Relatório de Classificação Binária e Multiclasse

UC: Inteligência Artificial
Profª: Edyene Cely Amaro Oliveira

Leonardo Araújo Alvarenga - RA 322218544

INTRODUÇÃO

Base de Dados Binária

DATASET SELECIONADO:
https://www.kaggle.com/datasets/eswarchandt/phishing-website-detector

A base de dados utilizada contém uma coleção de URLs com mais de 11.000 sites. Cada amostra tem 30 parâmetros de site e um rótulo de classe que o identifica como um site de phishing ou não (1 ou 0). Essa base inclui várias features que podem ser extraídas dos URLs, como a presença de palavras suspeitas, caracteres especiais, números, redirecionamentos, entre outros. Além disso, outras características, como informações do domínio e análise de conteúdo também podem ser consideradas. A quantidade de características da base selecionada é de 31, são elas: *Index, UsingIP, LongURL, ShortURL, Symbol@, Redirecting//, PrefixSuffix-, SubDomains, HTTPS, DomainRegLen, Favicon, NonStdPort, HTTPSDomainURL, RequestURL, AnchorURL, LinksInScriptTags, ServerFormHandler, InfoEmail, AbnormalURL, WebsiteForwarding, StatusBarCust, DisableRightClick, UsingPopupWindow, IframeRedirection, AgeofDomain, DNSRecording, WebsiteTraffic, PageRank, GoogleIndex, LinksPointingToPage, StatsReportclass*. A saída esperada do modelo treinado é uma classificação binária indicando se o site é legítimo ou phishing, com o objetivo de identificar eficientemente sites fraudulentos e proteger os usuários contra golpes online.

Resumo das categorias:

Index representação numérica que identifica a posição ou valor de um elemento dentro de um conjunto ou sistema de referência. (0 \ 11.053)

UsingIP utilização de IP.

LongURL serviço online que permite expandir e visualizar URLs encurtadas, revelando a URL original por trás delas.

ShortURL serviço que encurta URLs longas em links curtos para facilitar o compartilhamento.

Symbol@ representa uma referência ou ponteiro para a variável ou objeto atual.

Redirecting// direcionar ou redirecionar algo para outro destino ou local.

PrefixSuffix- método linguístico que se refere ao uso de um prefixo (um afixo que é adicionado antes de uma palavra) e um sufixo (um afixo que é adicionado após uma palavra) para modificar ou criar uma nova palavra.

SubDomains domínios secundários que fazem parte de um domínio principal, permitindo a criação de endereços web específicos e independentes sob um único domínio.

HTTPS HyperText Transfer Protocol Secure protocolo de comunicação seguro que utiliza criptografia para proteger os dados transmitidos entre um navegador web e um servidor, garantindo assim a confidencialidade e integridade das informações.

DomainRegLen medida do comprimento do registro de domínio, geralmente em número de caracteres.

Favicon ícone pequeno que representa um website e é exibido na aba do navegador e na barra de favoritos.

NonStdPort porta de comunicação em um sistema de computador que não segue os padrões de portas comuns e é usada para fins especiais ou personalizados.

HTTPSDomainURL URL de domínio que usa uma conexão segura criptografada (HTTPS) para garantir a segurança e privacidade dos dados transmitidos.

RequestURL função ou propriedade que retorna a URL ou endereço de uma solicitação feita em um contexto de programação.

AnchorURL plataforma de encurtamento de URLs que permite aos usuários reduzirem o tamanho de URLs longas em links curtos e personalizáveis.

LinksInScriptTags técnica de desenvolvimento web onde links para arquivos JavaScript são incorporados diretamente em tags de script HTML, permitindo a execução de código JavaScript diretamente na página web.

ServerFormHandler manipulador de formulários usado em programação de servidores para processar e lidar com dados submetidos através de formulários enviados por clientes.

InfoEmail serviço de correio eletrônico que oferece aos usuários a capacidade de enviar e receber mensagens eletrônicas, compartilhar informações e se comunicar de forma rápida e eficiente através da internet.

AbnormalURL URL ou endereço de internet que possui características ou comportamentos considerados incomuns ou fora do padrão.

