Modelos Preditivos

Vinicius Sampaio

January 2022

1 Introdução

Nesta atividade iremos aplicar diversos métodos preditivos ao dataset MovieLens e verificaremos a capacidade de cada um desses métodos de predizer a avaliação média de cada filme no dataset.

2 Modelos Preditivos

Modelos preditivos são modelos que têm a capacidade de predizer uma categoria ou um valor dado informações sobre uma observação. Tais modelos na maioria das vezes se baseiam em um processo de aprendizado supervisionado para conseguirem fazer suas predições.

Aprendizado Supervisionado é um processo de aprendizado de máquina onde os dados utilizados no treinamento já estão classificados, ou seja, as predições já foram feitas - ou observadas. Desta forma o modelo pode, através de várias gerações, calcular o erro de suas predições de forma exata e corrigi-las.

Nesta atividade iremos utilizar as técnicas: KNN, Naive Bayes, Árvore de decisão, MLP e SVM.

2.1 KNN

K-nearest neighbor, também conhecido como algoritmo KNN, é um algoritmo não paramétrico que classifica pontos de dados com base em sua proximidade e associação a outros dados disponíveis[1]. Este algoritmo assume que pontos de dados semelhantes podem ser encontrados próximos uns dos outros. Como resultado, ele busca calcular a distância entre os pontos de dados, geralmente através da distância euclidiana, e então atribui uma categoria com base na categoria ou média mais frequente.

Sua facilidade de uso e baixo tempo de cálculo o tornam um algoritmo preferido pelos cientistas de dados, mas à medida que o conjunto de dados de teste cresce, o tempo de processamento aumenta, tornando-o menos atraente para tarefas de classificação. O KNN é normalmente usado para mecanismos de recomendação e reconhecimento de imagem.

2.2 Naive Bayes

Naive Bayes é uma abordagem de classificação que adota o princípio da independência condicional de classe do Teorema de Bayes[1]. Isso significa que a presença de um recurso não afeta a presença de outro na probabilidade de um determinado resultado, e cada preditor tem um efeito igual nesse resultado. Existem três tipos de classificadores Naïve Bayes: Multinomial Naïve Bayes, Bernoulli Naïve Bayes e Gaussian Naïve Bayes. Essa técnica é usada principalmente na classificação de texto, identificação de spam e sistemas de recomendação.

2.3 Árvore de Decisão

Árvores de decisão são modelos de suporte à decisão que classificam padrões usando uma sequência de regras bem definidas. São grafos semelhantes a árvores em que cada nó de ramificação representa uma opção entre várias alternativas e cada nó folha representa um resultado das escolhas cumulativas.[4]

2.4 MLP

O perceptron multicamada (MLP) é um suplemento da rede neural feed forward. Ela consiste em três tipos de camadas – a camada de entrada, a camada de saída e a camada oculta. A camada de entrada recebe o sinal de entrada a ser processado. A tarefa necessária, como previsão e classificação, é realizada pela camada de saída. Um número arbitrário de camadas ocultas que são colocadas entre as camadas de entrada e saída são o verdadeiro mecanismo computacional do MLP. Semelhante a uma rede feed forward em um MLP, os dados fluem na direção direta da camada de entrada para a camada de saída. Os neurônios no MLP são treinados com o algoritmo de aprendizado de retropropagação. As MLPs são projetadas para aproximar qualquer função contínua e podem resolver problemas que não são linearmente separáveis. Os principais casos de uso do MLP são classificação de padrões, reconhecimento, previsão e aproximação.[2]

2.5 SVM

Uma máquina de vetor de suporte é um modelo de aprendizado supervisionado popular desenvolvido por Vladimir Vapnik, usado para classificação e regressão de dados. Dito isso, normalmente é aproveitado para problemas de classificação, construindo um hiperplano onde a distância entre duas classes de pontos de dados é máxima. Esse hiperplano é conhecido como limite de decisão, separando as classes de pontos de dados (por exemplo, laranjas versus maçãs) em ambos os lados do plano.

3 Metodologia

3.1 Preparação dos Dados

O dataset MovieLens 25M[3] foi escolhido como o dataset a ser utilizado. O dataset contém informações sobre filmes, como título, avaliações de usuários e tags e scores genome, que representam o quão bem um filme se adequa a determinada categoria (tag).

Antes de que qualquer predição ou análise possa ser feita, os dados do dataset precisam ser organizados. Os dados vieram separados em 4 arquivos diferentes. Um arquivo com os filmes e seus respectivos ids, outro com o rating (de 0 a 5) de cada usuário e o id do filme, um arquivo com id do filme, id de genome tag e genome score e finalmente um arquivo com id de genome tag e nome da tag.

Foi feita uma união dos arquivos de forma que cada linha representa um filme com as seguintes características: id do filme, título, gênero, média das avaliações e uma coluna para cada tag genome contendo seu score.

No fim desta etapa temos um dataset de shape = (13816, 1129) e uma coluna com as avaliações médias.

3.2 Criação das classes

Os valores da média de avaliação dos resultados foram transformados em 3 classes ('bad', 'ok', 'good') da seguinte forma:

- filmes que tiveram a avaliação média maior do que 3.5 foram classificados como "good";
- filmes que tiveram a avaliação média entre 2.5 e 3.5 foram classificados como "ok"
- filmes que tiveram a avaliação média menor que 2.5 foram classificados como "bad"

3.3 PCA

Dada a quantidade enorme de colunas no dataset foi aplicado PCA. Inicialmente foi plotado o gráfico da variância explicada para determinar a quantidade de features no final do processo

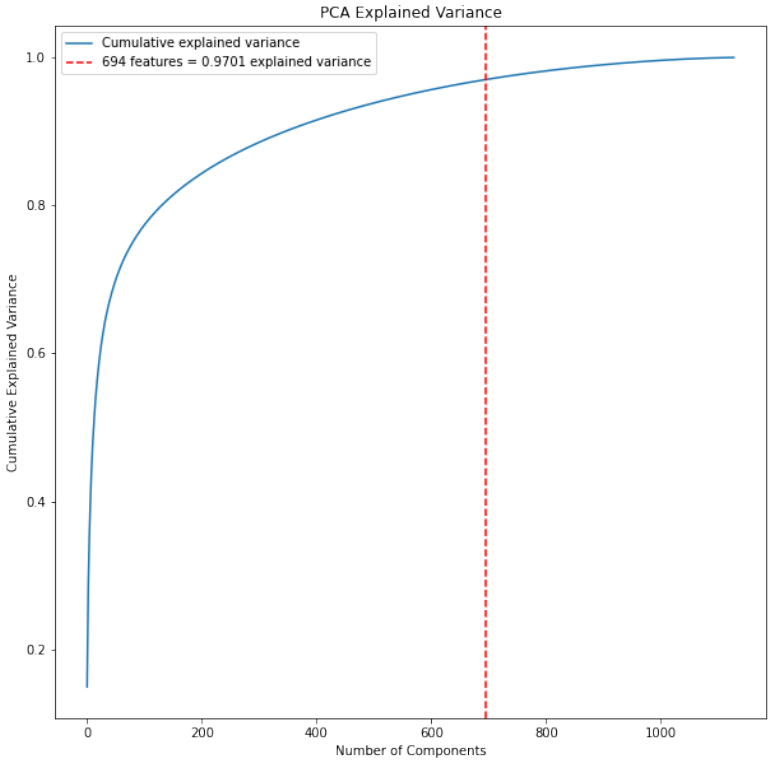


Figure 1: Variância Explicada

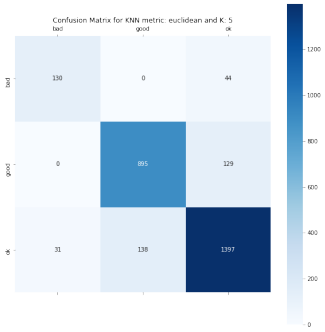
Foi determinado que uma variância de 0.97 seria adequada para prosseguir, resultando em 694 features, uma redução de $38,5\%$

4 Resultados

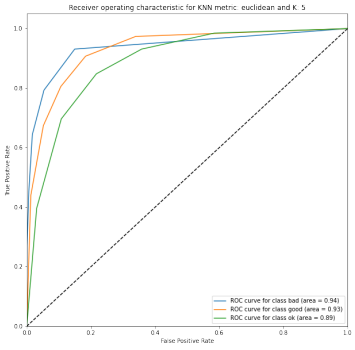
4.1 KNN

4.1.1 euclidean distance K = 5

	precision	recall	f1-score	support
bad	0.81	0.75	0.78	174
good	0.87	0.87	0.87	1024
ok	0.89	0.89	0.89	1566
accuracy			0.88	2764
macro avg	0.85	0.84	0.85	2764
weighted avg	0.88	0.88	0.88	2764



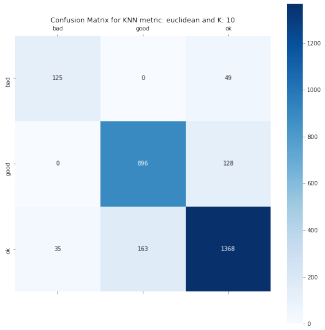
(a) Matriz de Confusão



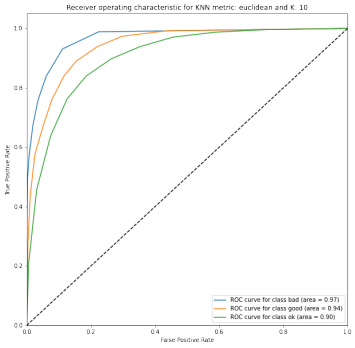
(b) ROC AUC

4.1.2 euclidean distance K = 10

	precision	recall	f1-score	support
bad	0.78	0.72	0.75	174
good	0.85	0.88	0.86	1024
ok	0.89	0.87	0.88	1566
accuracy			0.86	2764
macro avg	0.84	0.82	0.83	2764
weighted avg	0.86	0.86	0.86	2764



