Trabalho Final - Machine Learning

WATER POTABILITY

Discentes: Ângelo Resplandes,
Áquila Morais, Luís Otávio

Resumo do Projeto

- **Objetivo:** Implementar e comparar 3 algoritmos para **detecção de anomalias** em qualidade de água.
- **Dataset:** Water Potability (Kaggle) - 3.276 amostras, 9 variáveis.
- **Algoritmos:** Isolation Forest, LOF, Autoencoder (com sistema de consenso por votação).
- **Pré-processamento:** KNN Imputer, StandardScaler, PCA para visualização.

Domínio de Aplicação

- **Saúde Pública e Saneamento** - Monitoramento e Controle de Qualidade da Água para Consumo Humano.
- **Contexto:** Sistemas de tratamento de água urbanos e rurais.
- **Aplicação:** Detecção de anomalias em parâmetros físico-químicos.
- **Impacto:** Garantia de água potável segura para consumo humano.

Base de Dados: Water Quality

- **Fonte:** Kaggle - Water Potability Dataset.
- **Registros:** 3.276 amostras de água.
- **Características:** pH, Dureza, Sólidos Dissolvidos, Cloretos, Sulfatos, Condutividade, Matéria Orgânica, Trihalometanos, Turbidez, Potabilidade.
- **Desafio:** Dados faltantes e desbalanceamento de classes.

Base de Dados: Water Quality

ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
NaN	204.890.455	20.791.318.98	7.300.212	368.516.441	564.308.654	10.379.783	86.990.970	2.963.135	0
3.716.080	129.422.921	18.630.057.85	6.635.246	NaN	592.885.359	15.180.013	56.329.076	4.500.656	0
8.099.124	224.236.259	19.909.541.73	9.275.884	NaN	418.606.213	16.868.637	66.420.093	3.055.934	0
8.316.766	214.373.394	22.018.417.44	8.059.332	356.886.136	363.266.516	18.436.524	100.341.674	4.628.771	0
9.092.223	181.101.509	17.978.986.33	6.546.600	310.135.738	398.410.813	11.558.279	31.997.993	4.075.075	0
5.584.087	188.313.324	28.748.687.73	7.544.869	326.678.363	280.467.916	8.399.735	54.917.862	2.559.708	0
10.223.862	248.071.735	28.749.716.54	7.513.408	393.663.396	283.651.634	13.789.695	84.603.556	2.672.989	0
8.635.849	203.361.523	13.672.091.76	4.563.009	303.309.771	474.607.645	12.363.817	62.798.309	4.401.425	0
NaN	118.988.579	14.285.583.85	7.804.174	268.646.941	389.375.566	12.706.049	53.928.846	3.595.017	0
11.180.284	227.231.469	25.484.508.49	9.077.200	404.041.635	563.885.481	17.927.806	71.976.601	4.370.562	0

Classificações de anomalias

pH

- pH ácido (< 6.5): Corrosão de tubulações, dissolução de metais
- pH alcalino (> 9.0): Descarga industrial, contaminação química

Turbidez

- Alta turbidez (> 5 NTU): Chuvas intensas (runoff), descarga de esgoto
- Turbidez + pH normal: Partículas suspensas sem alteração química

Condutividade

- Alta condutividade (> 1500 µS/cm): intrusão salina, mineralização excessiva
- Baixa condutividade (< 30 µS/cm): Água destilada/purificada em excesso

Trihalometanos (THMs)

- THMs elevados (> 100 µg/L): Excesso de cloro + matéria orgânica alta,
- Correlação: pH ↑, temperatura ↑, carbono orgânico ↑ → THMs ↑

Cloraminas

- Excesso (> 3.0 mg/L): Sobredosagem de desinfetante
- Deficiência (< 0.3 mg/L): Risco microbiológico, descarga, diluição

Sulfatos

- Sulfatos altos (> 400 mg/L): Drenagem ácida de minas, efluentes industriais (papel, química), decomposição de matéria orgânica

Carbono Orgânico (TOC)

- TOC elevado (> 10 mg/L): Descarga de esgoto, decomposição de vegetação
- Correlação: TOC ↑ → THMs ↑ (quando clorada)

Dureza

- Dureza extrema (> 500 mg/L): Dissolução de calcário/dolomita

Sólidos Dissolvidos (TDS)

- TDS alto (> 1000 mg/L): Mineralização natural excessiva, intrusão salina

Pipeline de Pré-processamento

1. **Tratamento de Valores Ausentes:** KNN Imputer (k=5)
2. **Normalização:** StandardScaler (μ=0, σ=1)
3. **Redução de Dimensionalidade:** PCA para visualização 2D

Objetivo:
Preparar um conjunto de dados (possivelmente incompleto e com escalas diferentes) para ser visualizado graficamente em um plano 2D.

Algoritmos Implementados

1. **Isolation Forest:** Baseado em isolamento de observações
2. **Local Outlier Factor (LOF):** Baseado em densidade local
3. **Autoencoder:** Rede neural para reconstrução

Objetivo:
Detectar amostras de água com características anômalas que possam indicar problemas de qualidade ou contaminação.