WebsiteForwarding redirecionamento de um site para outro endereço na web.

StatusBarCust biblioteca de código aberto que permite personalizar a barra de status em aplicativos Android.

DisableRightClick função que desabilita o clique com o botão direito do mouse em elementos de uma página web.

UsingPopupWindow método que permite criar e manipular janelas de pop-up em um aplicativo ou site de forma programática.

IframeRedirection técnica de redirecionamento de página da web usando um elemento iframe incorporado em outra página, geralmente para fins de rastreamento, publicidade ou redirecionamento malicioso.

AgeofDomain plataforma online que oferece informações detalhadas sobre a idade e histórico de domínios de websites.

DNSRecording registro que contém informações sobre a resolução de nomes de domínio em endereços IP, permitindo a navegação na Internet usando URLs amigáveis em vez de endereços numéricos.

WebsiteTraffic volume de visitantes que acessam um site em um determinado período de tempo.

PageRank algoritmo de classificação usado pelo Google para determinar a relevância de uma página da web com base em sua popularidade e nas ligações de outras páginas que apontam para ela.

GoogleIndex algoritmo de classificação usado pelo Google para determinar a relevância de uma página da web com base em sua popularidade e nas ligações de outras páginas que apontam para ela.

LinksPointingToPage métrica que representa o número de links externos que direcionam para uma página específica na web.

StatsReport class classe que encapsula dados estatísticos em um único objeto em linguagens de programação, como JavaScript.

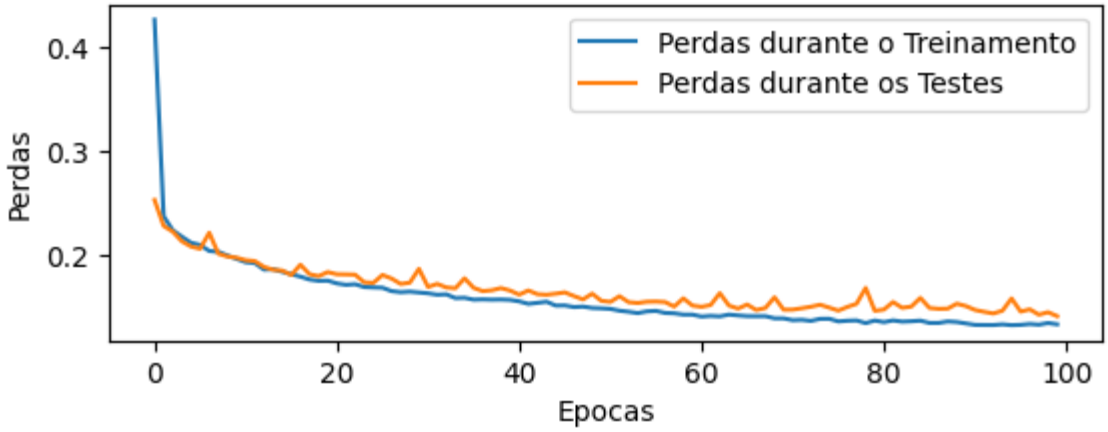
CÓDIGO UTILIZADO:

https://colab.research.google.com/drive/1QZv3kdsFG2aF3_joRjwT5mLuvDPltCmp?usp=sharing

A escolha de utilizar uma rede neural se deu em razão de sua capacidade de aprendizado avançada, tornando mais eficiente a previsão de ocorrência de phishing em um site. Para garantir a qualidade do processo, foram realizadas algumas etapas de pré-processamento nos dados, como a remoção da coluna Index, a substituição dos valores -1 por 0 e a normalização das características, com o objetivo de melhorar o desempenho do modelo durante a fase de treinamento.

O modelo de rede neural utilizado consiste em quatro camadas, sendo três camadas ocultas com ativação “ReLU” e uma camada de saída com ativação sigmóide. O número de neurônios nas camadas ocultas foi definido como 10, 15 e 15, respectivamente. Durante o treinamento, foram utilizadas 50 épocas e um tamanho de lote de 32. A escolha da função de ativação “ReLU” se deve ao fato de que a base de dados já não possuía valores negativos, o que torna essa função a mais indicada para o modelo.

