UNIVERSIDADE FEDERAL DO PARANÁ

SETOR DE TECNOLOGIA

CURSO DE INFORMÁTICA BIOMÉDICA

ANNA CAROLINE BOZZI

IMPACTOS DA BASE DE APRENDIZAGEM

CURITIBA

2020

Introdução

Esse relatório possui análise referente aos classificadores:

- KNN
- Naïve Bayes
- Linear Discriminant Analysis
- Logistic Regression
- Perceptron

Com dados previamente fornecidos para treino e teste, *test.txt* e *train.txt,* os quais possuem para teste 58646 dados com 132 características cada e 2000 dados com 132 características cada para treino. Foram feitas as seguintes considerações, que serão respondidas a diante:

1. Comparação do desempenho desses classificadores em função da disponibilidade de base de treinamento.
2. Indicação do classificador que tem o melhor desempenho com poucos dados < 1000 exemplos.
3. Indicação do classificador com melhor desempenho com todos os dados.
4. Classificador mais rápido para classificar os 58k exemplos de teste.
5. Análise das matrizes de confusão. E verificação dos erros para todos os classificadores quando todos eles utilizam toda a base de teste.

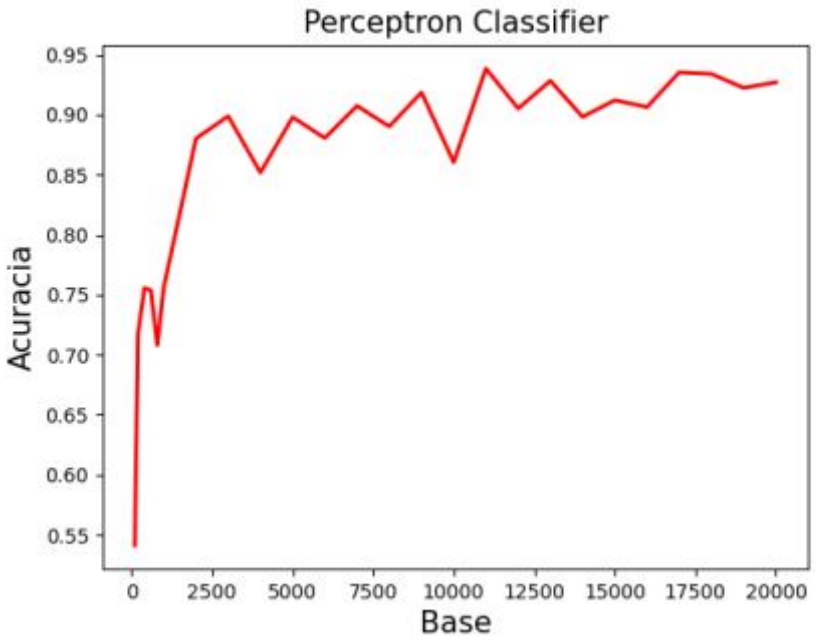
Resultados e Análise

1. Comparação do desempenho dos classificadores em função da disponibilidade de base de treinamento:

Foi alimentado o classificador com blocos de 1000 exemplos, iniciados em 1000 e terminando em 58000.

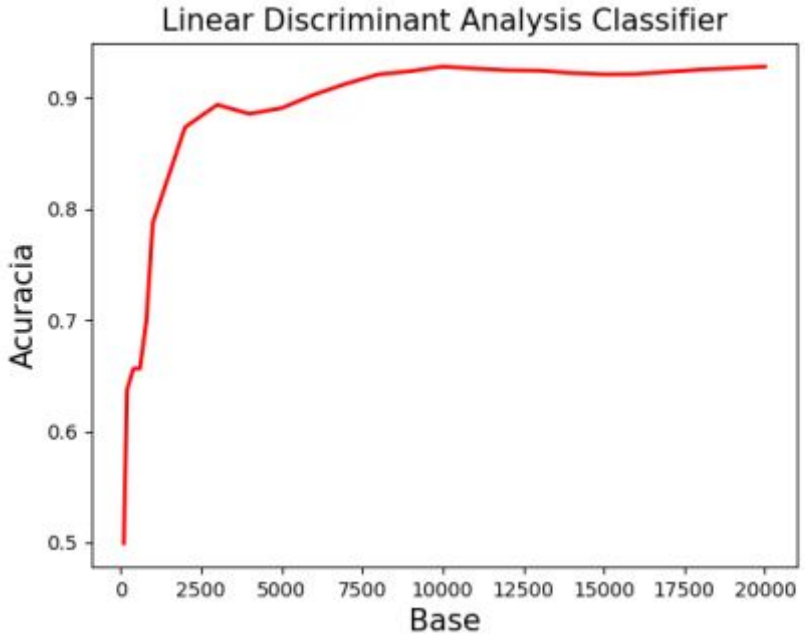
Perceptron:

Iniciando com a análise do *Perceptron*, é possível verificar no gráfico a seguir, que a Acurácia máxima verificada foi de 93,5 com uma base de 17000 dados de teste, e à partir desse valor o aumento da base não interferiu mais positivamente na Acurácia.



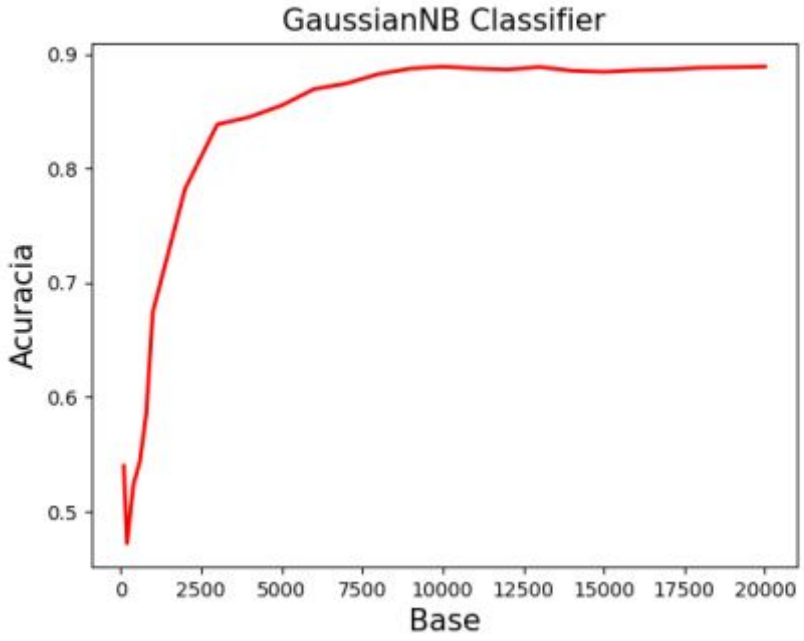
Linear Discriminant Analysis:

A seguir temos o *Linear Discriminant Analysis*, conforme o gráfico a seguir, a Acurácia máxima verificada foi de 92,7, para 1000 dados da base de teste, à partir daí foi observado novamente para 20000 dados de teste. Ou seja, à partir de 10000 dados não houve significativa mudança de acurácia que justificasse usar os 20000 dados de treinamento.



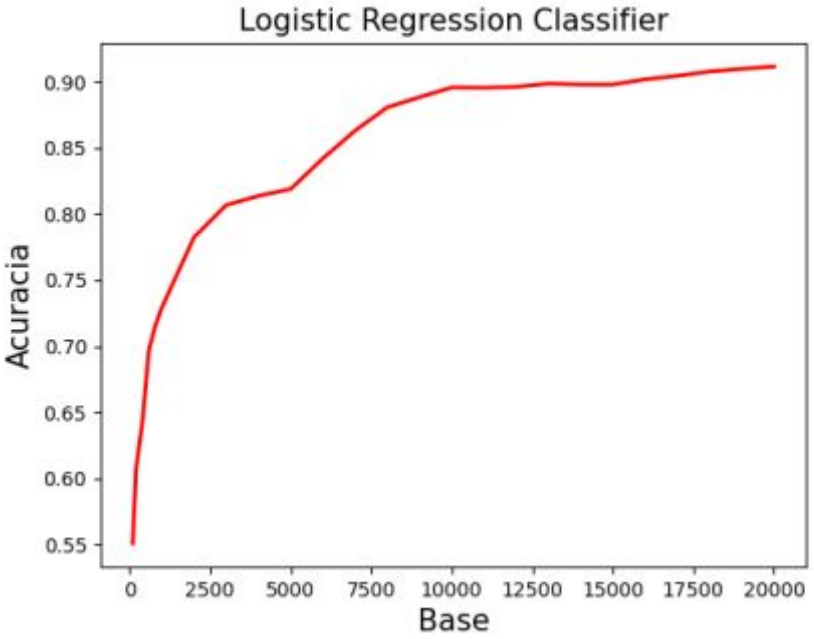
Naive Bayes:

O classificador *GaussianND*, apresentou Acurácia máxima de 88,9 para as base de 10000 dados de teste, e novamente verifica em 20000 dados, conforme gráfico a seguir. Assim como no caso do *Linear Discriminant Analysis,* não houve desempenho suficiente para justificar o uso da base inteira de testes.



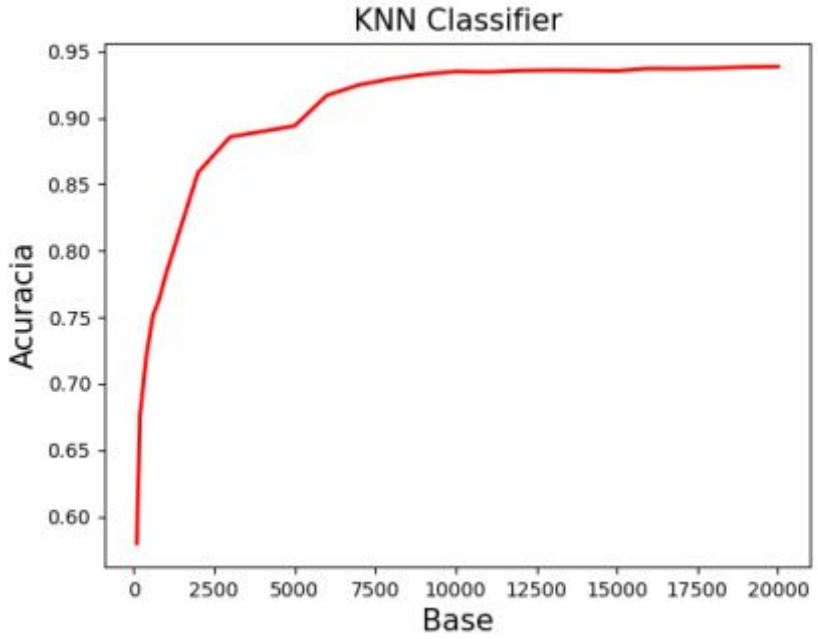
Logistic Regression:

O classificador *Logistic Regression*, apresentou Acurácia máxima de 91,1 para as base de 2000 dados de teste conforme gráfico apresentado a seguir.



KNN:

Por fim a apresentação dos resultados do classificador *KNN*, a Acurácia máxima observada foi de 93,8, para a base de 20000 dados de teste conforme gráfico a seguir. A partir de 10000 dados de teste houve acréscimo pequeno de acurácia até atingir a máxima em 20000 dados.

