Projeto 3 – Algoritmo KNN

Apresentação da Análise do Dataset "Glass Identification"

Giovane Bruno Nardari – 21000244

João Roberto Crespi Junior - 21003801

Introdução

- O Dataset selecionado pelo grupo foi o "Glass Identification", criado com o intuito de treinar algoritmos de inteligência artificial para a tarefa de classificação de diferentes tipos de vidro.
- Os dados contidos no Dataset foram coletados pelo Serviço de Ciência Forense dos Estados Unidos.
- O Dataset contém 214 exemplares, que apresentam sete atributos que tratam da composição química e do índice de refração das amostras de vidro coletadas.

Classes e Atributos

- As classes do Dataset, que, neste caso, representam os 7 diferentes tipos de vidro que os dados abrangem, são, em inglês:

1. ***Building windows float-processed*** **(Tipo 1);**
2. ***Building windows non-float-processed*** **(Tipo 2);**
3. ***Vehicle windows float-processed*** **(Tipo 3);**
4. ***Vehicle windows non-float-processed*** **(Tipo 4);**
5. ***Containers*** **(Tipo 5);**
6. ***Tableware*** **(Tipo 6);**
7. ***Headlamps*** **(Tipo 7);**

Classes e Atributos

- Já os atributos analisados para classificar os exemplares do Dataset são:

1. **Índice de Refração (RI);**
2. **Sódio (Na);**
3. **Magnésio (Mg);**
4. **Alumínio(Al);**
5. **Silício (Si);**
6. **Potássio (K);**
7. **Cálcio (Ca);**
8. **Bário (Ba) ;**
9. **Ferro (Fe);**

Visualização e Análise

- Inicialmente, foi feita uma procura por dados ruidosos que poderiam existir no Dataset e também foram verificadas algumas medidas estatísticas.
- As primeiras análises feitas nos permitiram concluir que não existia nenhum dado faltante entre os exemplares do Dataset, mas havia uma disparidade de exemplares entre as classes, como mostra o gráfico a seguir:

