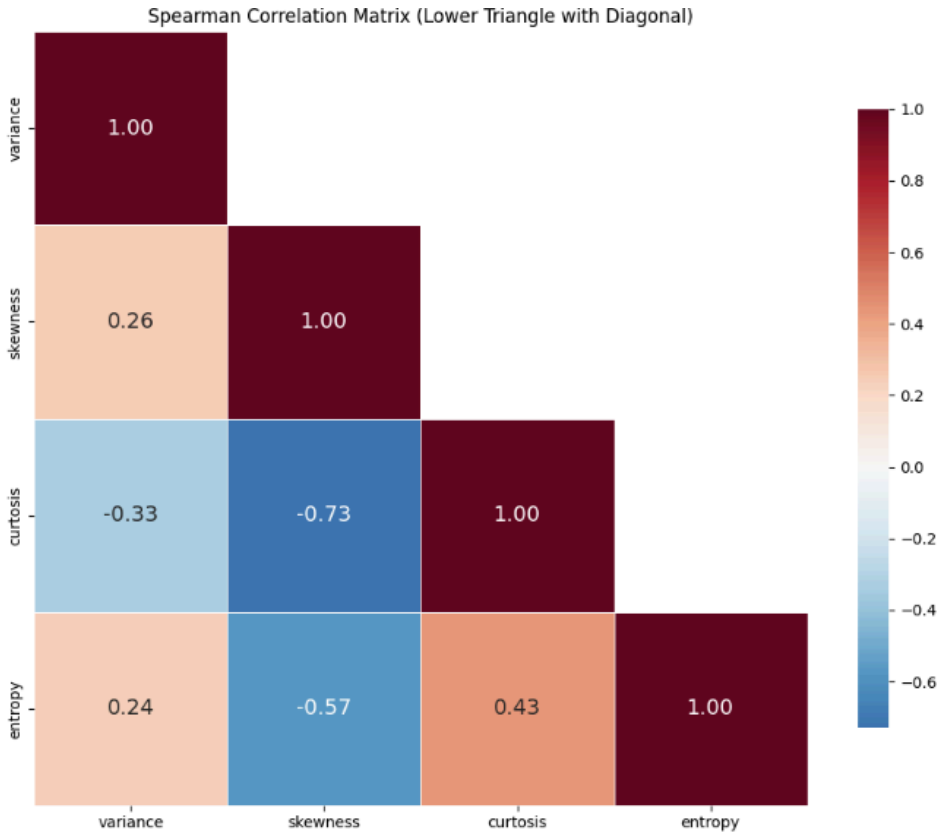
Análise de Autenticidade de Cédulas Bancárias com Redes Neurais

Autor: João Gabriel Argento e Lucas Martins
Disciplina: Redes Neurais
Instituição: Universidade do Estado do Rio de Janeiro (UERJ)

Este relatório apresenta os resultados obtidos a partir da aplicação de um mapa de Kohonen (*Self-Organizing Map - SOM*), uma rede neural artificial de aprendizado não supervisionado, para a tarefa de análise e agrupamento de dados visando identificar padrões em cédulas genuínas e falsificadas. O estudo foi conduzido utilizando o conjunto de dados Banknote Authentication, disponibilizado no repositório UCI Machine Learning, composto por medidas extraídas de imagens de notas bancárias. Cada instância no conjunto representa uma cédula, caracterizada por atributos derivados de *Transformadas Wavelet*, como variância, assimetria, curtose e entropia da imagem.
Além da análise de agrupamento com o mapa de Kohonen, também são abordados aspectos da Análise Exploratória de Dados (AED) realizada previamente, com o intuito de contextualizar as variáveis utilizadas, suas distribuições e correlações. O objetivo principal é investigar a capacidade do *SOM* em distinguir entre cédulas autênticas e falsas, identificando possíveis *clusters* naturais nos dados e avaliando a eficácia da técnica para detecção de fraudes monetárias.