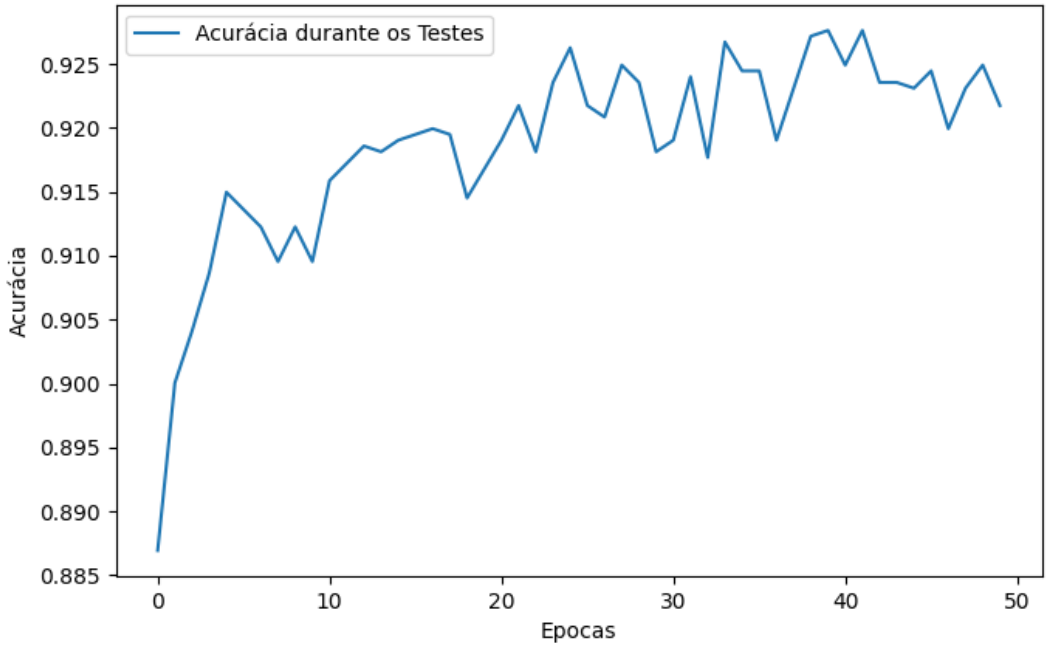
GRÁFICO DE CUSTO



Durante o treinamento do nosso modelo, experimentamos uma perda inicial mais elevada do que na fase de testes. No entanto, à medida que as épocas avançavam, essa perda diminuiu gradualmente até atingir níveis similares aos da fase de testes.

Sendo assim, durante o treinamento do modelo houve uma melhora progressiva na sua performance, medida pela diminuição da perda.