Visualização e Análise

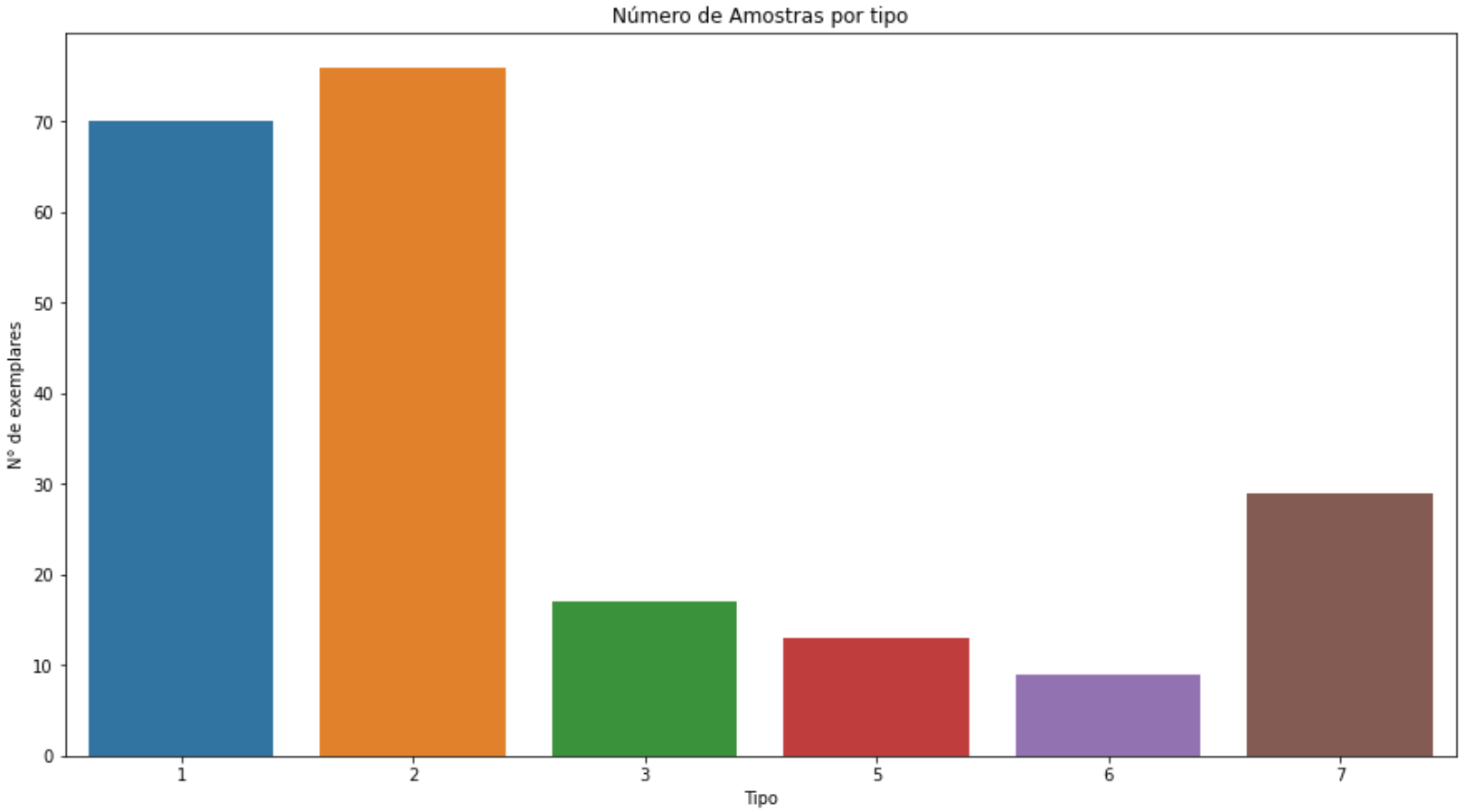

Figura 1 - Gráfico de barras contabilizando o número total de exemplares para cada tipo.