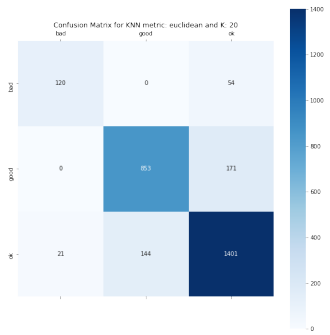
(a) Matriz de Confusão



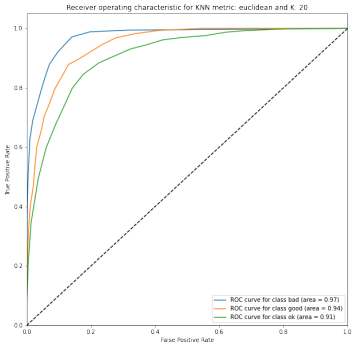
(b) ROC AUC

4.1.3 euclidean distance K = 20

	precision	recall	f1-score	support
bad	0.85	0.69	0.76	174
good	0.86	0.83	0.84	1024
ok	0.86	0.89	0.88	1566
accuracy			0.86	2764
macro avg	0.86	0.81	0.83	2764
weighted avg	0.86	0.86	0.86	2764



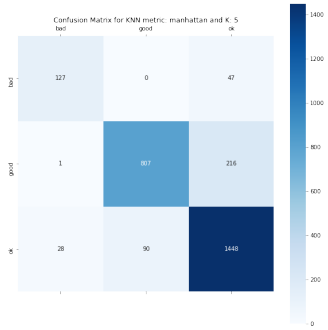
(a) Matriz de Confusão



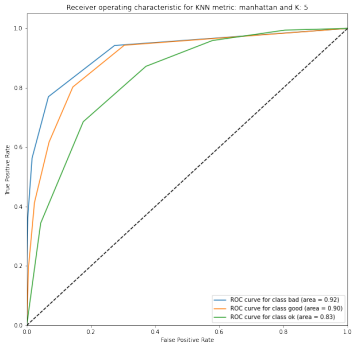
(b) ROC AUC

4.1.4 manhattan distance K = 5

	precision	recall	f1-score	support
bad	0.81	0.73	0.77	174
good	0.90	0.79	0.84	1024
ok	0.85	0.92	0.88	1566
accuracy			0.86	2764
macro avg	0.85	0.81	0.83	2764
weighted avg	0.86	0.86	0.86	2764



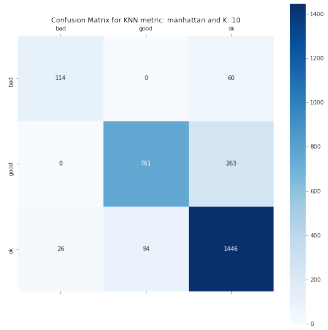
(a) Matriz de Confusão



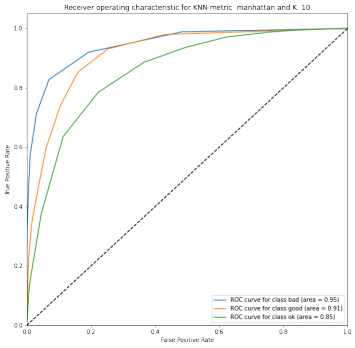
(b) ROC AUC

4.1.5 manhattan distance K = 10

	precision	recall	f1-score	support
bad	0.81	0.66	0.73	174
good	0.89	0.74	0.81	1024
ok	0.82	0.92	0.87	1566
accuracy			0.84	2764
macro avg	0.84	0.77	0.80	2764
weighted avg	0.84	0.84	0.84	2764



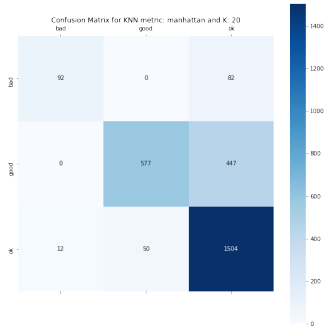
(a) Matriz de Confusão



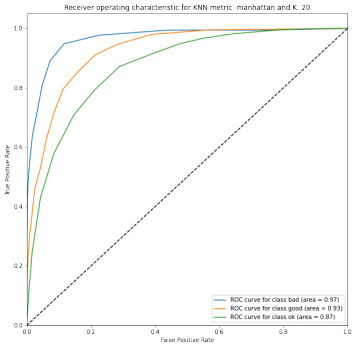
(b) ROC AUC

4.1.6 manhattan distance K = 20

	precision	recall	f1-score	support
bad	0.88	0.53	0.66	174
good	0.92	0.56	0.70	1024
ok	0.74	0.96	0.84	1566
accuracy			0.79	2764
macro avg	0.85	0.68	0.73	2764
weighted avg	0.82	0.79	0.77	2764



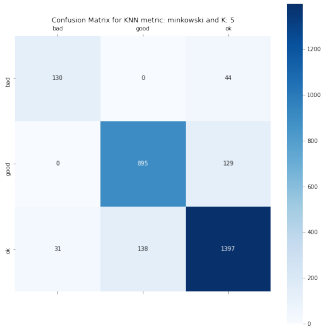
(a) Matriz de Confusão



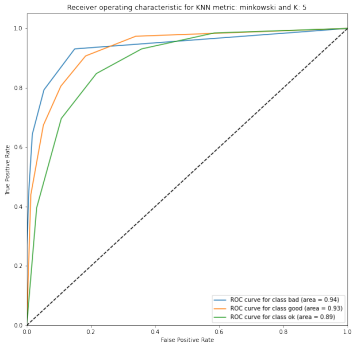
(b) ROC AUC

4.1.7 minkowski distance K = 5

	precision	recall	f1-score	support
bad	0.81	0.75	0.78	174
good	0.87	0.87	0.87	1024
ok	0.89	0.89	0.89	1566
accuracy			0.88	2764
macro avg	0.85	0.84	0.85	2764
weighted avg	0.88	0.88	0.88	2764



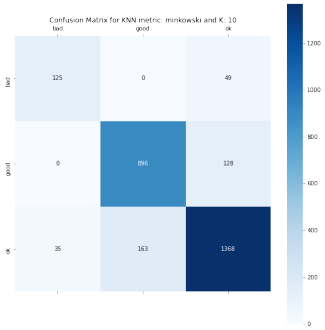
(a) Matriz de Confusão



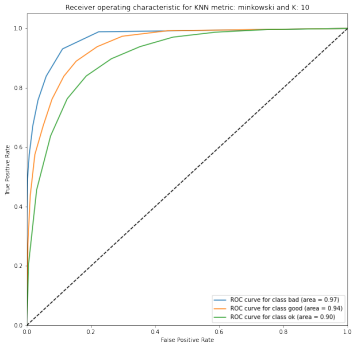
(b) ROC AUC

4.1.8 minkowski distance K = 10

	precision	recall	f1-score	support
bad	0.78	0.72	0.75	174
good	0.85	0.88	0.86	1024
ok	0.89	0.87	0.88	1566
accuracy			0.86	2764
macro avg	0.84	0.82	0.83	2764
weighted avg	0.86	0.86	0.86	2764



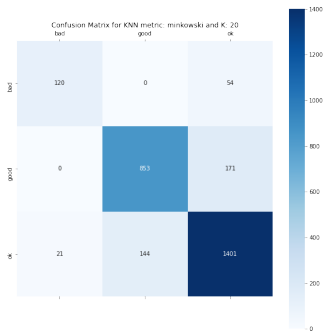
(a) Matriz de Confusão



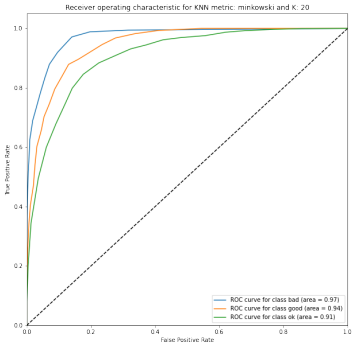
(b) ROC AUC

4.1.9 minkowski distance K = 20

	precision	recall	f1-score	support
bad	0.85	0.69	0.76	174
good	0.86	0.83	0.84	1024
ok	0.86	0.89	0.88	1566
accuracy			0.86	2764
macro avg	0.86	0.81	0.83	2764
weighted avg	0.86	0.86	0.86	2764



(a) Matriz de Confusão

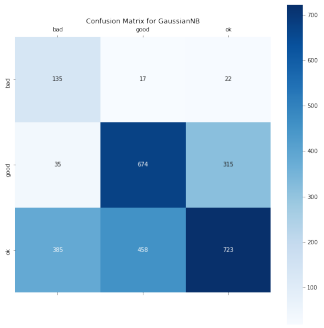


(b) ROC AUC

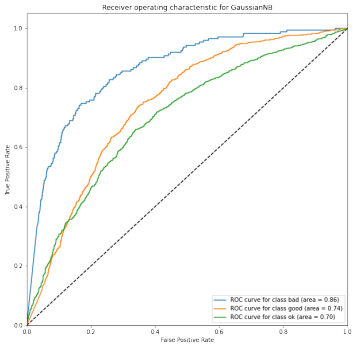
4.2 Naive Bayes

4.2.1 GaussianNB

	precision	recall	f1-score	support
bad	0.24	0.78	0.37	174
good	0.59	0.66	0.62	1024
ok	0.68	0.46	0.55	1566
accuracy			0.55	2764
macro avg	0.50	0.63	0.51	2764
weighted avg	0.62	0.55	0.57	2764