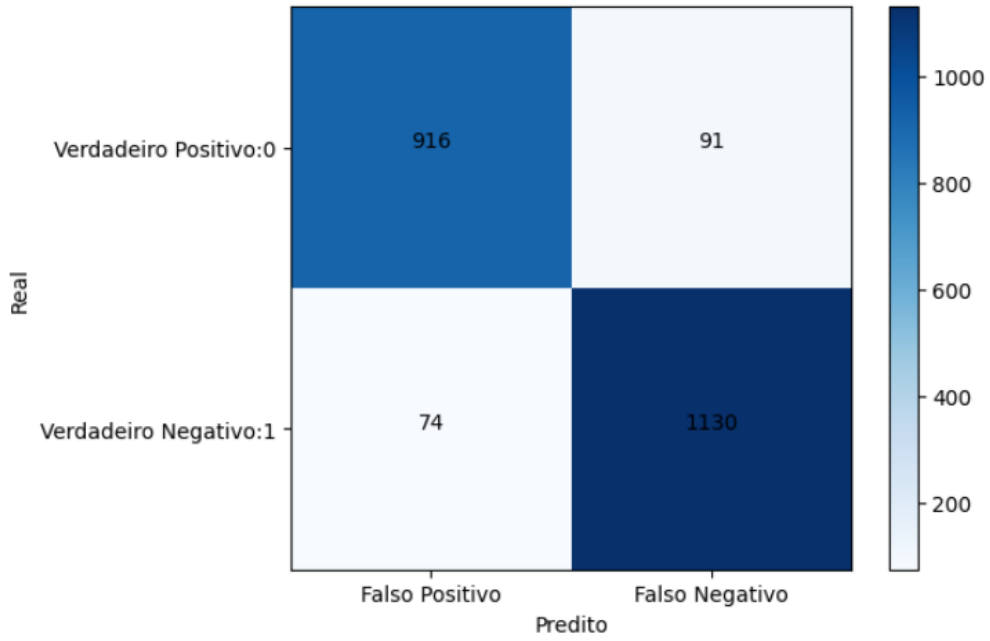
GRÁFICO DE DESEMPENHO



Podemos dizer que a nossa acurácia, teve maior elevação nas primeiras épocas mas após atingir seu máximo acabou variando de acordo com as épocas.

Nesse caso, o nosso modelo apresentou um boa acurácia inicial e foi aumentando gradualmente até chegar no seu pico. Após isso, a acurácia permaneceu oscilando.

MATRIZ DE CONFUSÃO



Após analisarmos a matriz de confusão, podemos inferir que o modelo apresentou um desempenho satisfatório, com um número maior de acertos do que erros.

Essa observação sugere que o modelo foi capaz de identificar corretamente as amostras positivas e negativas, o que é um indicativo de sua eficácia.

RESULTADO DAS MÉTRICAS

Accuracy: **0.914971**

Precision: **0.907752**

Recall: **0.944355**

F1 score: **0.925692**

Cohen's Kappa: **0.826417**

ROC AUC: **0.910900**

Com base nos resultados obtidos, podemos concluir que o modelo apresentou um bom desempenho na tarefa de classificação. O modelo acertou cerca de 91,5% das previsões, o que é uma proporção consideravelmente alta. Além disso, cerca de 90,8% das previsões positivas do modelo estão corretas e cerca de 94,4% dos exemplos positivos foram previstos corretamente, o que indica um bom desempenho do modelo em relação à classe positiva. O F1 score de 0,925692 indica que o modelo tem um bom equilíbrio entre precisão e recall, o que é importante para a classificação correta dos exemplos positivos e negativos. O coeficiente de Cohen's Kappa de 0,826417 indica uma boa

concordância entre os rótulos verdadeiros e as previsões do modelo. Finalmente, a área sob a curva ROC de 0,910900 indica que o modelo tem uma boa capacidade de distinguir entre exemplos positivos e negativos. Portanto, podemos concluir que o modelo apresentou um desempenho geral satisfatório e pode ser considerado útil para a tarefa de classificação.

CONCLUSÃO

Com base no banco de dados Phishing, que consiste em informações sobre a presença ou ausência de Phishing em sites, treinamos nosso modelo para fazer previsões precisas. Para garantir que o modelo produzisse bons resultados, realizamos um tratamento cuidadoso dos dados, fazendo as alterações necessárias. Depois de adequar os dados para o nosso modelo, obtivemos resultados satisfatórios, com um alto desempenho na matriz de confusão, que mostra a quantidade de previsões corretas. Além disso, as métricas também apresentaram resultados positivos, com percentuais acima de 90%, o que nos permitiu concluir que o modelo está pronto para ser implementado na fase de produção.

Base de Dados Multiclasse

DATASET SELECIONADO:

https://www.kaggle.com/datasets/vittoriogiatti/diamondprices

Este conjunto de dados clássicos contém os preços e outros atributos de quase 54.000 diamantes. Cada registro é um diamante aleatório com suas próprias características.

Nessa base é utilizado os 4Cs de um diamante, que se refere ao corte do diamante, cor, clareza e peso de quilate (Carat).