Visualização e Análise

- Também foi realizado uma serie de *boxplots* para observação de *outliers*

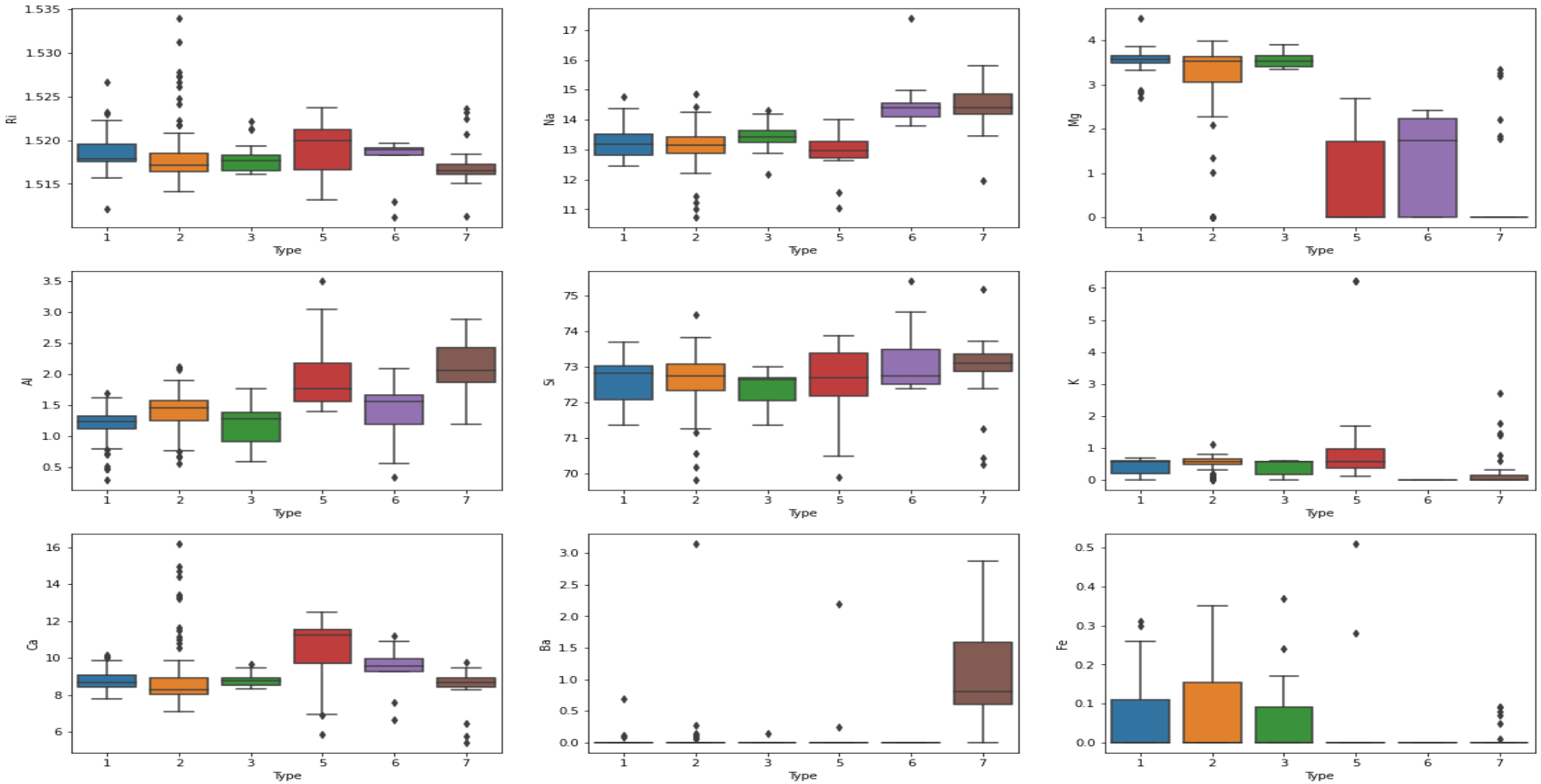

Figura 2 – *Boxplots* comparando os tipos de vidro para cada um de seus atributos.

Visualização e Análise

- Após essas constatações, os exemplares do tipo 1, 2 e 3 ainda geravam dificuldades para serem distinguidos.
- Em busca de resolver esse problema, foi criado um gráfico da distribuição normal desses tipos, para melhor diferenciação dos mesmos.