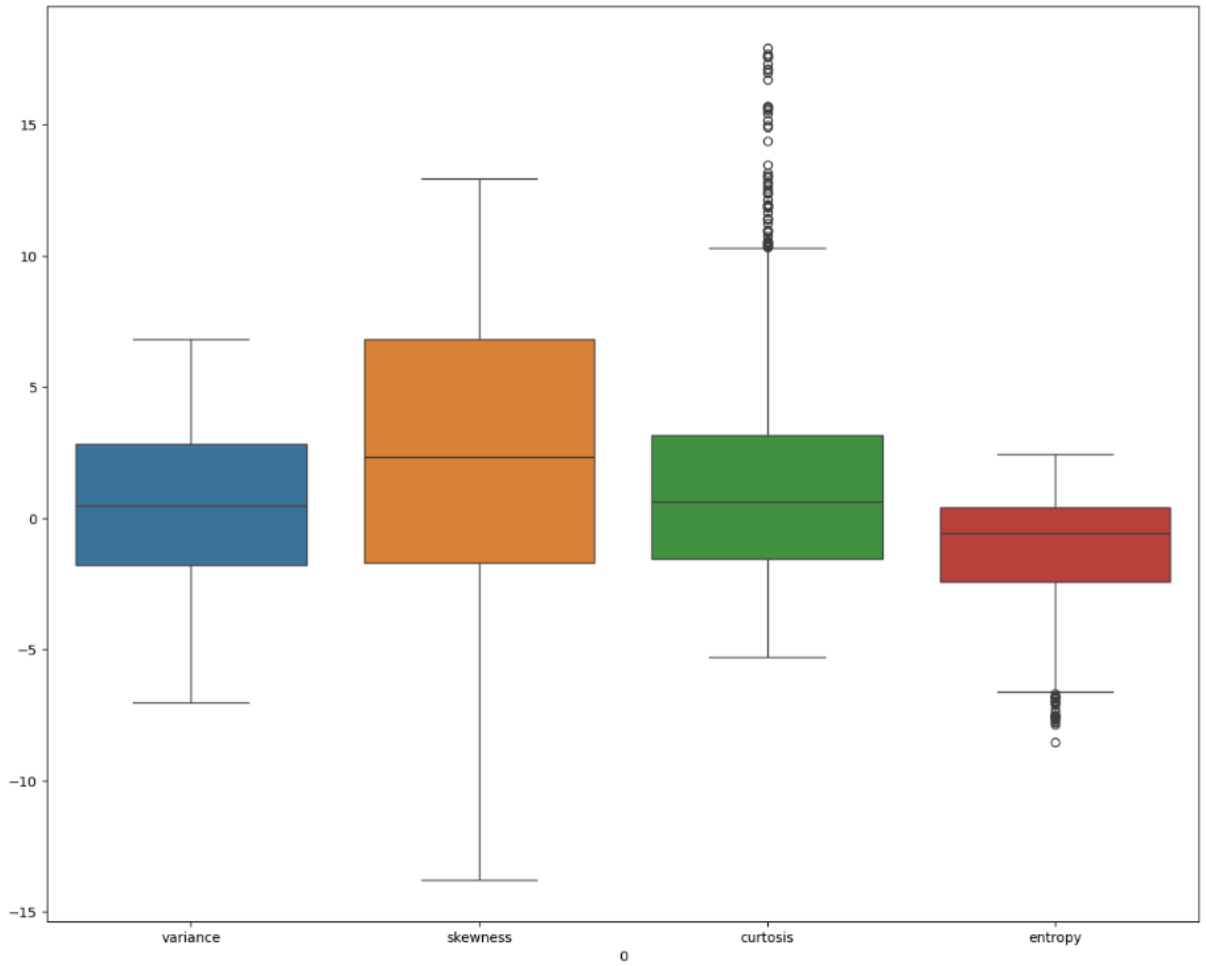
Análise Exploratória de Dados

Para iniciar a análise das características das variáveis numéricas da base de dados, foi utilizada a correlação de Spearman, que permite identificar relações não lineares entre elas. Essa escolha se justifica pelo fato de redes neurais serem especialmente eficazes na modelagem de padrões complexos e não lineares presentes nos dados.



