidos entre a faixa de 0 e 1. Após essa normalização, os dados de treinamento são inseridos no algoritmo. Foram utilizadas 3000 iterações com uma taxa de aprendizagem de 0.2. Tal número

de iterações foi utilizado devido ao fato de após essa quantidade, o gradiente ser praticamente nulo. Não foi utilizada a regularização nem validação cruzada.
Como métrica global de avaliação do desempenho do classificador, foi utilizada a medida F1. Além disso também foi calculada a matriz de confusão. A medida F1 foi de 0.89, o que indica um bom desempenho do classificador na separação das classes. A matriz de confusão pode ser vista na Figura 6.

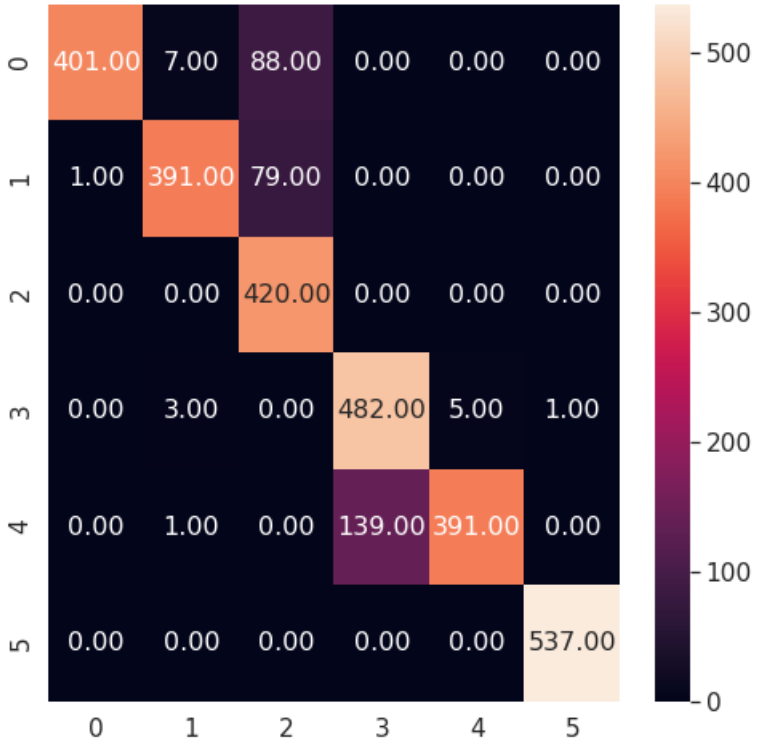

Figura 6: Matriz de confusão para o caso da classificação multi-classe.

É possível ver que a maior parte das classificações foi feita corretamente. Entretanto, uma quantidade significativa de erros na classificação pode ser encontrada nas regiões em roxo. Tais regiões contribuem para o resultado do F1 não ser um resultado mais próximo de 1.

b) K-Nearest Neighbour

O modelo K-nearest neighbour para classificação foi testado no conjunto de dados. Devido a questões intrínsecas ao modelo não foi realizada a etapa de treinamento. Entretanto nos dados de teste foi avaliado o desempenho do classificador para um dado valor de K. Foram testados 7 valores. São eles: 1, 3, 6, 10, 30, 60, 100. Para cada um desses valores foi calculada a medida F1 juntamente com a respectiva matriz de confusão. Não foi realizada a normalização dos dados, bem como não foi realizada a validação cruzada.

A Figura 7 mostra a matriz de confusão para o valor de K igual a 1. A medida F1 obtida para esse valor de K foi de 0.876.

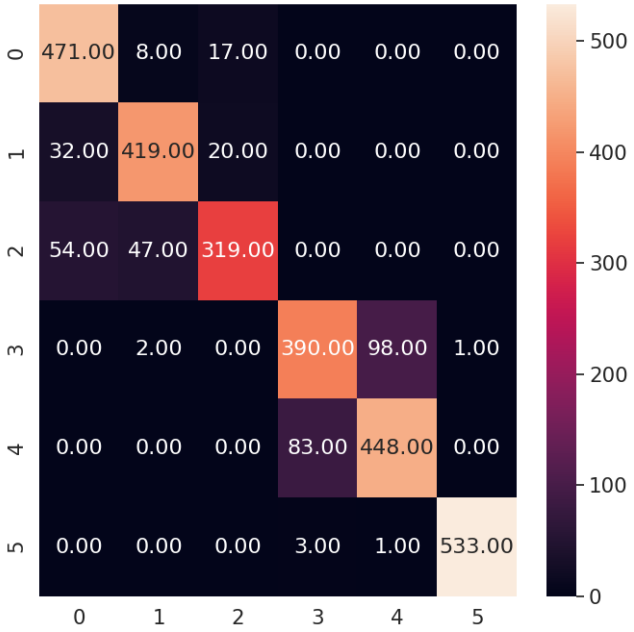

Figura 7: Matriz de confusão para o caso da classificação multi-classe utilizando KNN com K igual a 1.

A Figura 8 mostra a matriz de confusão para o valor de K igual a 3. A medida F1 obtida para esse valor de K foi de 0.89.