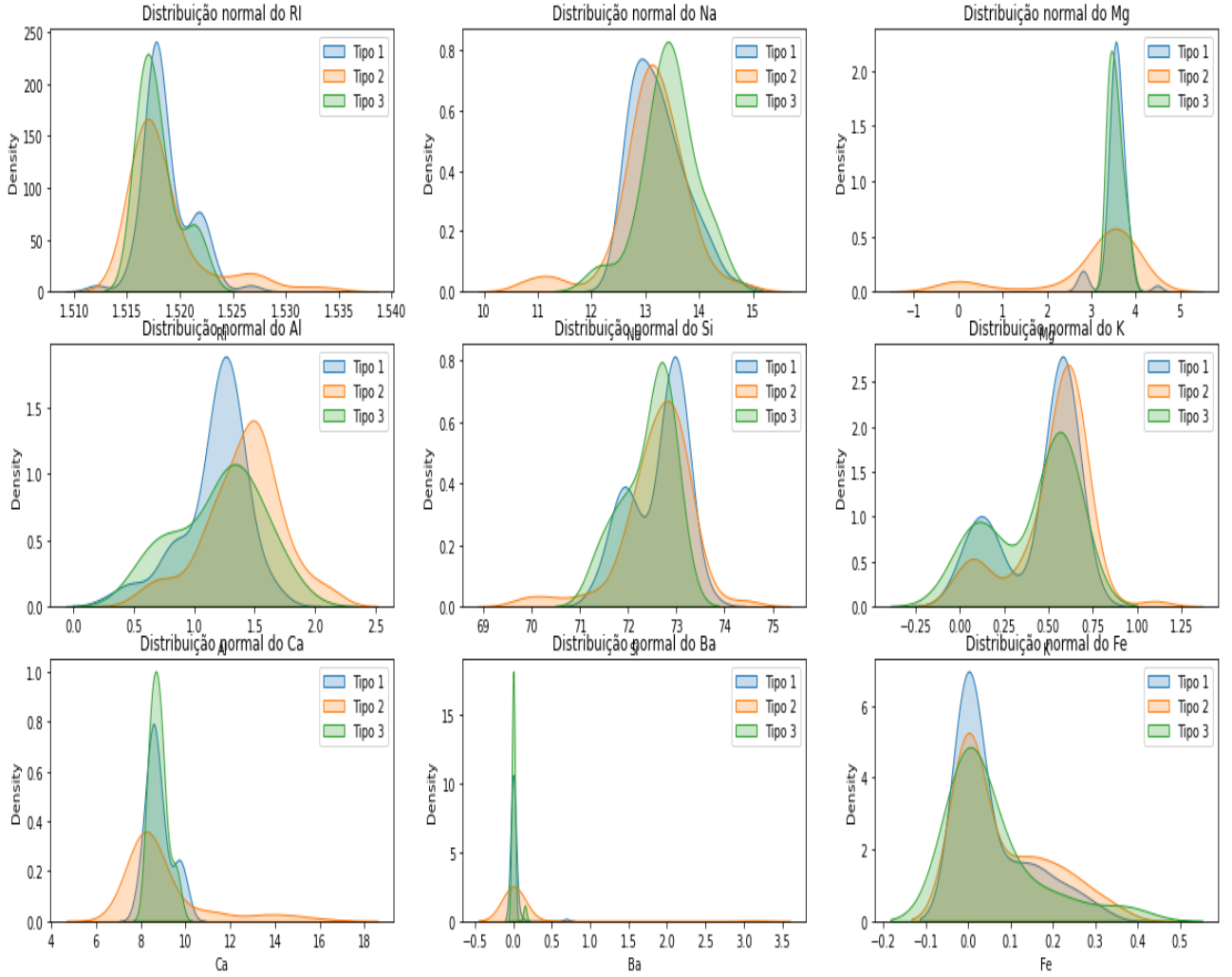

Figura 3 – Distribuição normal dos atributos para os tipos 1, 2 e 3.

Algoritmo K-Nearest Neighbors

- Para a tarefa de classificação dos exemplares, foi aplicado o algoritmo KNN, que foi, ao longo do projeto, adaptado, em busca de resultados que fossem satisfatórios, tanto no "score", quanto na verdadeira representação da realidade.
- Vários testes foram realizados, alterando o Dataset e a configuração de treinamento dos dados.
- Após alguns testes, identificamos as duas melhores configurações: o dataset original, com o "score" variando entre 44,2% e 93%, e o dataset sem os *outliers* (selecionados usando o critério do intervalo interquartil).

Algoritmo K-Nearest Neighbors

- Os gráficos a seguir mostram a distribuição dos scores obtidos sobre os dois Datasets, variando o número "k" para diferentes amostras de treino e teste.