- **Cortar**: De todos os 4Cs, o corte de diamante tem o maior efeito na beleza de um diamante. Ao determinar a qualidade do corte, o graduador de diamantes avalia a habilidade do cortador na formação do diamante. Quanto mais preciso o diamante é cortado, mais cativante o diamante é para os olhos.
- **Cor**: A cor dos diamantes com qualidade de gema ocorre em muitos tons. Na faixa de incolor a amarelo claro ou marrom claro. Diamantes incolores são os mais raros. Outras cores naturais (azul, vermelho, rosa, por exemplo) são conhecidas como "gosta" e sua classificação de cores é diferente da dos diamantes incolores brancos
- **Clareza**: Os diamantes podem ter características internas conhecidas como inclusões ou características externas conhecidas como manchas. Diamantes sem inclusões ou manchas são raros; no entanto, a maioria das características só pode ser vista com ampliação.
- **Carat(Quilate):** O quilate é o peso físico do diamante medido em quilates métricos. Um quilate é igual a 1/5 grama e é subdividido em 100 pontos. O peso do quilate é o grau mais objetivo dos 4Cs.

Resumo das categorias:

preço preço em dólares americanos (\ $ 326-- \ $ 18.823)

quilate peso do diamante (0,2 - 5,01)

cortar qualidade do corte (Justo, Bom, Muito Bom, Premium, Ideal)

cor cor de diamante, de J (pior) a D (melhor)

clareza uma medida de quão claro o diamante é (I1 (pior), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (melhor))

x comprimento em mm (0 - 10,74)

y largura em mm (0 - 58,9)

z profundidade em mm (0 - 31,8)

profundidade porcentagem de profundidade total = z / média (x, y) = 2 * z / (x + y) (43-79)

mesa largura da parte superior do diamante em relação ao ponto mais largo (43-95)

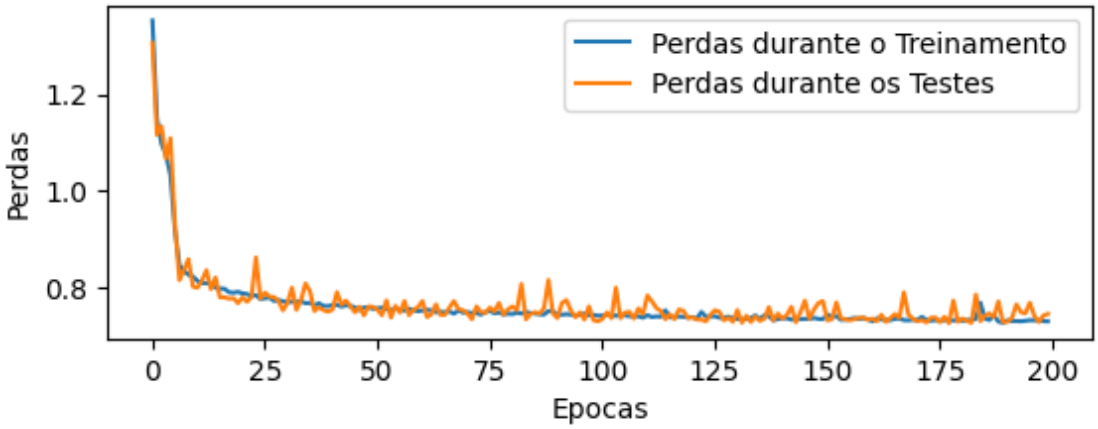
CÓDIGO UTILIZADO:

https://drive.google.com/file/d/1nw7e1u-cWchA7V_VXQCmgPtpvY5Hm8HG/view?usp=sharing

A escolha da rede neural foi baseada em sua capacidade avançada de aprendizado, o que permitiria uma previsão mais precisa da qualidade do corte (Justo, Bom, Muito Bom, Premium, Ideal). Para garantir a qualidade do processo, foram realizadas etapas de pré-processamento nos dados, como a remoção da coluna "Unnamed" e a reorganização das colunas para que a coluna "cut" ficasse como última. Além disso, a coluna "color" teve seus valores alterados de letras (de J até D) para números (de 0 a 6, onde 0 é o pior e 6 é o melhor). O mesmo foi feito com a coluna de clareza, onde os resultados foram enquadrados em uma escala de 1 a 7, sendo 1 o pior e 7 o melhor. Para o modelo gerar suas previsões, as respostas da coluna "cut" foram convertidas em um array binário (por exemplo, 1 0 0 0, 0 0 0 1, 0 0 1 1, 1 1 1 1). Com essas modificações, o modelo foi capaz de gerar previsões precisas.