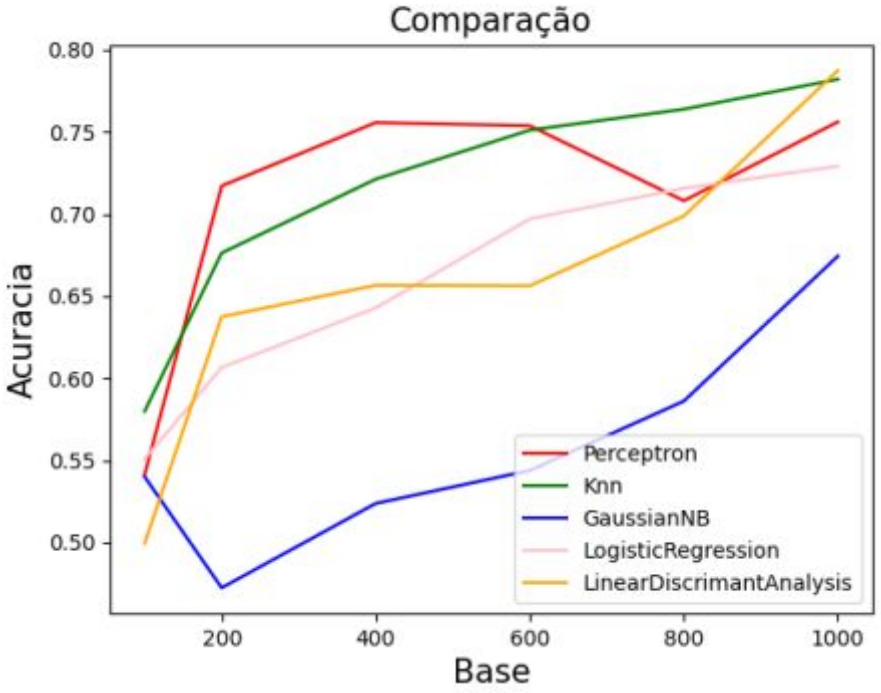


2. Classificador com o melhor desempenho para poucos dados:

Para essa análise foi testado todos os classificadores considerando valores de base abaixo de 1000 dados.

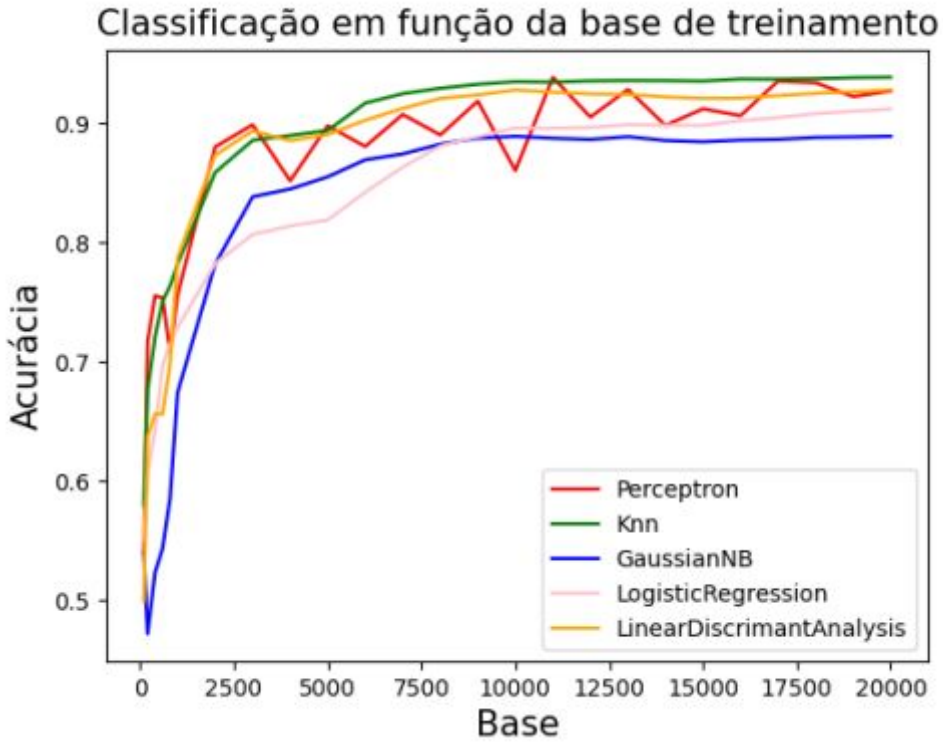
É possível verificar no gráfico a seguir a comparação entre todos os classificadores, *KNN, Naïve Bayes, Linear Discriminant Analysis, Logistic Regression, Perceptron,* o que apresentou o melhor desempenho geral para poucos dados, <1000, foi o KNN, conforme verifica-se no gráfico a seguir. O Perceptron apresentou bons resultados também, porém houve uma queda significativa a partir da base de 800 dados. A métrica utilizado para análise de desempenho foi a médias das Acurácias de cada classificador, segue:

- KNN: 71
- Perceptron: 70
- Logistic Regression: 65
- Linear Discriminant Analysis: 65
- Gauss para Naive Bayes: 55



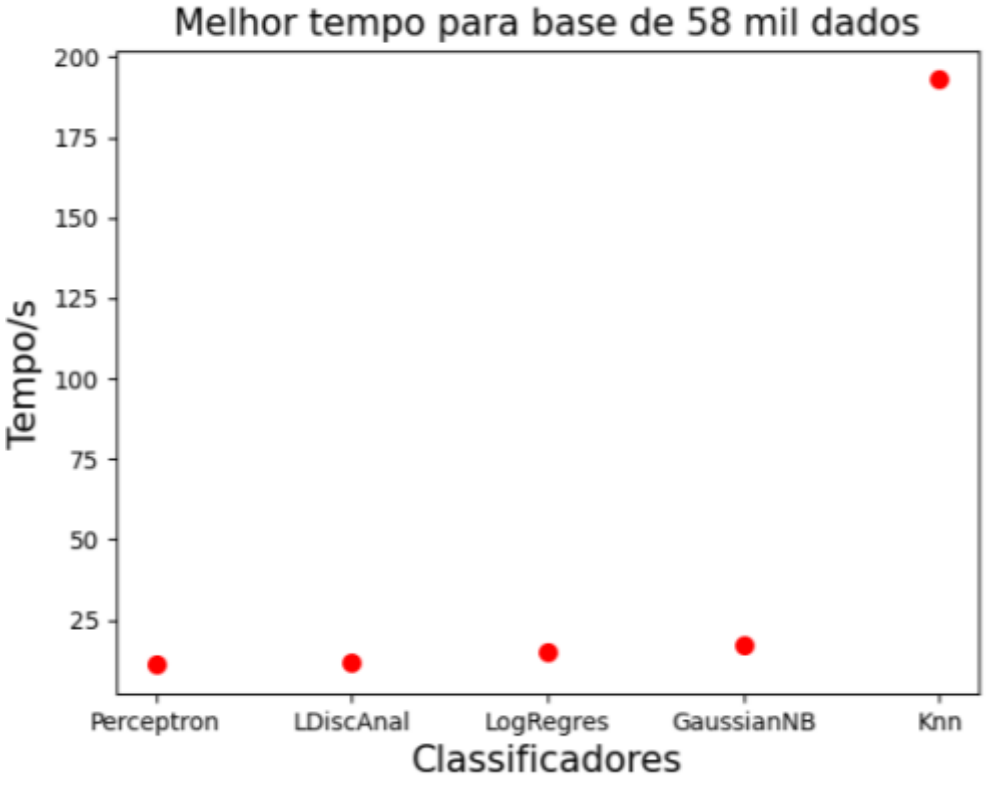
3. Indicação do classificador com o melhor desempenho com todos os dados de aprendizagem:

É possível verificar no gráficos que se segue, a comparação dos 5 classificadores, o KNN foi o classificador que apresentou melhor desempenho para todos os dados de teste, considerando a acurácia como métrica de avaliação.



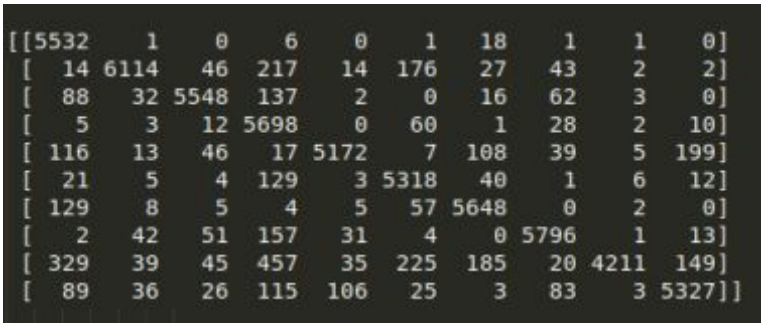
4. Indicação do classificador mais rápido para classificar os 58000 dados de teste:

Para a análise de parâmetro foi utilizada a função *time* na compilação dos arquivos e dados. O classificador mais rápido para 58000 dados de teste foi o *Perceptron,* com 11 segundos de execução, seguido dos *Linear Discriminant*, *Analysis,* 12 segundos ,*Logistic Regression,* 15 segundos , *GaussianND ,* 17 segundos, e por fim o que levou mais tempo, *knn,* 193 segundos. No gráfico a seguir é possível visualizar o comparativo de tempo:



5. Análise das matrizes de confusão, e verificação dos erros para todos os classificadores quando todos eles utilizam toda a base de teste:

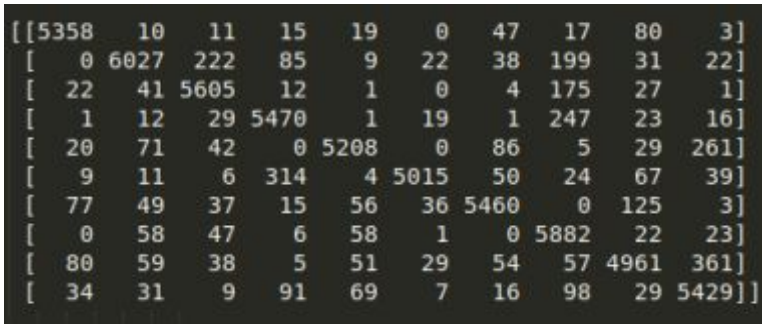
Perceptron:



```
[[5532    1    0    6    0    1   18    1    1    0]
 [  14 6114   46  217   14  176   27   43    2    2]
 [  88   32 5548  137    2    0   16   62    3    0]
 [   5    3   12 5698    0   60    1   28    2   10]
 [ 116   13   46   17 5172    7  108   39    5  199]
 [  21    5    4  129    3 5318   40    1    6   12]
 [ 129    8    5    4    5   57 5648    0    2    0]
 [   2   42   51  157   31    4    0 5796    1   13]
 [ 329   39   45  457   35  225  185   20 4211  149]
 [  89   36   26  115  106   25    3   83    3 5327]]
```

Esse classificador possui erros mais frequentes nas classificações do 1, 9 e do 10. Com mais erros de 9.

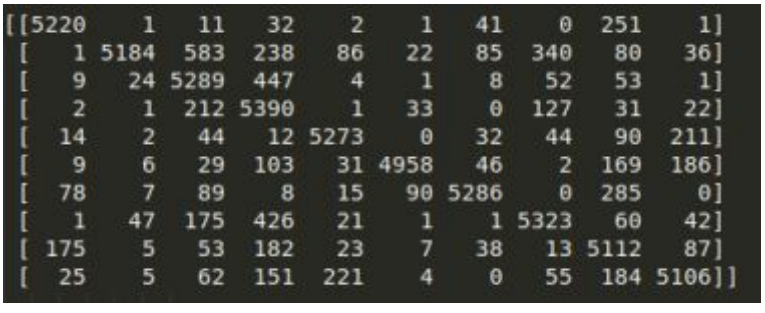
Linear Discriminant Analysis:



```
[[5358   10   11   15   19    0   47   17   80    3]
 [   0 6027  222   85    9   22   38  199   31   22]
 [  22   41 5605   12    1    0    4  175   27    1]
 [   1   12   29 5470    1   19    1  247   23   16]
 [  20   71   42    0 5208    0   86    5   29  261]
 [   9   11    6  314    4 5015   50   24   67   39]
 [  77   49   37   15   56   36 5460    0  125    3]
 [   0   58   47    6   58    1    0 5882   22   23]
 [  80   59   38    5   51   29   54   57 4961  361]
 [  34   31    9   91   69    7   16   98   29 5429]]
```

Assim como o perceptron, os erros mais frequentes são na classificação do 1,9 e 10. Com mais erros no 1.