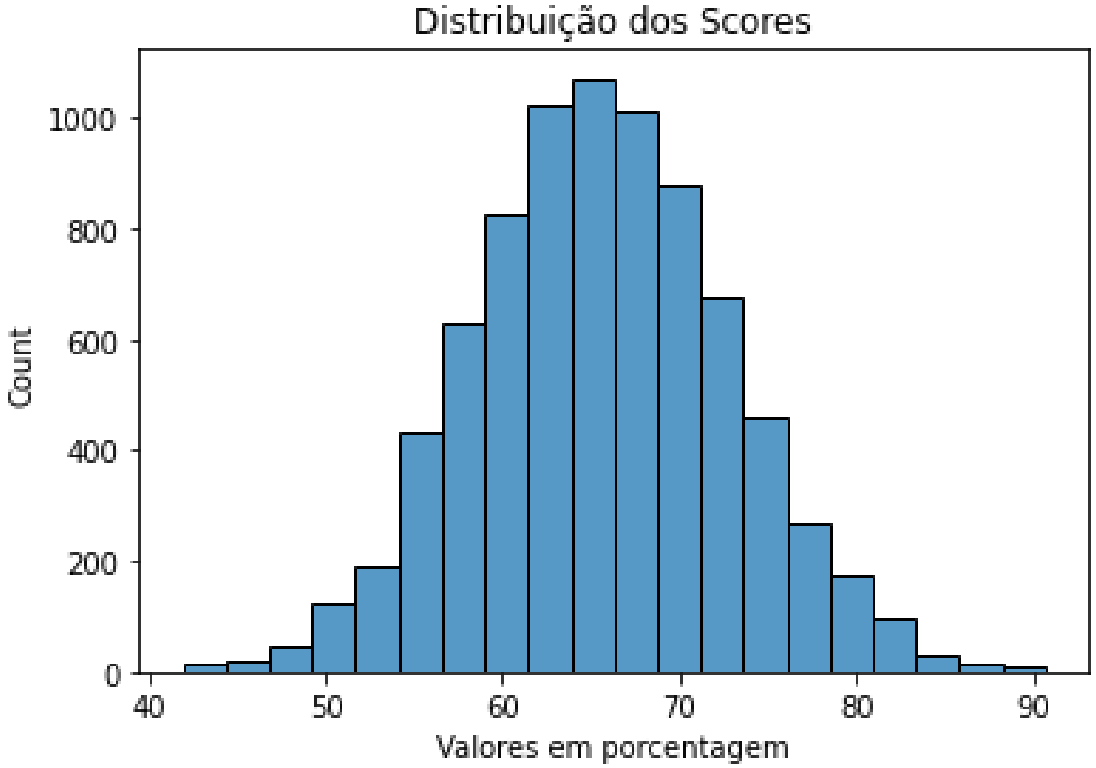

Figura 4 - Distribuição dos "scores" obtidos para diferentes "k" dentro do algoritmo KNN, aplicado sobre o Dataset original.

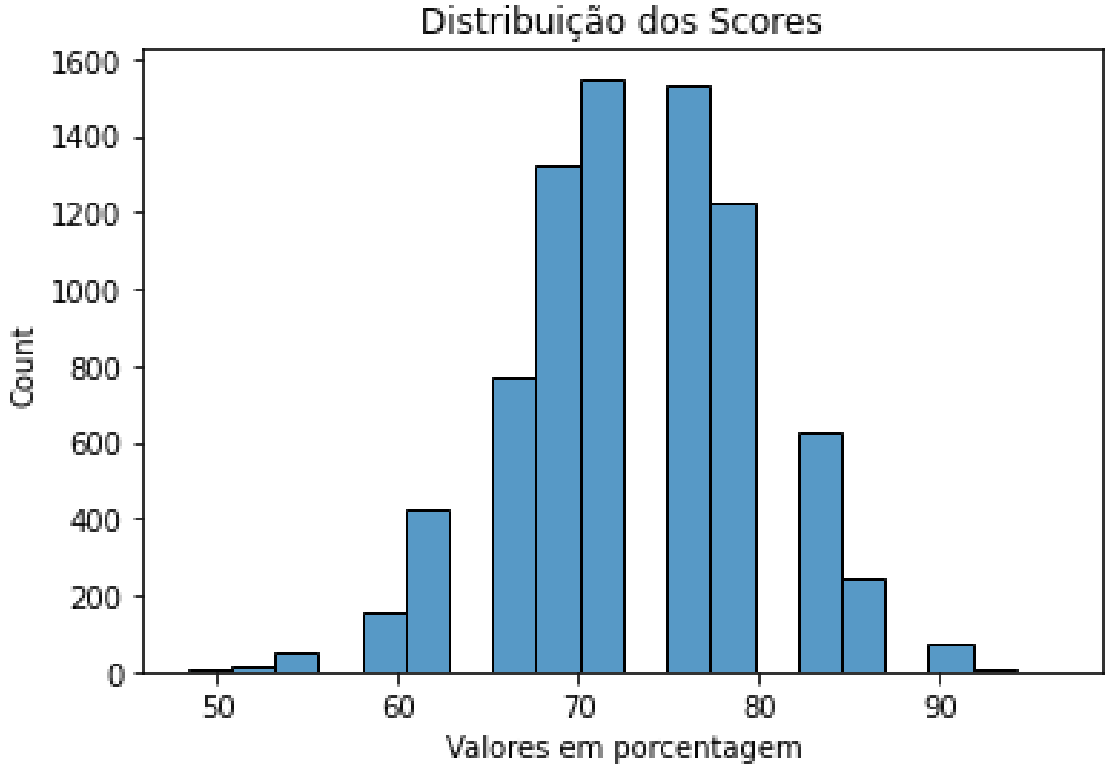

Figura 5 - Distribuição dos "scores" obtidos para diferentes "k" dentro do algoritmo KNN, aplicado sobre o Dataset sem *outliers*.