A matriz de correlação de Spearman revelou algumas relações importantes entre as variáveis numéricas analisadas. A correlação mais forte foi entre a assimetria (**skewness**) e a curtose (**curtosis**), com um valor de -0,73, indicando que, quando a distribuição dos dados fica mais inclinada para um lado, ela tende a se tornar menos "pontuda" ou concentrada no centro. Isso mostra que essas duas medidas, que descrevem o formato dos dados, costumam variar em sentidos opostos.
Entre as correlações moderadas, observamos que distribuições com maior variação (variância) tendem a ser menos "pontudas" (curtose), com uma correlação de -0,33. Além disso, quando os dados ficam mais inclinados para um lado (maior assimetria), costumam apresentar menos imprevisibilidade ou desordem (entropia), refletido em uma correlação de -0,57. Já a relação entre curtose e entropia foi positiva (0,43), sugerindo que distribuições mais "pontudas" tendem a ser um pouco mais imprevisíveis.
As outras relações identificadas foram mais fracas. Por exemplo, a variância teve uma correlação fraca e positiva com a assimetria (0,26) e com a entropia (0,24), mostrando que ela varia de forma quase independente das outras características, trazendo informações diferentes para a análise.
No geral, os resultados destacam que a relação mais marcante é entre assimetria e curtose, que tendem a se comportar de forma oposta. A entropia aparece ligada, de modo moderado, às medidas que descrevem o formato dos dados. E, por não existirem correlações perfeitas, cada variável contribui com algum detalhe único para a compreensão completa da base de dados.

A seguir, apresenta-se um gráfico de *boxplot* das variáveis explicativas da base de dados. Essa visualização permite analisar a distribuição, dispersão e possíveis *outliers* de cada variável, oferecendo insights sobre sua variabilidade e simetria.



Como podemos ver nos *boxplots* acima, as variáveis curtose (**curtosis**) e entropia (**entropy**) apresentam alguns valores que ficam muito distantes da maioria dos dados, são os chamados *outliers*. Esses valores extremos podem indicar casos atípicos ou situações especiais que merecem atenção na análise. Já as variáveis "variance" e "skewness" não mostram esses valores tão distantes, sugerindo um comportamento mais estável nesses aspectos.
Além disso, observamos que a variável "skewness", que mede o grau de assimetria da distribuição, apresenta a maior variação entre os dados. Isso significa que os valores de "skewness" mudam bastante de uma observação para outra, e que essa distribuição tende a ser menos equilibrada em torno do centro (mais "torta").
Por fim, a variável "variance" chama atenção por ter a distribuição mais simétrica entre todas, ou seja, seus valores estão mais equilibrados entre as duas metades do conjunto de dados, sem puxar muito para um lado ou para o outro.

A seguir, são apresentados os resultados do **teste de hipóteses não paramétrico *Mann-Whitney U***, aplicado para comparar a distribuição de cada variável explicativa entre as classes 0 (cédulas genuínas) e 1 (cédulas falsificadas) da base Banknote Authentication. Esse teste é especialmente útil quando não se pode assumir que os dados seguem uma distribuição normal, pois testes não paramétricos não dependem de pressupostos sobre a forma da distribuição dos dados, o que os torna mais robustos em cenários como este.
O teste foi escolhido justamente por essa característica: garantir uma análise estatística confiável mesmo quando as variáveis podem apresentar assimetrias, curtoses elevadas ou outras violações da normalidade.