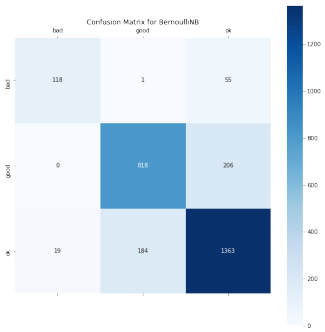
(a) Matriz de Confusão



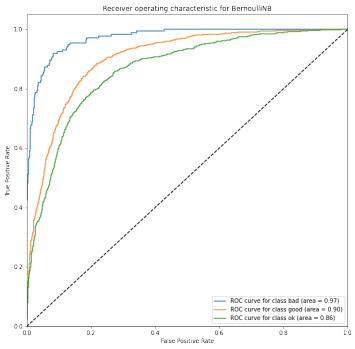
(b) ROC AUC

4.2.2 BernoulliNB

	precision	recall	f1-score	support
bad	0.86	0.68	0.76	174
good	0.82	0.80	0.81	1024
ok	0.84	0.87	0.85	1566
accuracy			0.83	2764
macro avg	0.84	0.78	0.81	2764
weighted avg	0.83	0.83	0.83	2764



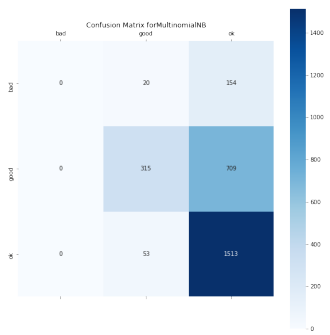
(a) Matriz de Confusão



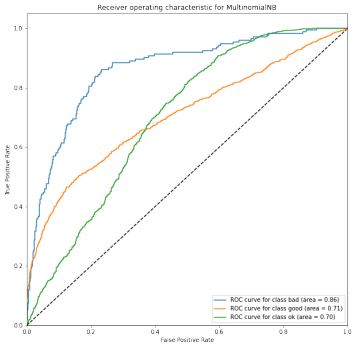
(b) ROC AUC

4.2.3 MultinomialNB

	precision	recall	f1-score	support
bad	0.00	0.00	0.00	174
good	0.81	0.31	0.45	1024
ok	0.64	0.97	0.77	1566
accuracy			0.66	2764
macro avg	0.48	0.42	0.40	2764
weighted avg	0.66	0.66	0.60	2764



(a) Matriz de Confusão

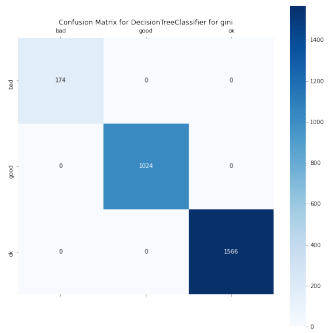


(b) ROC AUC

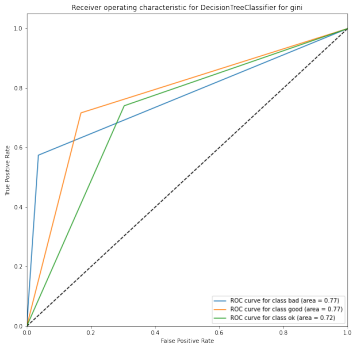
4.3 Árvore de Decisão

4.3.1 gini

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



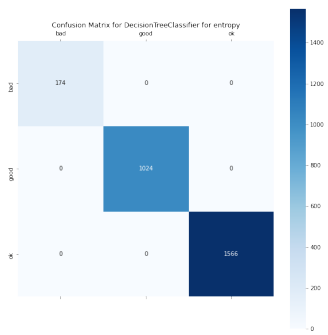
(a) Matriz de Confusão



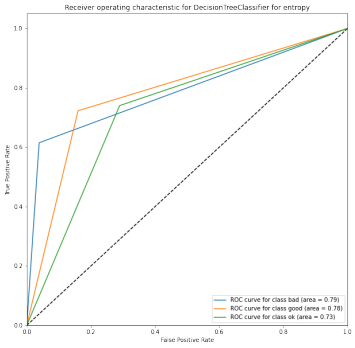
(b) ROC AUC

4.3.2 entropy

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



(a) Matriz de Confusão

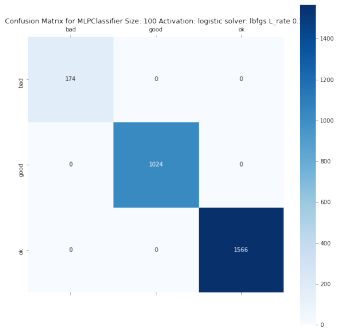


(b) ROC AUC

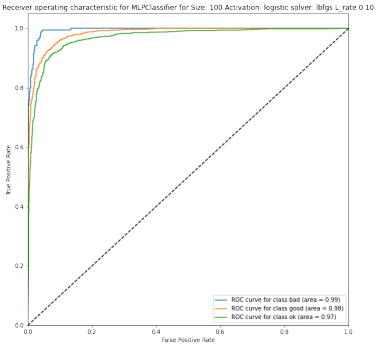
4.4 MLP

4.4.1 MLP Solver: lbfgs Ativação: logistic Hidden Layer Size:100

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



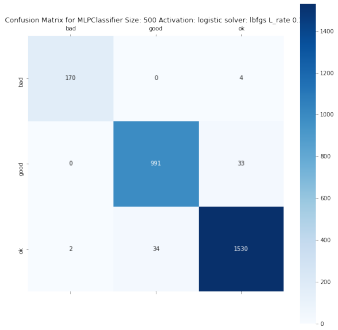
(a) Matriz de Confusão



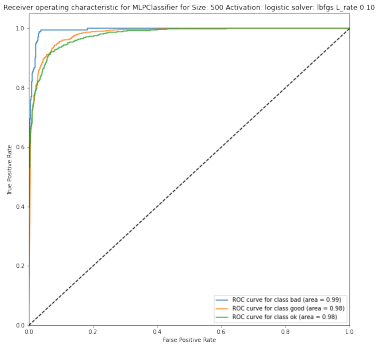
(b) ROC AUC

4.4.2 MLP Solver: lbfgs Ativação: logistic Hidden Layer Size:500

	precision	recall	f1-score	support
bad	0.99	0.98	0.98	174
good	0.97	0.97	0.97	1024
ok	0.98	0.98	0.98	1566
accuracy			0.97	2764
macro avg	0.98	0.97	0.98	2764
weighted avg	0.97	0.97	0.97	2764



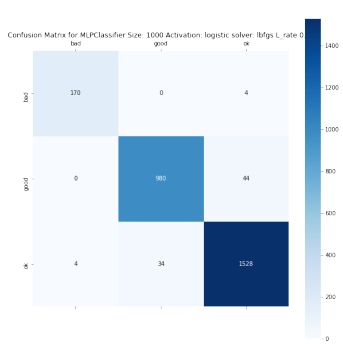
(a) Matriz de Confusão



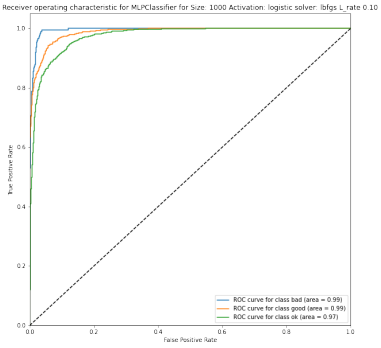
(b) ROC AUC

4.4.3 MLP Solver: lbfgs Ativação: logistic Hidden Layer Size:1000

	precision	recall	f1-score	support
bad	0.98	0.98	0.98	174
good	0.97	0.96	0.96	1024
ok	0.97	0.98	0.97	1566
accuracy			0.97	2764
macro avg	0.97	0.97	0.97	2764
weighted avg	0.97	0.97	0.97	2764