Algoritmo K-Nearest Neighbors

- Após a comparação dos testes realizados, foi decidido que, apesar do Dataset modificado apresentar melhora nos resultados, a redução de 32,7% dos dados (*Outliers*) poderiam, em verdade, traduzir um resultado irreal, o que nos leva a entender que o Dataset original deve ser a melhor opção para o algoritmo.
- O melhor resultado obtido foi utilizado para fazer uma matriz de confusão, o que nos levou a entender que os erros de classificação ocorridos, envolvem majoritariamente, os tipos 1, 2 e 3, erros esses, provavelmente, causados pela alta semelhança entre os exemplares dessas três classes, aliado a menor quantidade de exemplares do tipo 3.

Algoritmo K-Nearest Neighbors

- Os testes realizados nos deixam a conclusão de que seria possível classificar todas os tipos de vidro, com poucos erros, caso existisse um maior volume de dados coletados para serem oferecidos no treino do algoritmo.

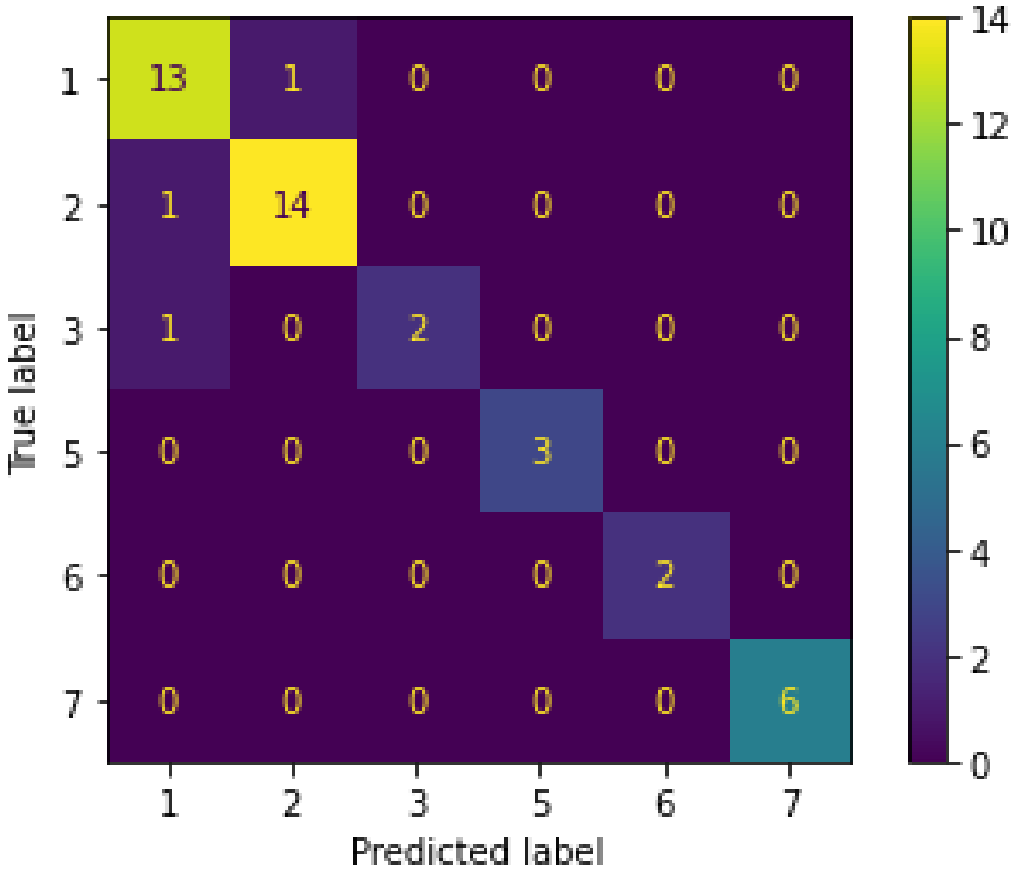

Figura 6 - Matriz de confusão para o melhor resultado obtido.

Agradecemos a sua atenção!
Estamos à disposição para quaisquer dúvidas!

Giovane Bruno Nardari

João Roberto Crespi Junior