A interpretação dos resultados baseia-se no **p-valor**, que representa a probabilidade de observar uma diferença entre as classes tão extrema quanto a verificada nos dados, caso a hipótese nula (de que as distribuições das classes são iguais) seja verdadeira. Se o p-valor for inferior ao nível de significância de 5% (0,05), rejeita-se a hipótese nula, indicando que há evidências estatísticas de que a variável apresenta diferenças relevantes entre cédulas genuínas e falsificadas. Dessa forma, o teste ajuda a identificar quais variáveis têm maior potencial discriminatório para a tarefa de autenticação.

```
Statistical tests for group differences:
================================================

variance:
  Mann-Whitney U p-value: 0.000000
  Significant difference: Yes

skewness:
  Mann-Whitney U p-value: 0.000000
  Significant difference: Yes

curtosis:
  Mann-Whitney U p-value: 0.022561
  Significant difference: Yes

entropy:
  Mann-Whitney U p-value: 0.225288
  Significant difference: No
```

Os resultados mostraram que as variáveis “variance” e “skewness” possuem p-valores extremamente baixos (próximos de zero), o que indica diferenças altamente significativas entre as classes. Isso significa que tanto a variabilidade dos valores quanto a assimetria da distribuição são características que distinguem fortemente cédulas verdadeiras de cédulas falsificadas, reforçando seu valor como atributos para classificação.
A variável “curtosis” também apresentou um p-valor abaixo do nível de significância convencional ($p \approx 0{,}022$), evidenciando que há uma diferença estatisticamente significativa entre as classes. Embora essa diferença não seja tão expressiva quanto nas variáveis anteriores, ela ainda contribui para diferenciar cédulas genuínas de falsificadas.
Por outro lado, a variável “entropy” apresentou um p-valor elevado ($\approx 0{,}225$), o que indica que não foi encontrada diferença estatisticamente significativa entre as classes para essa característica. Isso sugere que, no contexto dessa base de dados, a entropia não tem um papel relevante na separação entre cédulas verdadeiras e falsificadas, oferecendo menor poder discriminatório para a tarefa de autenticação.

Mapa de Kohonen (*Self-Organizing Map - SOM*)

Com base nas análises exploratórias apresentadas anteriormente — incluindo os *boxplots*, que apontaram possíveis *outliers* nas colunas “entropy” e “curtosis”, e o teste de hipóteses *Mann-Whitney U*, que identificou p-valor de aproximadamente 0,022 para “curtosis” (superior aos valores observados para “variance” e “skewness”, mas ainda indicando poder discriminatório) e um p-valor elevado ($\approx 0{,}225$) para “entropy” — foi possível orientar a **seleção de variáveis** para o modelo.

A etapa seguinte consiste na implementação da rede neural mapa de Kohonen, testando diferentes versões da base de dados para avaliar o impacto dessas variáveis no desempenho do modelo.
Serão consideradas quatro versões:

1) BASE 0: base completa sem remoções
2) BASE 1: base sem a coluna *entropy*
3) BASE 2: base sem a coluna *curtosis*
4) BASE 3: base sem ambas (*curtosis* e *entropy*)

Essa abordagem permitirá compreender de forma comparativa como a exclusão dessas variáveis influencia a capacidade do mapa de Kohonen em distinguir entre cédulas genuínas e falsificadas.