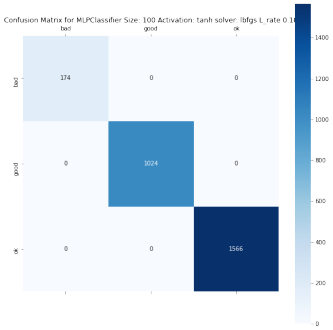
(a) Matriz de Confusão



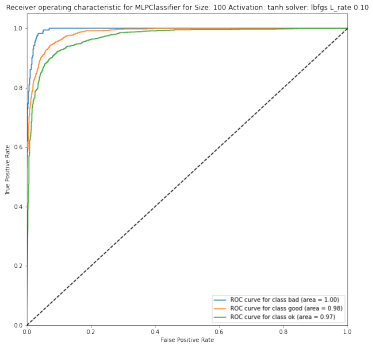
(b) ROC AUC

4.4.4 MLP Solver: lbfgs Ativação: tanh Hidden Layer Size:100

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



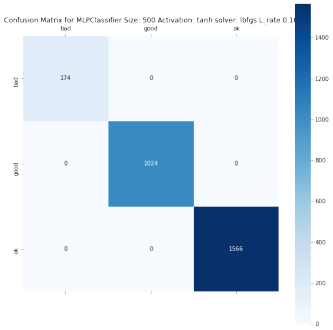
(a) Matriz de Confusão



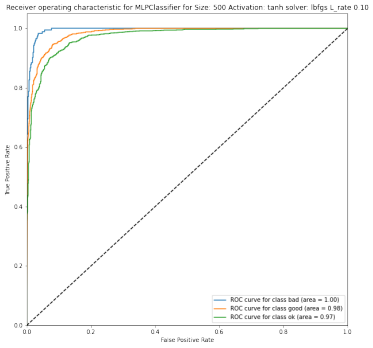
(b) ROC AUC

4.4.5 MLP Solver: lbfgs Ativação: tanh Hidden Layer Size:500

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



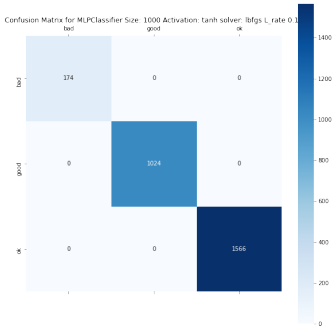
(a) Matriz de Confusão



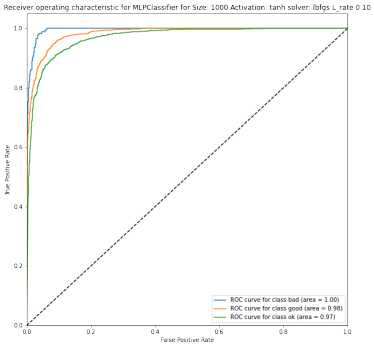
(b) ROC AUC

4.4.6 MLP Solver: lbfgs Ativação: tanh Hidden Layer Size:1000

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



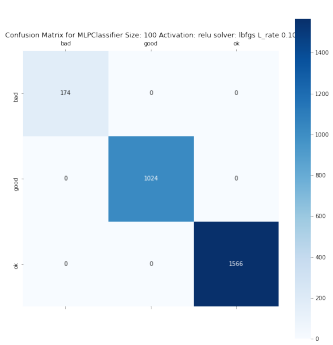
(a) Matriz de Confusão



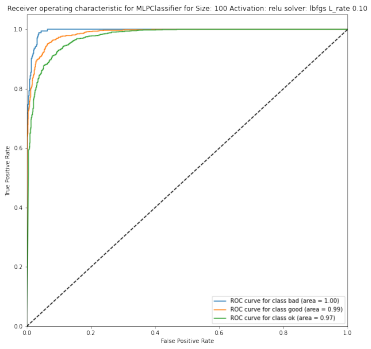
(b) ROC AUC

4.4.7 MLP Solver: lbfgs Ativação: relu Hidden Layer Size:100

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



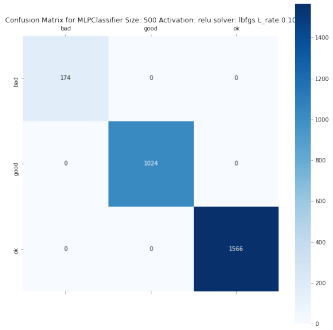
(a) Matriz de Confusão



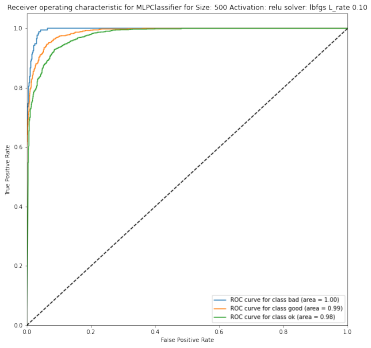
(b) ROC AUC

4.4.8 MLP Solver: lbfgs Ativação: relu Hidden Layer Size:500

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



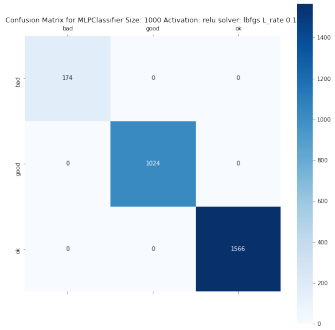
(a) Matriz de Confusão



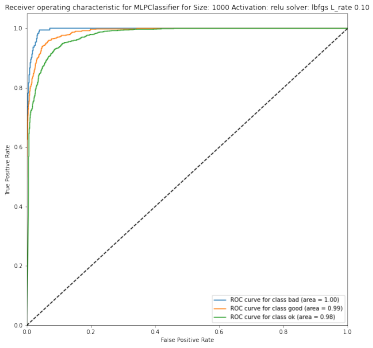
(b) ROC AUC

4.4.9 MLP Solver: lbfgs Ativação: relu Hidden Layer Size:1000

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



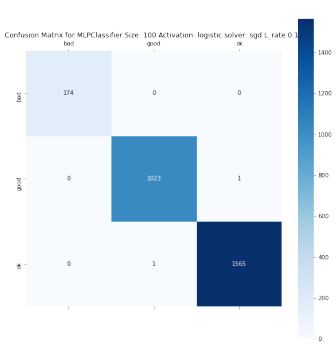
(a) Matriz de Confusão



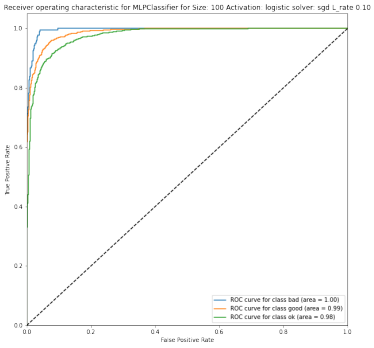
(b) ROC AUC

4.4.10 MLP Solver: sgd Ativação: logistic Hidden Layer Size:100

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



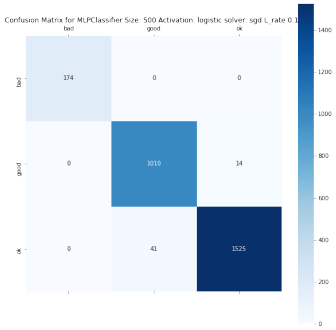
(a) Matriz de Confusão



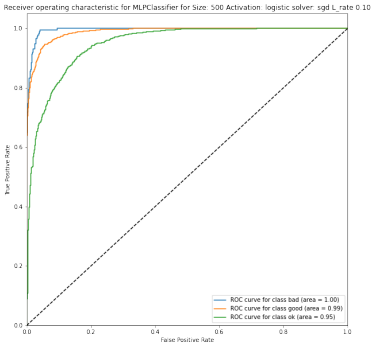
(b) ROC AUC

4.4.11 MLP Solver: sgd Ativação: logistic Hidden Layer Size:500

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	0.96	0.99	0.97	1024
ok	0.99	0.97	0.98	1566
accuracy			0.98	2764
macro avg	0.98	0.99	0.99	2764
weighted avg	0.98	0.98	0.98	2764



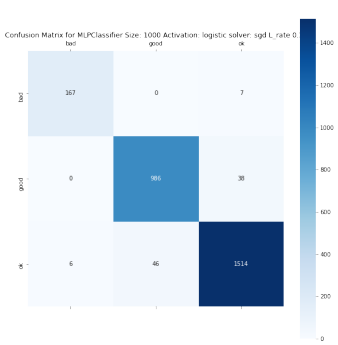
(a) Matriz de Confusão



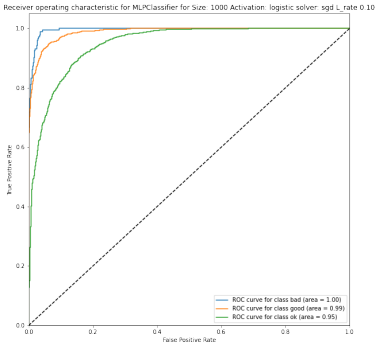
(b) ROC AUC

4.4.12 MLP Solver: sgd Ativação: logistic Hidden Layer Size:1000

	precision	recall	f1-score	support
bad	0.97	0.96	0.96	174
good	0.96	0.96	0.96	1024
ok	0.97	0.97	0.97	1566
accuracy			0.96	2764
macro avg	0.96	0.96	0.96	2764
weighted avg	0.96	0.96	0.96	2764