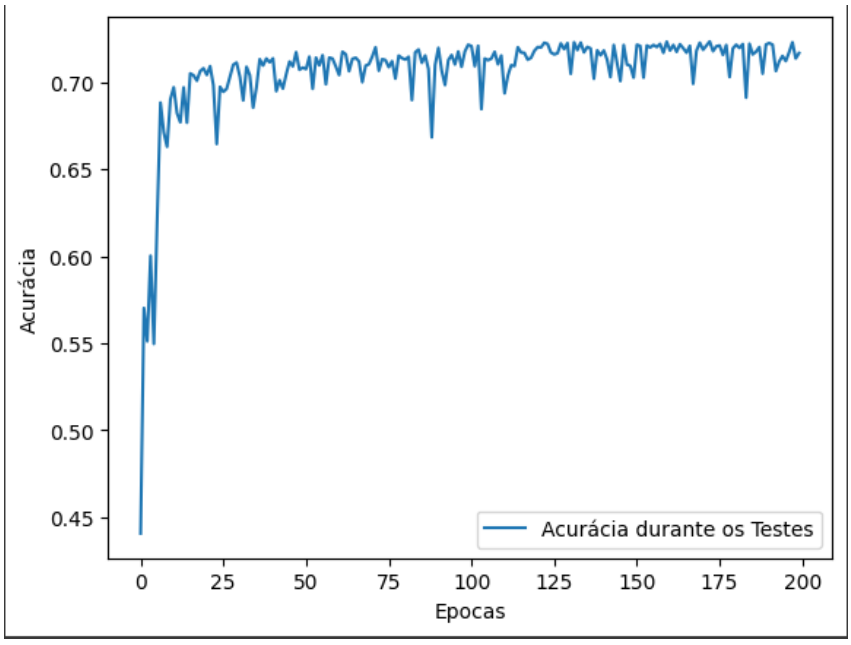
O modelo de rede neural utilizado consiste em quatorze camadas, sendo 13 camadas ocultas com ativação “ReLU” e uma camada de saída com ativação softmax. O número de neurônios nas camadas ocultas foi definido como: 10,15,5,15,5,15,5,15,15,15,15,15,5, respectivamente dando um total de 165 neurônios . Durante o treinamento, foram utilizadas 200 épocas e um tamanho de lote de 32. A escolha da função de ativação “ReLU”se deve ao fato de que a base de dados já não possuía valores negativos, o que torna essa função a mais indicada para o modelo.