Metodologia

A preparação dos dados começou com a divisão do conjunto original em três partes, de forma a manter a mesma proporção entre cédulas genuínas e falsificadas. Foram reservados 5% dos dados exclusivamente para testes, enquanto o restante foi novamente separado, destinando 15% para validação do modelo. Em seguida, os dados de treino e validação passaram por um processo de normalização de escala, usando uma técnica chamada *StandardScaler*, que ajusta os valores para que fiquem em uma faixa similar e evita que variáveis com números muito altos ou baixos influenciem mais no resultado.
Na etapa seguinte, foi configurada uma rede neural do tipo *Mapa Auto-Organizável (SOM)* de Kohonen. Essa rede foi montada como uma grade de 10x10 neurônios, cada um começando com valores iniciais escolhidos aleatoriamente. O treinamento foi definido com parâmetros básicos configurados com valores padrão (*default*): uma taxa de aprendizado inicial de 0,4, duração de 20.000 ciclos (iterações) e um raio que determina quais neurônios vizinhos são influenciados durante o aprendizado. Também foi usada a Análise de Componentes Principais (PCA) para reduzir a quantidade de informações e permitir visualizar os dados em apenas duas dimensões.
Durante o treinamento, a rede foi ajustada em etapas: a cada ciclo, um exemplo dos dados de treino era escolhido, e o neurônio mais parecido com esse exemplo — chamado de BMU (Melhor Unidade de Correspondência) — era identificado usando uma medida de distância. Depois, esse neurônio e seus vizinhos tinham seus valores ajustados para ficarem ainda mais próximos do exemplo analisado. Para calcular essa influência, foi usada uma função chamada Mexican Hat, e, conforme o treinamento avançava, a taxa de aprendizado ia diminuindo gradualmente. Em determinados intervalos, foram gerados gráficos mostrando como os neurônios estavam sendo ajustados, permitindo acompanhar a evolução do treinamento.
Por fim, diferentes versões do *SOM* foram comparadas, cada uma utilizando conjuntos de características distintos. Essa comparação avaliou três aspectos principais: o Erro de Quantização, que indica a precisão do mapa ao representar os dados; o Erro Topográfico, que verifica se a organização dos dados originais foi preservada no mapa; e a Acurácia de Classificação, que reflete a capacidade do modelo em distinguir corretamente cédulas genuínas das falsificadas. Vale destacar que não foi realizada a busca de hiperparâmetros nem a remoção dos possíveis *outliers* identificados anteriormente em *boxplots* na etapa de análise exploratória, pois a própria seleção de variáveis feita nessa mesma etapa, baseada em *boxplots* e no teste de hipóteses *Mann-Whitney U*, já se mostrou suficiente para alcançar resultados muito satisfatórios de acurácia no conjunto de validação, conforme será apresentado na seção de resultados do modelo. Essa análise conjunta permitiu identificar qual versão apresentou o melhor desempenho geral.