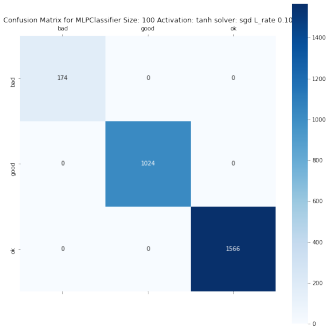
(a) Matriz de Confusão



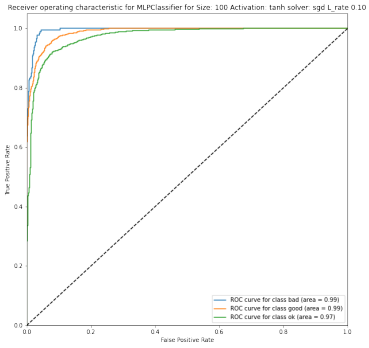
(b) ROC AUC

4.4.13 MLP Solver: sgd Ativação: tanh Hidden Layer Size:100

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



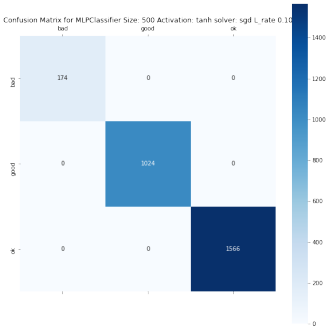
(a) Matriz de Confusão



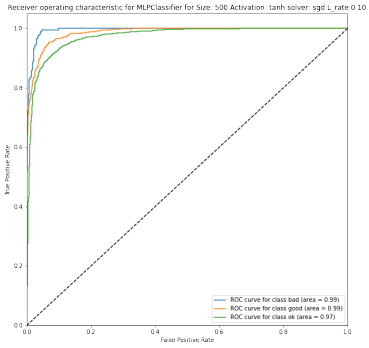
(b) ROC AUC

4.4.14 MLP Solver: sgd Ativação: tanh Hidden Layer Size:500

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



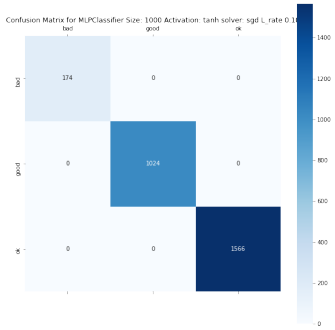
(a) Matriz de Confusão



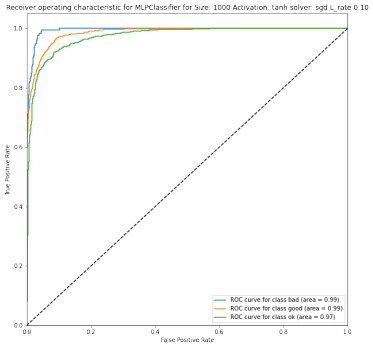
(b) ROC AUC

4.4.15 MLP Solver: sgd Ativação: tanh Hidden Layer Size:1000

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



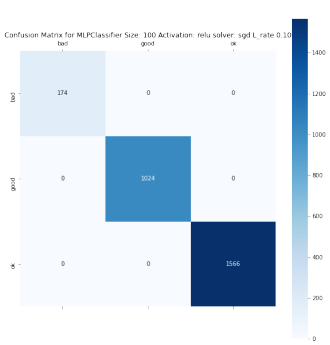
(a) Matriz de Confusão



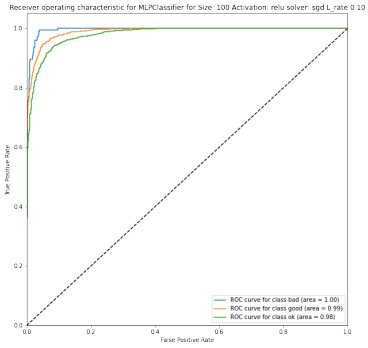
(b) ROC AUC

4.4.16 MLP Solver: sgd Ativação: relu Hidden Layer Size:100

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



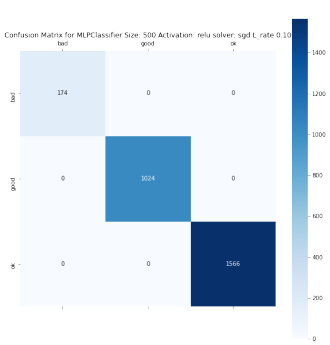
(a) Matriz de Confusão



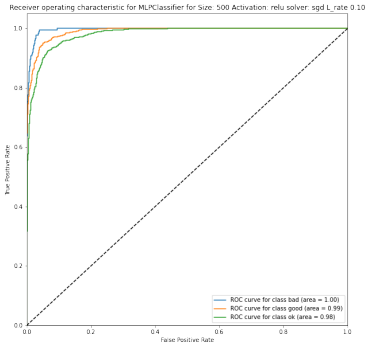
(b) ROC AUC

4.4.17 MLP Solver: sgd Ativação: relu Hidden Layer Size:500

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



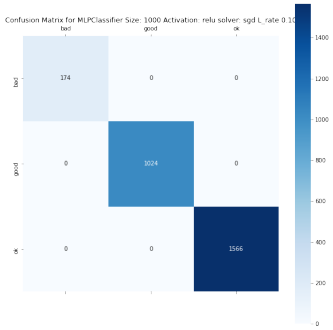
(a) Matriz de Confusão



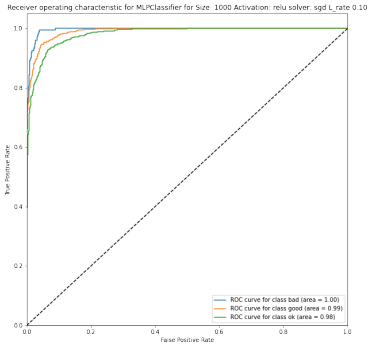
(b) ROC AUC

4.4.18 MLP Solver: sgd Ativação: relu Hidden Layer Size:1000

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



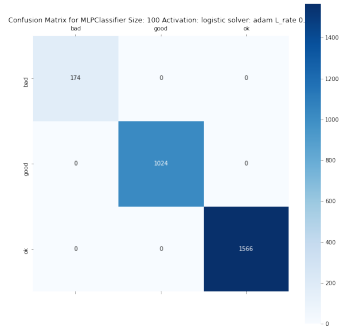
(a) Matriz de Confusão



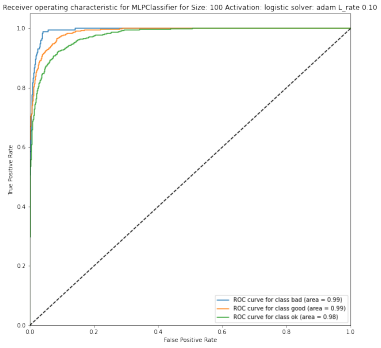
(b) ROC AUC

4.4.19 MLP Solver: adam Ativação: logistic Hidden Layer Size:100

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



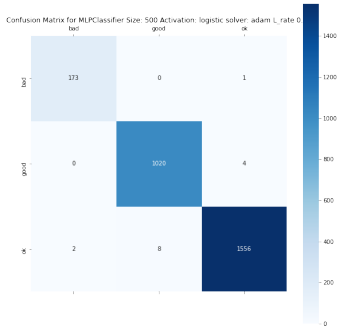
(a) Matriz de Confusão



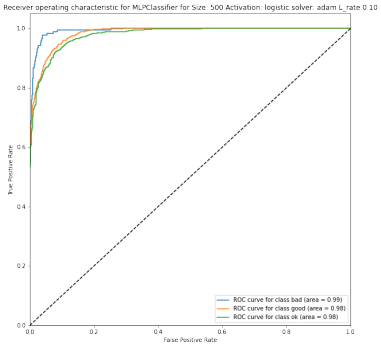
(b) ROC AUC

4.4.20 MLP Solver: adam Ativação: logistic Hidden Layer Size:500

	precision	recall	f1-score	support
bad	0.99	0.99	0.99	174
good	0.99	1.00	0.99	1024
ok	1.00	0.99	1.00	1566
accuracy			0.99	2764
macro avg	0.99	0.99	0.99	2764
weighted avg	0.99	0.99	0.99	2764



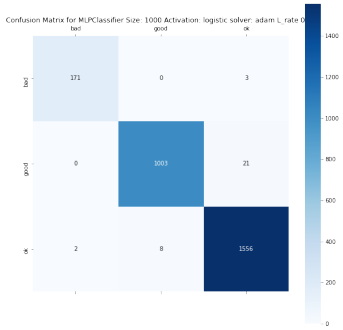
(a) Matriz de Confusão



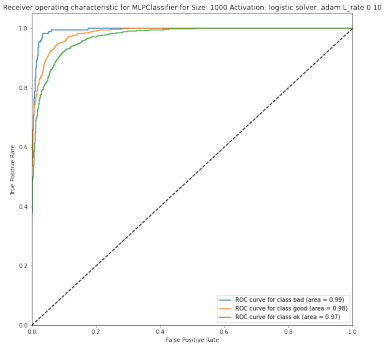
(b) ROC AUC

4.4.21 MLP Solver: adam Ativação: logistic Hidden Layer Size:1000

	precision	recall	f1-score	support
bad	0.99	0.98	0.99	174
good	0.99	0.98	0.99	1024
ok	0.98	0.99	0.99	1566
accuracy			0.99	2764
macro avg	0.99	0.99	0.99	2764
weighted avg	0.99	0.99	0.99	2764



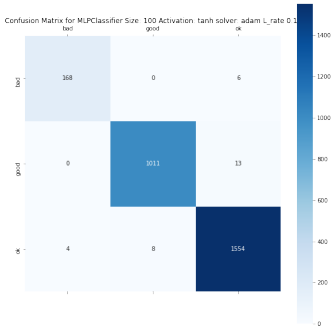
(a) Matriz de Confusão



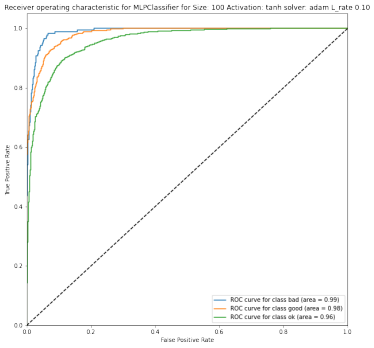
(b) ROC AUC

4.4.22 MLP Solver: adam Ativação: tanh Hidden Layer Size:100

	precision	recall	f1-score	support
bad	0.98	0.97	0.97	174
good	0.99	0.99	0.99	1024
ok	0.99	0.99	0.99	1566
accuracy			0.99	2764
macro avg	0.99	0.98	0.98	2764
weighted avg	0.99	0.99	0.99	2764



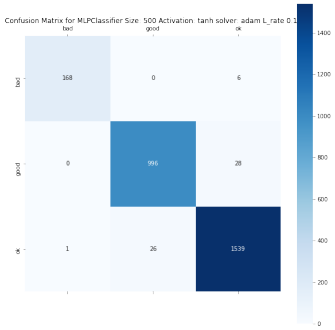
(a) Matriz de Confusão



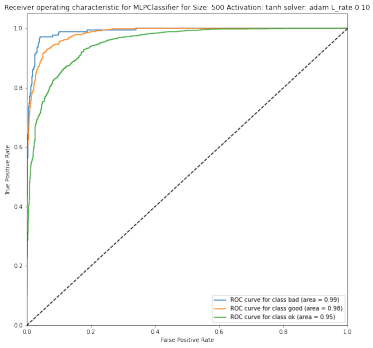
(b) ROC AUC

4.4.23 MLP Solver: adam Ativação: tanh Hidden Layer Size:500

	precision	recall	f1-score	support
bad	0.99	0.97	0.98	174
good	0.97	0.97	0.97	1024
ok	0.98	0.98	0.98	1566
accuracy			0.98	2764
macro avg	0.98	0.97	0.98	2764
weighted avg	0.98	0.98	0.98	2764