GRÁFICO DE CUSTO



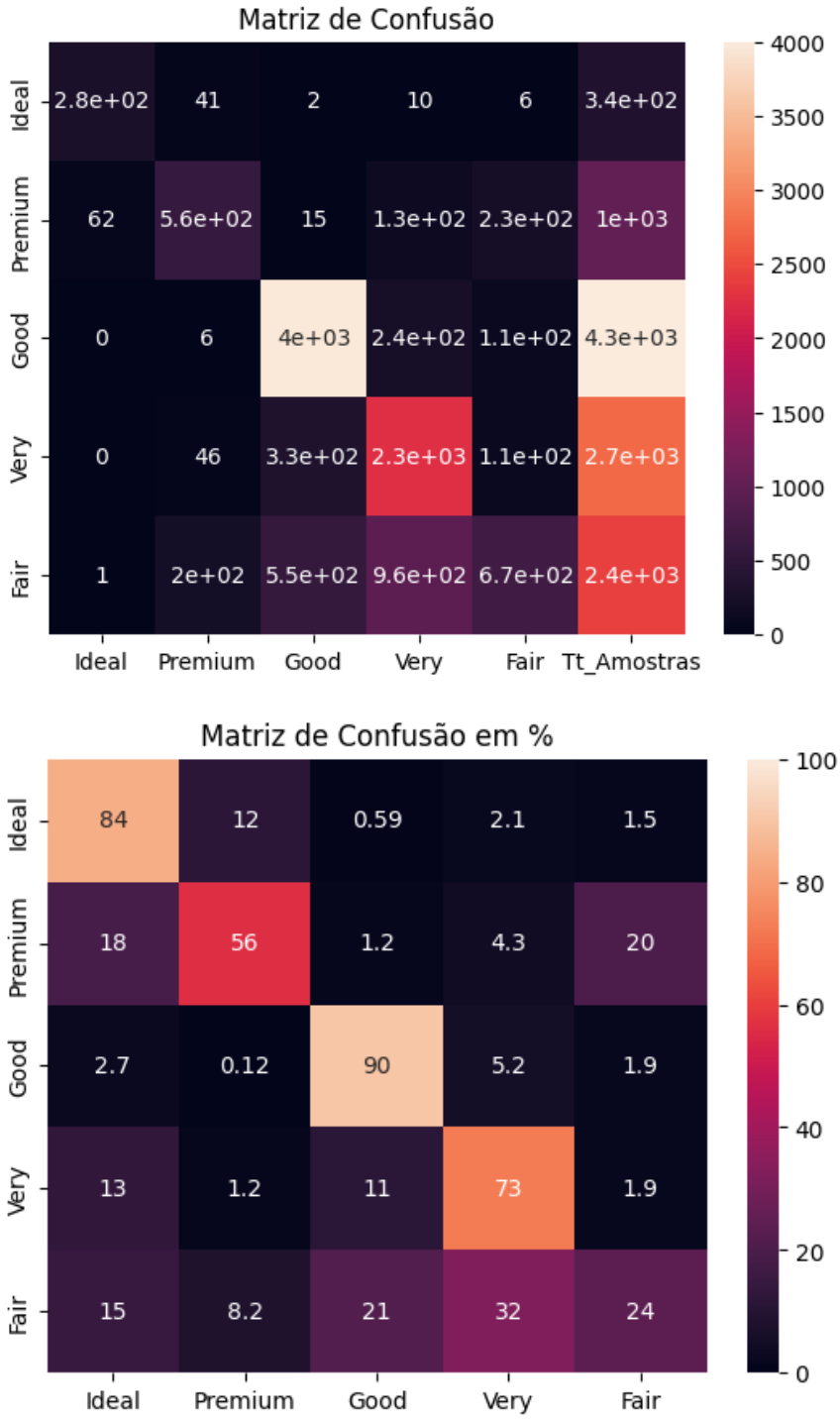
Ao analisar o gráfico, é possível observar que as perdas durante o treinamento e teste apresentaram resultados similares. No entanto, com algumas variações a partir da época 75 no conjunto de testes. Ambos terminaram com uma perda de 70%. Com base nesses resultados, podemos concluir que o nosso modelo não está apresentando uma previsão precisa dos valores reais.

GRÁFICO DE DESEMPENHO



Ao analisar o gráfico, é possível observar que a acurácia do modelo não apresenta mais melhorias significativas após a época 25, o que indica a existência de problemas em suas previsões. Embora a acurácia esteja acima de 70%, é perceptível que alguma anomalia pode estar prejudicando a obtenção de resultados mais precisos pelo modelo.

MATRIZ DE CONFUSÃO



Ao analisarmos a matriz de confusão, podemos observar que o modelo apresenta um desempenho inferior na identificação dos valores "Fair" e "Premium". Em particular, a categoria "Fair" é onde o modelo mais erra, confundindo-a com a sua maior parte na categoria "Very Good". Em contrapartida, os resultados para as outras categorias apresentam uma taxa de acerto consideravelmente alta. Esses resultados podem indicar a presença de algum problema na base de dados utilizada ou nas camadas de neurônios do modelo, que está prejudicando seu desempenho.