A Figura 9 mostra a matriz de confusão para o valor de K igual a 6. A medida F1 obtida para esse valor de K foi de 0.9.

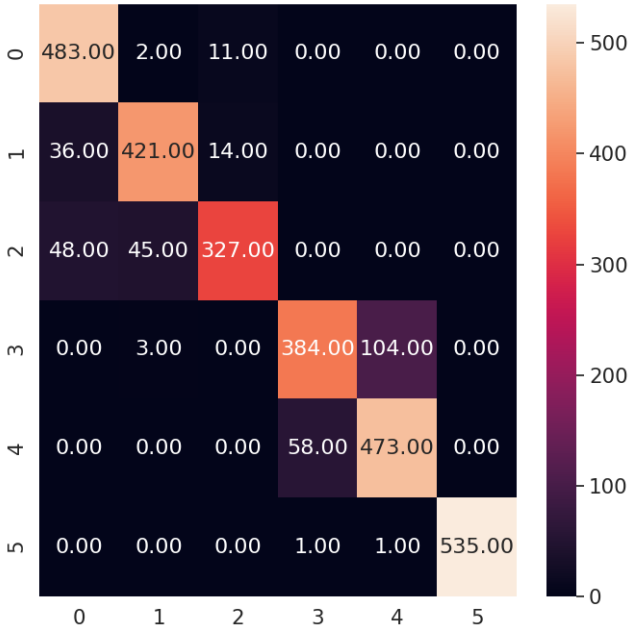

Figura 8: Matriz de confusão para o caso da classificação multi-classe utilizando KNN com K igual a 3.

A Figura 10 mostra a matriz de confusão para o valor de K igual a 10. A medida F1 obtida para esse valor de K foi de 0.904.

A Figura 11 mostra a matriz de confusão para o valor de K igual a 30. A medida F1 obtida para esse valor de K foi de 0.899.

A Figura 12 mostra a matriz de confusão para o valor de K igual a 60. A medida F1 obtida para esse valor de K foi de 0.89.

A Figura 13 mostra a matriz de confusão para o valor de K igual a 100. A medida F1 obtida para esse valor de K foi de 0.886.

De posse dessas métricas é possível ver que dentre os valores testados o valor de K igual a 10 foi aquele que resultou em um maior valor de F1, indicando um melhor desempenho do modelo. Em uma comparação entre os modelos de KNN e de regressão logística, pode-se ver que ambos os modelos tiveram desempenhos similares, entretanto para os valores de K iguais a 6 e 10 o KNN obteve um desempenho ligeiramente melhor.

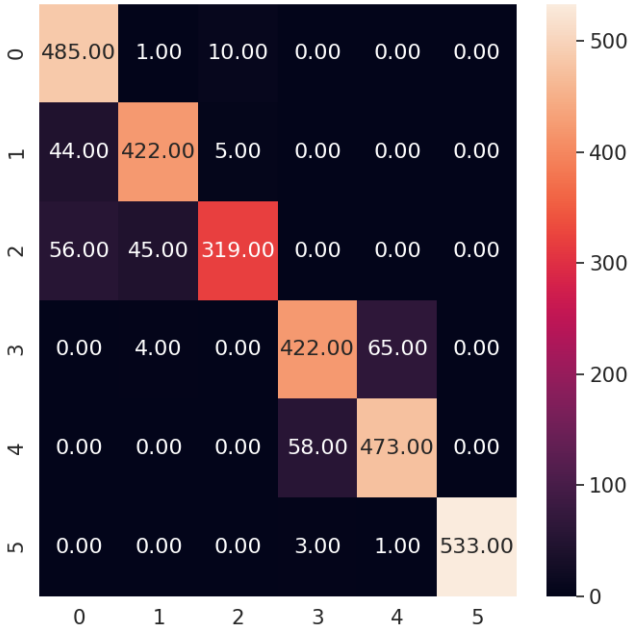

Figura 9: Matriz de confusão para o caso da classificação multi-classe utilizando KNN com K igual a 6.

Um outro ponto a ser destacado é o das regiões da matriz de confusão que resultaram em classificações erradas. Na regressão logística as classes que resultaram em um maior erro de classificação foram as classes 4 e 3. Entretanto no algoritmo de KNN além das classes 4 e 3, outras classes como 2 e 0 resultaram um error consideráveis de classificação.