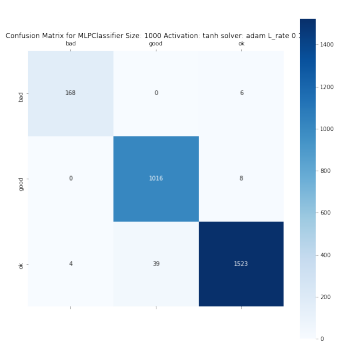
(a) Matriz de Confusão



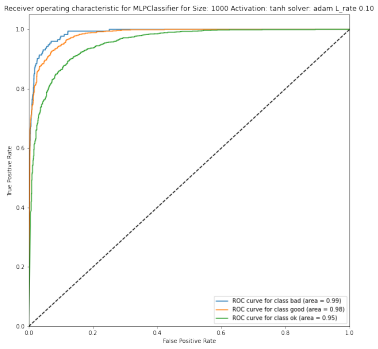
(b) ROC AUC

4.4.24 MLP Solver: adam Ativação: tanh Hidden Layer Size:1000

	precision	recall	f1-score	support
bad	0.98	0.97	0.97	174
good	0.96	0.99	0.98	1024
ok	0.99	0.97	0.98	1566
accuracy			0.98	2764
macro avg	0.98	0.98	0.98	2764
weighted avg	0.98	0.98	0.98	2764



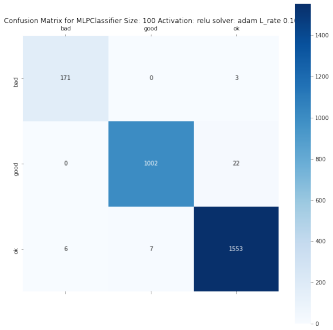
(a) Matriz de Confusão



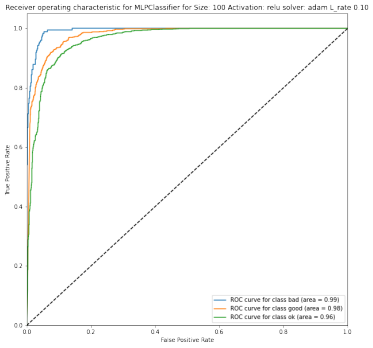
(b) ROC AUC

4.4.25 MLP Solver: adam Ativação: relu Hidden Layer Size:100

	precision	recall	f1-score	support
bad	0.97	0.98	0.97	174
good	0.99	0.98	0.99	1024
ok	0.98	0.99	0.99	1566
accuracy			0.99	2764
macro avg	0.98	0.98	0.98	2764
weighted avg	0.99	0.99	0.99	2764



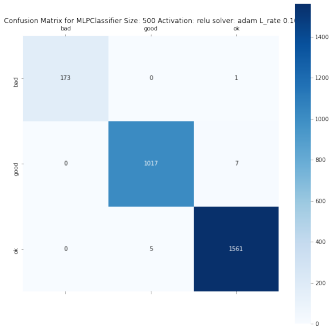
(a) Matriz de Confusão



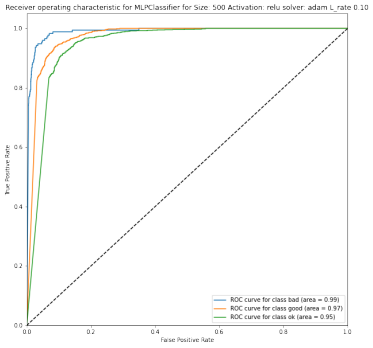
(b) ROC AUC

4.4.26 MLP Solver: adam Ativação: relu Hidden Layer Size:500

	precision	recall	f1-score	support
bad	1.00	0.99	1.00	174
good	1.00	0.99	0.99	1024
ok	0.99	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	0.99	1.00	2764
weighted avg	1.00	1.00	1.00	2764



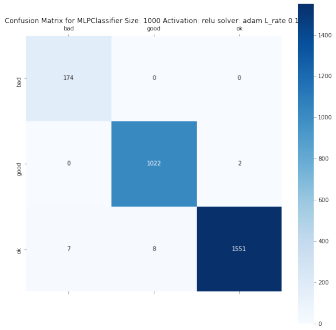
(a) Matriz de Confusão



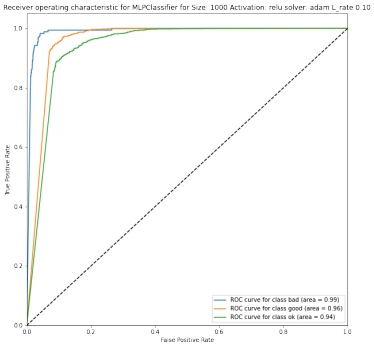
(b) ROC AUC

4.4.27 MLP Solver: adam Ativação: relu Hidden Layer Size:1000

	precision	recall	f1-score	support
bad	0.96	1.00	0.98	174
good	0.99	1.00	1.00	1024
ok	1.00	0.99	0.99	1566
accuracy			0.99	2764
macro avg	0.98	1.00	0.99	2764
weighted avg	0.99	0.99	0.99	2764



(a) Matriz de Confusão

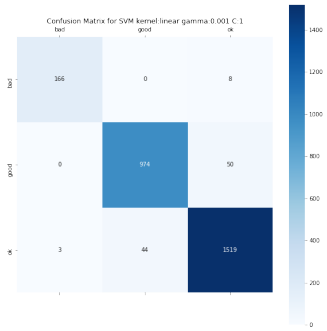


(b) ROC AUC

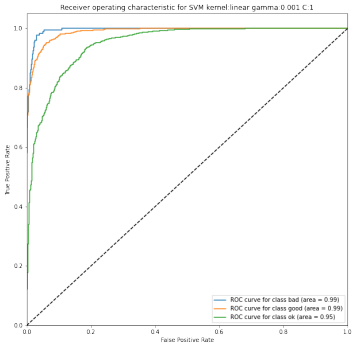
4.5 SVM

4.5.1 SVM Kernel: linear C: 1 Gamma: 0.001

	precision	recall	f1-score	support
bad	0.98	0.95	0.97	174
good	0.96	0.95	0.95	1024
ok	0.96	0.97	0.97	1566
accuracy			0.96	2764
macro avg	0.97	0.96	0.96	2764
weighted avg	0.96	0.96	0.96	2764



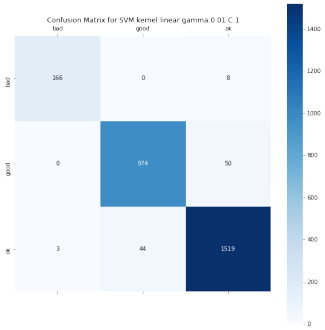
(a) Matriz de Confusão



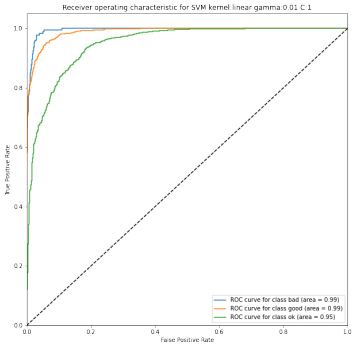
(b) ROC AUC

4.5.2 SVM Kernel: linear C: 1 Gamma: 0.01

	precision	recall	f1-score	support
bad	0.98	0.95	0.97	174
good	0.96	0.95	0.95	1024
ok	0.96	0.97	0.97	1566
accuracy			0.96	2764
macro avg	0.97	0.96	0.96	2764
weighted avg	0.96	0.96	0.96	2764



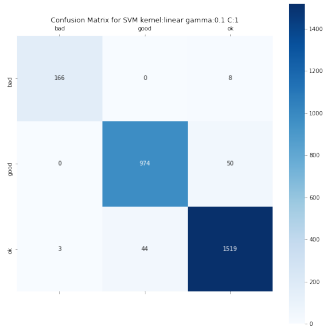
(a) Matriz de Confusão



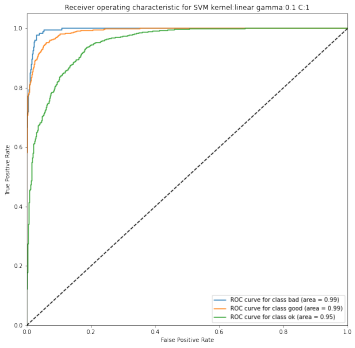
(b) ROC AUC

4.5.3 SVM Kernel: linear C: 1 Gamma: 0.1

	precision	recall	f1-score	support
bad	0.98	0.95	0.97	174
good	0.96	0.95	0.95	1024
ok	0.96	0.97	0.97	1566
accuracy			0.96	2764
macro avg	0.97	0.96	0.96	2764
weighted avg	0.96	0.96	0.96	2764