RESULTADO DAS MÉTRICAS

```
|------------------|
|Accuracy: 0.68    |
|------------------|
|Precision: 0.74   |
|------------------|
|Recall: 0.68      |
|------------------|
|F1-score: 0.71    |
|------------------|

Relatório de classificação

              precision    recall  f1-score   support

       Ideal       0.82      0.82      0.82       339
     Premium       0.67      0.56      0.61      1000
        Good       0.83      0.90      0.86      4320
        Very       0.66      0.73      0.69      2736
        Fair       0.63      0.24      0.35      2393

   micro avg       0.74      0.68      0.71     10788
   macro avg       0.72      0.65      0.67     10788
weighted avg       0.72      0.68      0.68     10788
 samples avg       0.68      0.68      0.68     10788
```

Com base nos resultados apresentados, podemos concluir que o modelo de classificação possui uma **acurácia de 0.68**, o que significa que ele acerta aproximadamente 68% das previsões. Além disso, o modelo apresenta um desempenho desigual para as diferentes categorias de diamantes, com uma **precisão variando entre 0.63 e 0.83** e um **recall entre 0.24 e 0.90**. O **F1-score, que é uma medida que combina precisão e recall, tem um valor médio de 0.71.**

Olhando para o relatório de classificação, podemos ver que o modelo apresenta um desempenho muito bom para as categorias "Ideal","Very Good ","Good", com uma precisão e recall de 0.82, 0.83, 0.90 e 0.66, 0.73 respectivamente. Já a categoria “Premium” e "Fair" apresenta um desempenho bem inferior, com uma precisão e recall de 0.67, 0.56 e 0.63, 0.24, respectivamente.

Em geral, podemos dizer que o modelo tem um desempenho aceitável, mas ainda pode ser melhorado, especialmente para a categoria “Premium” e "Fair". É possível que seja necessário ajustar os parâmetros do modelo, utilizar mais dados para treinamento ou considerar a utilização de um modelo diferente para obter resultados mais precisos.

CONCLUSÃO

Ao utilizar a base de dados diamonds, que contém informações detalhadas sobre diamantes, foi possível realizar previsões quanto ao seu tipo de corte, classificados como (Justo, Bom, Muito Bom, Premium, Ideal). Para isso, foi necessário realizar algumas modificações na base, como excluir a coluna (Unnamed: 0), que servia como índice, e reorganizar as colunas para facilitar a separação das características das respostas. Além disso, os valores das colunas (color e clarity), que eram do tipo String, precisaram ser substituídos por valores numéricos de 0 a 6 e de 0 a 7, respectivamente. Também foi necessário normalizar os valores de x e transformar as respostas da coluna cut em arrays binários.

No entanto, após realizarmos os testes e treinamentos, percebemos que nosso modelo não obteve resultados muito bons. Concluímos que talvez seja necessário realizar ajustes na base de dados ou na forma de pré-processar os dados, ou até mesmo utilizar um modelo diferente do que foi utilizado, para obter resultados melhores e assim poder introduzi-lo na fase de produção.

CONCLUSÕES DO RELATÓRIO

O relatório solicitado utilizando as métricas de avaliação em Machine Learning foi devidamente concluído, alcançando os objetivos propostos. Foram utilizados dois bancos de dados (binário e multiclasse) para analisar e comparar o desempenho de diferentes algoritmos na tarefa de classificação de dados.

Os objetivos do relatório foram cumpridos com sucesso, permitindo uma compreensão detalhada do desempenho dos modelos utilizados em ambos os bancos de dados. Foram aplicadas métricas de avaliação padrão, incluindo acurácia, precisão, F1-score, Recall, matriz de confusão, AUC ROC e Kappa, para avaliar a performance dos modelos.

Os resultados obtidos indicam um desempenho satisfatório dos modelos na tarefa de classificação, com boa capacidade de generalização para novos dados. Ao comparar os resultados obtidos nos dois bancos de dados, observou-se que o modelo apresentou um desempenho ligeiramente melhor no banco de dados binário em relação ao multiclasse. Isso pode ser explicado pela maior complexidade e variabilidade dos dados multiclasse, que tornam a tarefa de classificação mais desafiadora.

Em conclusão, o relatório apresentou uma análise abrangente do que foi proposto pela UC. Os objetivos indicados foram alcançados com sucesso, fornecendo insights importantes sobre o desempenho dos modelos e áreas de melhoria.