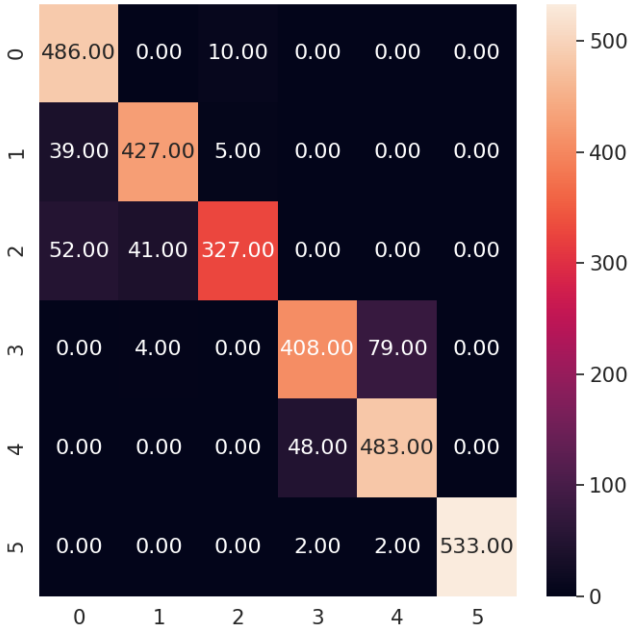

Figura 10: Matriz de confusão para o caso da classificação multi-classe utilizando KNN com K igual a 10.

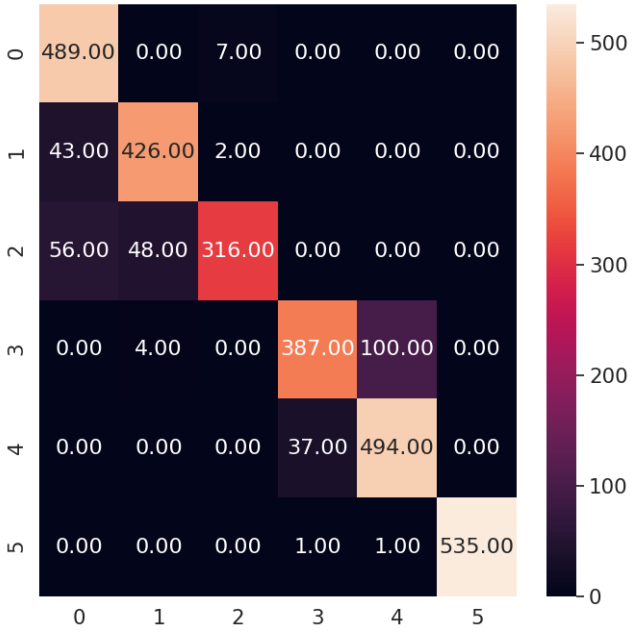

Figura 11: Matriz de confusão para o caso da classificação multi-classe utilizando KNN com K igual a 30.

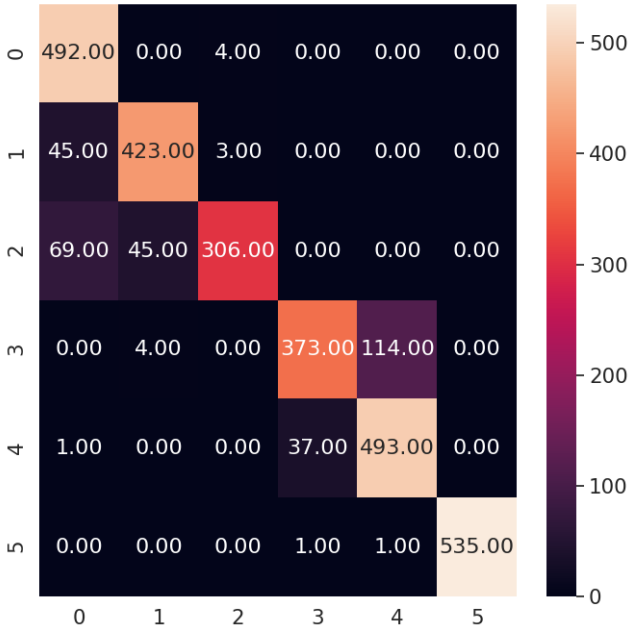

Figura 12: Matriz de confusão para o caso da classificação multi-classe utilizando KNN com K igual a 60.

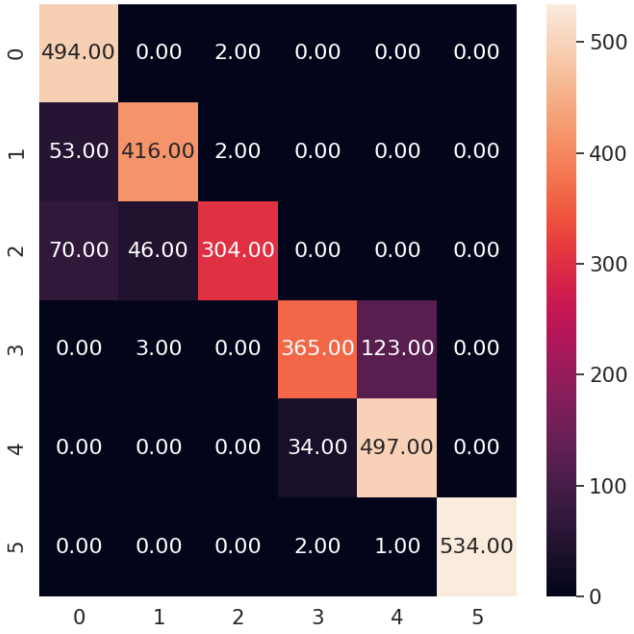

Figura 13: Matriz de confusão para o caso da classificação multi-classe utilizando KNN com K igual a 100.