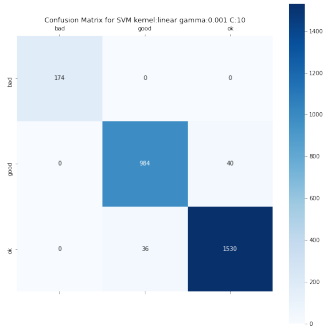
(a) Matriz de Confusão



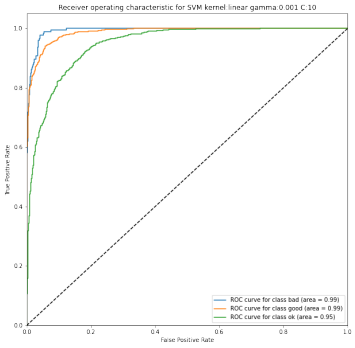
(b) ROC AUC

4.5.4 SVM Kernel: linear C: 10 Gamma: 0.001

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	0.96	0.96	0.96	1024
ok	0.97	0.98	0.98	1566
accuracy			0.97	2764
macro avg	0.98	0.98	0.98	2764
weighted avg	0.97	0.97	0.97	2764



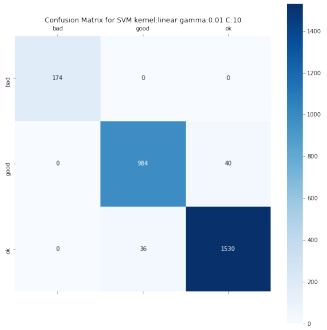
(a) Matriz de Confusão



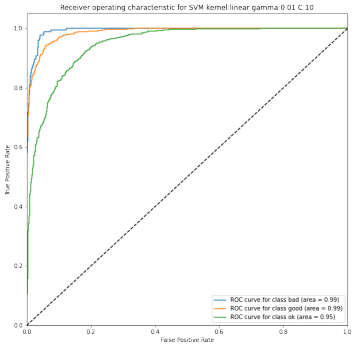
(b) ROC AUC

4.5.5 SVM Kernel: linear C: 10 Gamma: 0.01

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	0.96	0.96	0.96	1024
ok	0.97	0.98	0.98	1566
accuracy			0.97	2764
macro avg	0.98	0.98	0.98	2764
weighted avg	0.97	0.97	0.97	2764



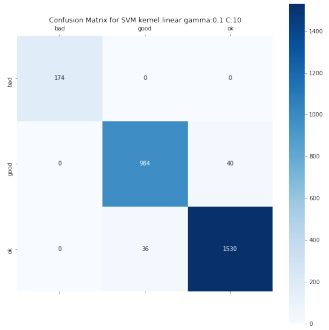
(a) Matriz de Confusão



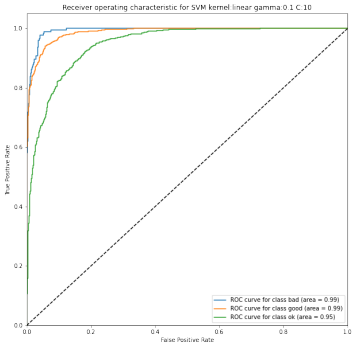
(b) ROC AUC

4.5.6 SVM Kernel: linear C: 10 Gamma: 0.1

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	0.96	0.96	0.96	1024
ok	0.97	0.98	0.98	1566
accuracy			0.97	2764
macro avg	0.98	0.98	0.98	2764
weighted avg	0.97	0.97	0.97	2764



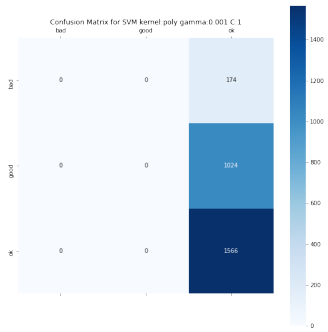
(a) Matriz de Confusão



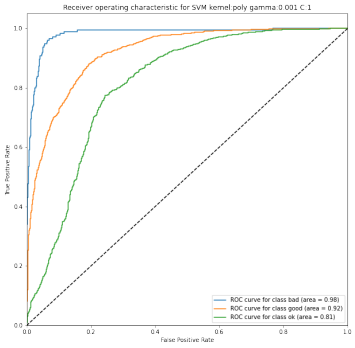
(b) ROC AUC

4.5.7 SVM Kernel: poly C: 1 Gamma: 0.001

	precision	recall	f1-score	support
bad	0.00	0.00	0.00	174
good	0.00	0.00	0.00	1024
ok	0.57	1.00	0.72	1566
accuracy			0.57	2764
macro avg	0.19	0.33	0.24	2764
weighted avg	0.32	0.57	0.41	2764



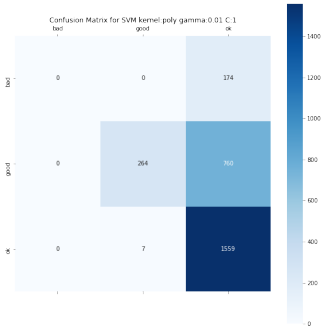
(a) Matriz de Confusão



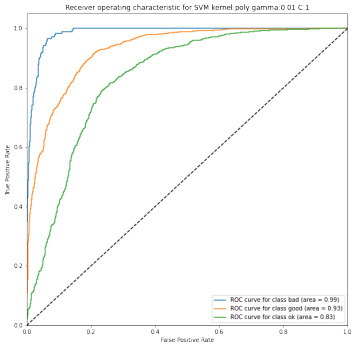
(b) ROC AUC

4.5.8 SVM Kernel: poly C: 1 Gamma: 0.01

	precision	recall	f1-score	support
bad	0.00	0.00	0.00	174
good	0.97	0.26	0.41	1024
ok	0.63	1.00	0.77	1566
accuracy			0.66	2764
macro avg	0.53	0.42	0.39	2764
weighted avg	0.72	0.66	0.59	2764



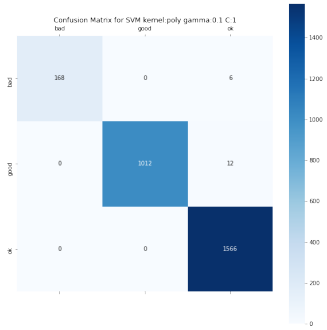
(a) Matriz de Confusão



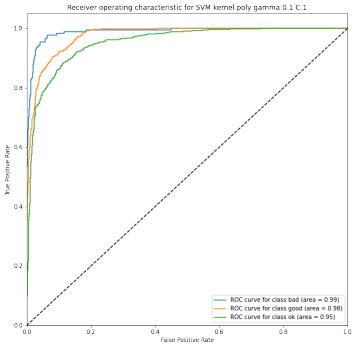
(b) ROC AUC

4.5.9 SVM Kernel: poly C: 1 Gamma: 0.1

	precision	recall	f1-score	support
bad	1.00	0.97	0.98	174
good	1.00	0.99	0.99	1024
ok	0.99	1.00	0.99	1566
accuracy			0.99	2764
macro avg	1.00	0.98	0.99	2764
weighted avg	0.99	0.99	0.99	2764



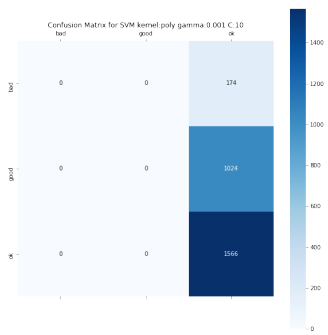
(a) Matriz de Confusão



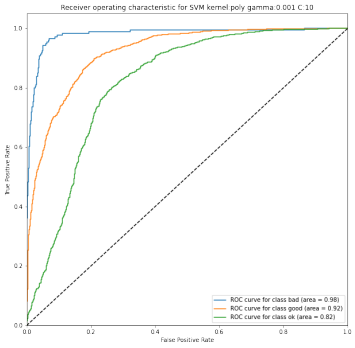
(b) ROC AUC

4.5.10 SVM Kernel: poly C: 10 Gamma: 0.001

	precision	recall	f1-score	support
bad	0.00	0.00	0.00	174
good	0.00	0.00	0.00	1024
ok	0.57	1.00	0.72	1566
accuracy			0.57	2764
macro avg	0.19	0.33	0.24	2764
weighted avg	0.32	0.57	0.41	2764



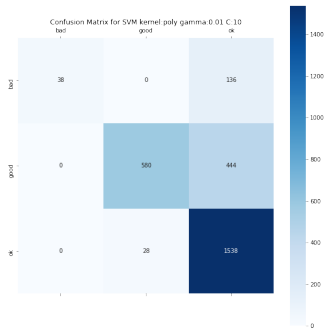
(a) Matriz de Confusão



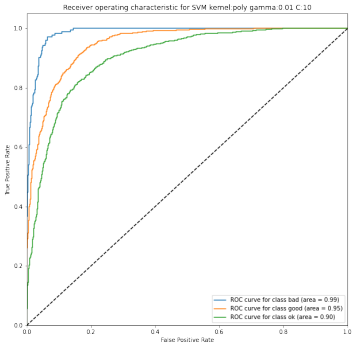
(b) ROC AUC

4.5.11 SVM Kernel: poly C: 10 Gamma: 0.01

	precision	recall	f1-score	support
bad	1.00	0.22	0.36	174
good	0.95	0.57	0.71	1024
ok	0.73	0.98	0.83	1566
accuracy			0.78	2764
macro avg	0.89	0.59	0.63	2764
weighted avg	0.83	0.78	0.76	2764