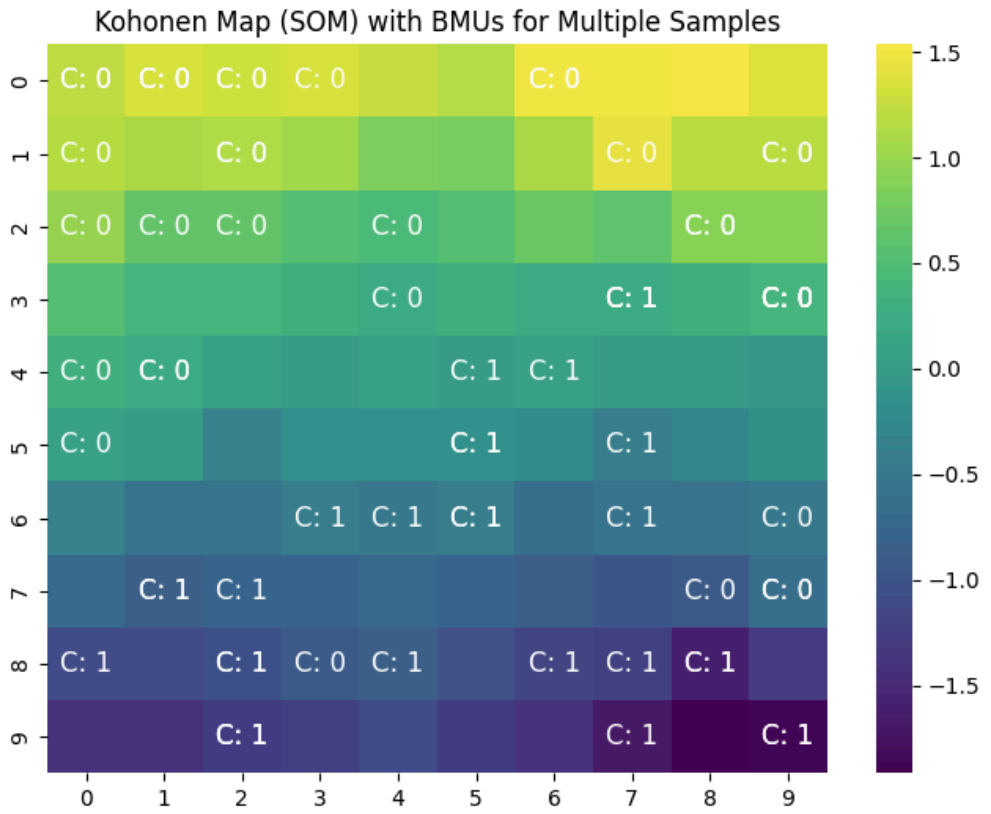
Resultados

Após concluir o processo de treinamento, a tabela abaixo contém os resultados das métricas do **conjunto de validação** (erro de quantização, erro topográfico e acurácia) das 4 versões da base criadas.

Rank	Versão da Base	Erro de Quantização	Erro Topográfico	Acurácia
1	BASE 1	0.2743	0.1707	1.0000
2	BASE 0	0.4345	0.1355	0.9847
3	BASE 2	0.3145	0.1220	0.9694
4	BASE 3	0.1522	0.1192	0.9337

Conforme a tabela acima, é possível observarmos o resultado expressivo da acurácia de **100%** no conjunto de validação da BASE 1 (sem a coluna “entropy”), garantindo o 1º lugar neste critério e demonstrando ser a configuração mais eficaz para distinguir corretamente as cédulas genuínas das falsificadas.
No Erro de Quantização, onde valores menores indicam melhor representação dos dados, o melhor desempenho foi da BASE 3 (sem “curtosis” e “entropy”), com o menor valor registrado (**0,1522**).
Essa mesma base também obteve o menor Erro Topográfico (**0,1192**), sugerindo maior preservação da topologia e coerência da estrutura dos dados no mapa.
De modo geral, a exclusão combinada das variáveis “curtosis” e “entropy” resultou em melhorias na qualidade do mapeamento e na preservação topológica. Contudo, o modelo de rede neural construído a partir da BASE 1 (mantendo todas as variáveis, exceto “entropy”) foi selecionado como o melhor modelo, por ter atingido acurácia máxima de 100% no conjunto de validação.

Em seguida, o modelo selecionado foi aplicado ao **conjunto de teste**, cujos resultados serão apresentados abaixo, juntamente com seu Mapa de Kohonen gerado.



Versão da Base	Erro de Quantização	Erro Topográfico	Acurácia
BASE 1	0.2906	0.2319	1.0000

Como é possível observar acima, o modelo novamente alcançou acurácia de 100% no conjunto de teste, mesma acurácia registrada no conjunto de validação. Houve apenas pequenas variações nos demais indicadores: o erro de quantização passou de 0.2743 na validação para 0.2906 no teste (aumento de 0.0163) e o erro topográfico aumentou de 0.1707 para 0.2319 (diferença de 0.0612). Esses resultados reforçam a consistência do modelo mesmo ao ser aplicado em dados que não participaram do treinamento.

Conclusão

Em síntese, os resultados demonstram que a aplicação do mapa de Kohonen foi eficaz para identificar padrões entre cédulas genuínas e falsificadas, especialmente após uma seleção criteriosa de variáveis orientada por análises de *boxplot* e pelo teste de hipóteses *Mann-Whitney U*. A exclusão apenas da coluna “entropy” (BASE 1) permitiu ao modelo alcançar acurácia de 100%, tanto no conjunto de validação quanto no conjunto de teste, evidenciando sua capacidade de generalização.

Vale ressaltar que não foi realizada a busca de hiperparâmetros nem a remoção dos possíveis *outliers* identificados durante a análise exploratória, pois a própria seleção de variáveis já se mostrou suficiente para obter resultados muito satisfatórios.

Esses achados reforçam a importância da análise exploratória no processo de modelagem e confirmam o potencial do *SOM* como ferramenta complementar na detecção de fraudes em cédulas bancárias.