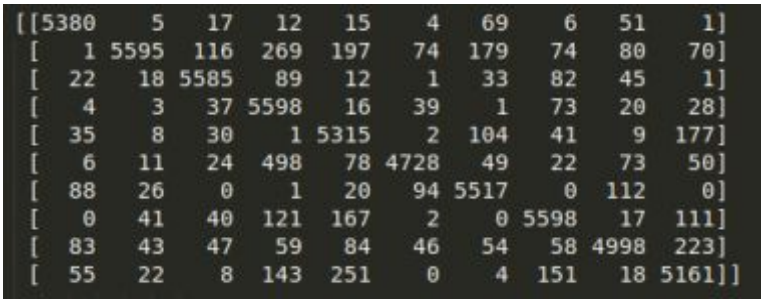
Naive Bayes:



```
[[5220    1   11   32    2    1   41    0  251    1]
 [   1 5184  583  238   86   22   85  340   80   36]
 [   9   24 5289  447    4    1    8   52   53    1]
 [   2    1  212 5390    1   33    0  127   31   22]
 [  14    2   44   12 5273    0   32   44   90  211]
 [   9    6   29  103   31 4958   46    2  169  186]
 [  78    7   89    8   15   90 5286    0  285    0]
 [   1   47  175  426   21    1    1 5323   60   42]
 [ 175    5   53  182   23    7   38   13 5112   87]
 [  25    5   62  151  221    4    0   55  184 5106]]
```

Assim como o perceptron, e o LDA os erros mais frequentes são na classificação do 1,9 e 10. Com mais erros no 1. Porém ainda erros consideráveis no 2.

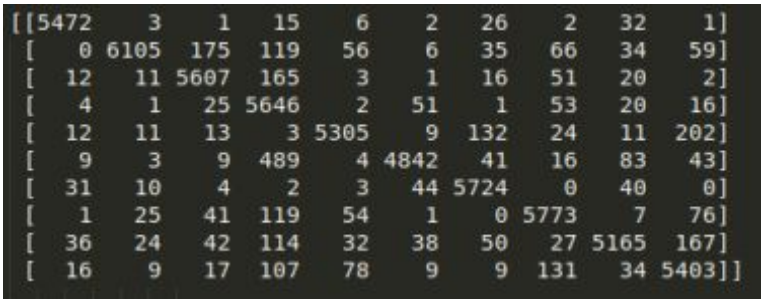
Logistic Regression:



```
[[5380    5   17   12   15    4   69    6   51    1]
 [   1 5595  116  269  197   74  179   74   80   70]
 [  22   18 5585   89   12    1   33   82   45    1]
 [   4    3   37 5598   16   39    1   73   20   28]
 [  35    8   30    1 5315    2  104   41    9  177]
 [   6   11   24  498   78 4728   49   22   73   50]
 [  88   26    0    1   20   94 5517    0  112    0]
 [   0   41   40  121  167    2    0 5598   17  111]
 [  83   43   47   59   84   46   54   58 4998  223]
 [  55   22    8  143  251    0    4  151   18 5161]]
```

Assim como nos anteriores os erros mais frequentes são na classificação do 1, 9 e 10. Com mais erros no 1.

KNN:



```
[[5472    3    1   15    6    2   26    2   32    1]
 [   0 6105  175  119   56    6   35   66   34   59]
 [  12   11 5607  165    3    1   16   51   20    2]
 [   4    1   25 5646    2   51    1   53   20   16]
 [  12   11   13    3 5305    9  132   24   11  202]
 [   9    3    9  489    4 4842   41   16   83   43]
 [  31   10    4    2    3   44 5724    0   40    0]
 [   1   25   41  119   54    1    0 5773    7   76]
 [  36   24   42  114   32   38   50   27 5165  167]
 [  16    9   17  107   78    9    9  131   34 5403]]
```

Essa classificador ja possui erros mais balanceados em quase todas as classes, com assertividades mais bem definidas que erros.

Conclusão

Apresentado todos os testes e resultado é possível verificar que os classificadores apresentam um comportamento parecido no decorrer da avaliação em relação a variação da base, o Naive Bayes apresenta um comportamento um pouco fora dos observados com bases menores, porém no decorrer do acréscimo da base ele segue o padrão. As acurácias observadas foram relativamente próximas, o KNN destacou-se em alguns testes, porém foi o que demorou mais tempo em todos os casos ao longo dos testes. O Perceptron apresentou um desempenho positivo maior ao longo dos testes, por ser o mais rápido e equilibrando com um bom desempenho.