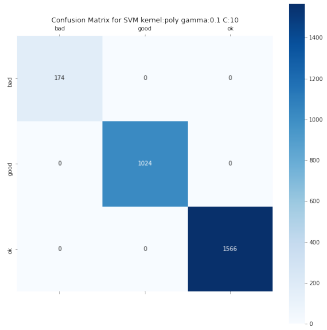
(a) Matriz de Confusão



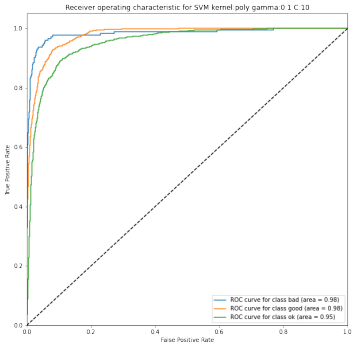
(b) ROC AUC

4.5.12 SVM Kernel: poly C: 10 Gamma: 0.1

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



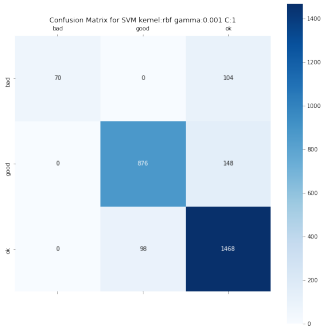
(a) Matriz de Confusão



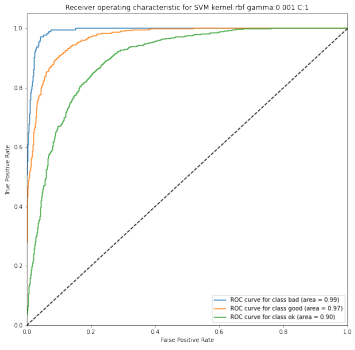
(b) ROC AUC

4.5.13 SVM Kernel: rbf C: 1 Gamma: 0.001

	precision	recall	f1-score	support
bad	1.00	0.40	0.57	174
good	0.90	0.86	0.88	1024
ok	0.85	0.94	0.89	1566
accuracy			0.87	2764
macro avg	0.92	0.73	0.78	2764
weighted avg	0.88	0.87	0.87	2764



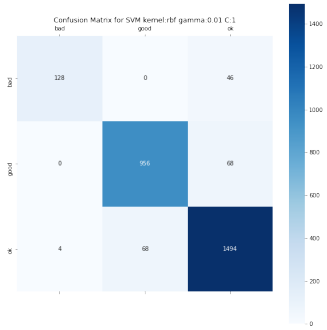
(a) Matriz de Confusão



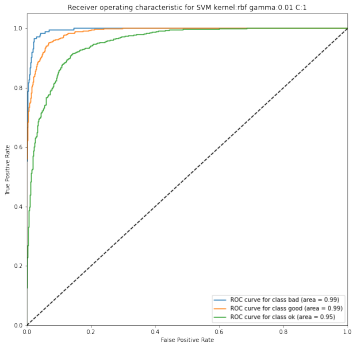
(b) ROC AUC

4.5.14 SVM Kernel: rbf C: 1 Gamma: 0.01

	precision	recall	f1-score	support
bad	0.97	0.74	0.84	174
good	0.93	0.93	0.93	1024
ok	0.93	0.95	0.94	1566
accuracy			0.93	2764
macro avg	0.94	0.87	0.90	2764
weighted avg	0.93	0.93	0.93	2764



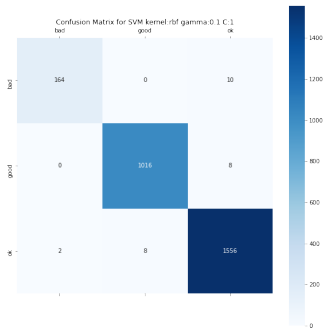
(a) Matriz de Confusão



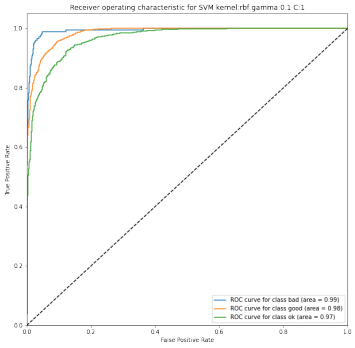
(b) ROC AUC

4.5.15 SVM Kernel: rbf C: 1 Gamma: 0.1

	precision	recall	f1-score	support
bad	0.99	0.94	0.96	174
good	0.99	0.99	0.99	1024
ok	0.99	0.99	0.99	1566
accuracy			0.99	2764
macro avg	0.99	0.98	0.98	2764
weighted avg	0.99	0.99	0.99	2764



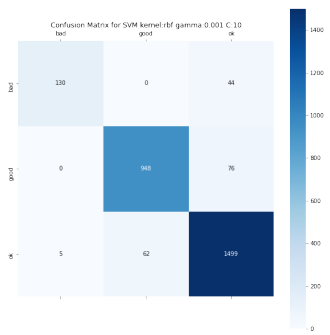
(a) Matriz de Confusão



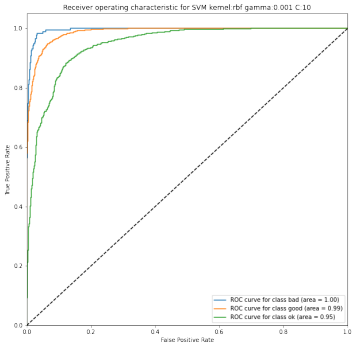
(b) ROC AUC

4.5.16 SVM Kernel: rbf C: 10 Gamma: 0.001

	precision	recall	f1-score	support
bad	0.96	0.75	0.84	174
good	0.94	0.93	0.93	1024
ok	0.93	0.96	0.94	1566
accuracy			0.93	2764
macro avg	0.94	0.88	0.90	2764
weighted avg	0.93	0.93	0.93	2764



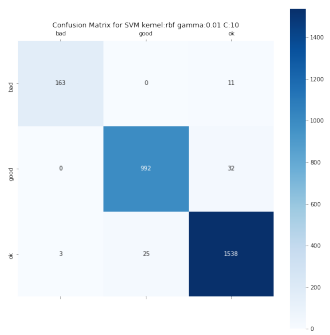
(a) Matriz de Confusão



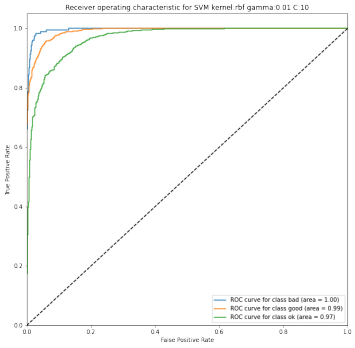
(b) ROC AUC

4.5.17 SVM Kernel: rbf C: 10 Gamma: 0.01

	precision	recall	f1-score	support
bad	0.98	0.94	0.96	174
good	0.98	0.97	0.97	1024
ok	0.97	0.98	0.98	1566
accuracy			0.97	2764
macro avg	0.98	0.96	0.97	2764
weighted avg	0.97	0.97	0.97	2764



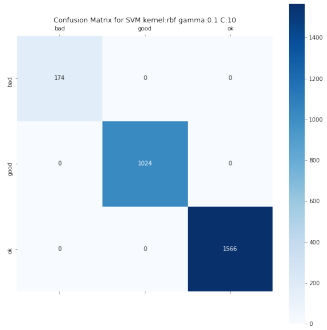
(a) Matriz de Confusão



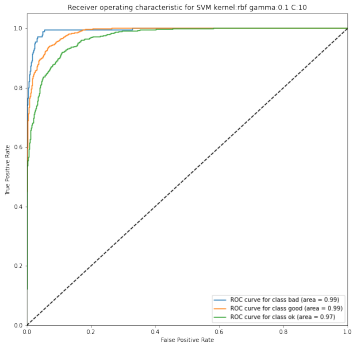
(b) ROC AUC

4.5.18 SVM Kernel: rbf C: 10 Gamma: 0.1

	precision	recall	f1-score	support
bad	1.00	1.00	1.00	174
good	1.00	1.00	1.00	1024
ok	1.00	1.00	1.00	1566
accuracy			1.00	2764
macro avg	1.00	1.00	1.00	2764
weighted avg	1.00	1.00	1.00	2764



(a) Matriz de Confusão



(b) ROC AUC

5 Conclusão

Como podemos ver nos resultados, a maioria das configuracões dos modelos produziu bons resultados, com o MLP, SVM, e Árvores de decisão não errando nenhum no dataset de teste. Vale ressaltar que a divisão entre treino e teste foi igual para todos os experimentos. Devido ao tempo muito elevado de treinamento de alguns modelos, como, por exemplo, o SVM, foi decidido não utilizar cross-validation

Vale ressaltar que a diferenca entre os resultados da matriz de confusão e os da curva ROC vem pois a análise ROC é feita primariamente com classificacão binária, enquanto o dataset utilizado é multiclasse. Para calcular a curva ROC foi feita uma análise das classes de forma One Over Rest, ou seja para cada classe i, i é considerado SIM e todas as outras são consideradas como NÃO. No fim temos uma quantidade de de curvas ROC igual a quantidade de classes, cada uma tendo passado por um treinamento separado.

Para trabalhos futuros a utilizacão de uma ferramenta de grid search para pesquisa dos melhores parâmetros de cada modelo de forma automática será bastante útil, pois a geracão de classification reports e curvas ROC para várias combinacões de parâmetros gera muitos arquivos e dificultam a análise dos resultados.

References

[1] What is supervised learning? — ibm. `https://www.ibm.com/cloud/learn/supervised-learning#:~:text=Supervised%20learning%2C%20also%20known%20as,data%20or%20predict0outcomes%20accurately`. (Accessed on 01/10/2022).

[2] S. Abirami and P. Chitra. Chapter fourteen - energy-efficient edge based real-time healthcare support system. In Pethuru Raj and Preetha Evangeline, editors, *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases*, volume 117 of *Advances in Computers*, pages 339–368. Elsevier, 2020.

[3] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), dec 2015.

[4] Joo Chuan Tong and Shoba Ranganathan. 5 - computational t cell vaccine design. In Joo Chuan Tong and Shoba Ranganathan, editors, *Computer-Aided Vaccine Design*, Woodhead Publishing Series in Biomedicine, pages 59–86. Woodhead Publishing, 2013.