Sistema de Consenso

O projeto utiliza um **sistema de votação por maioria:**

```
Consenso = (Isolation Forest + LOF + Autoencoder) ≥ 2
```

Apenas amostras identificadas por **≥2 dos 3 métodos** são consideradas anomalias finais.

Isolation Forest para Detecção de Anomalias

Baseia-se na premissa de que anomalias são poucas e diferentes, portanto mais fáceis de isolar.

Aspecto	Descrição
Princípio	Isolamento rápido de anomalias via particionamento recursivo
Complexidade	O(n log n) - linear
Métrica	Path Length (comprimento do caminho)
Vantagem	Eficiente para grandes volumes de dados

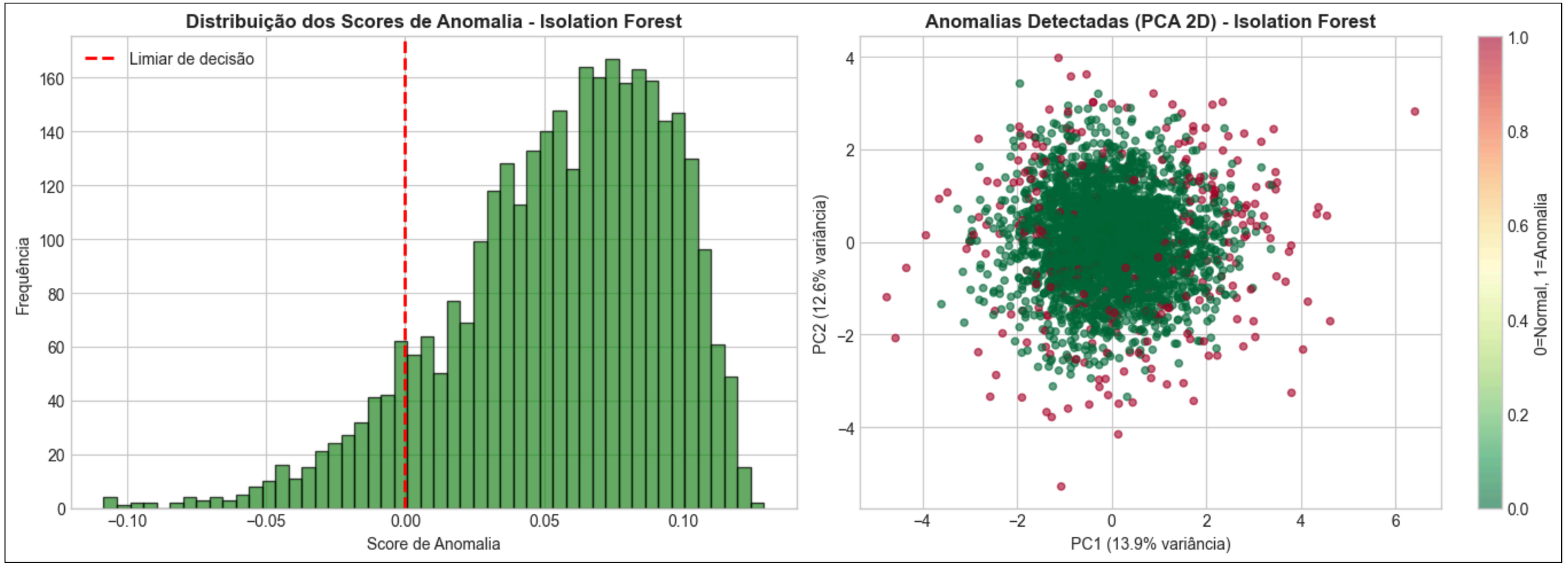
Isolation Forest para Detecção de Anomalias

Hiperparâmetros:
n_estimators = 100 (árvores)
contamination = 0.10 (10% anomalias)

ISOLATION FOREST
Parâmetros do modelo:
- Número de estimadores: 100
- Taxa de contaminação: 10.0%

Resultados:
- Total de amostras: 3276
- Anomalias detectadas: 328 (10.01%)
- Amostras normais: 2948 (89.99%)

Isolation Forest para Detecção de Anomalias



Local Outlier Factor (LOF)

Princípio de Funcionamento

k-Vizinhos Mais Próximos: Para cada ponto, identifica os k vizinhos mais próximos.

Distância de Alcançabilidade: Calcula a distância de alcançabilidade entre pontos.

Densidade Local de Alcançabilidade (LRD): Mede a densidade local de cada ponto.

Fator LOF: Compara a densidade local de um ponto com a de seus vizinhos.

Interpretação do Score LOF:

LOF ≈ 1: O ponto tem densidade similar aos vizinhos (normal).

LOF > 1: O ponto tem densidade menor que os vizinhos (possível anomalia).

LOF >> 1: Forte indicação de anomalia.

LOF para Detecção de Anomalias

Hiperparâmetros Utilizados

n_neighbors = 20: Número de vizinhos para calcular densidade local
contamination = 0.10: Proporção esperada de anomalias (10%)

LOCAL OUTLIER FACTOR (LOF)

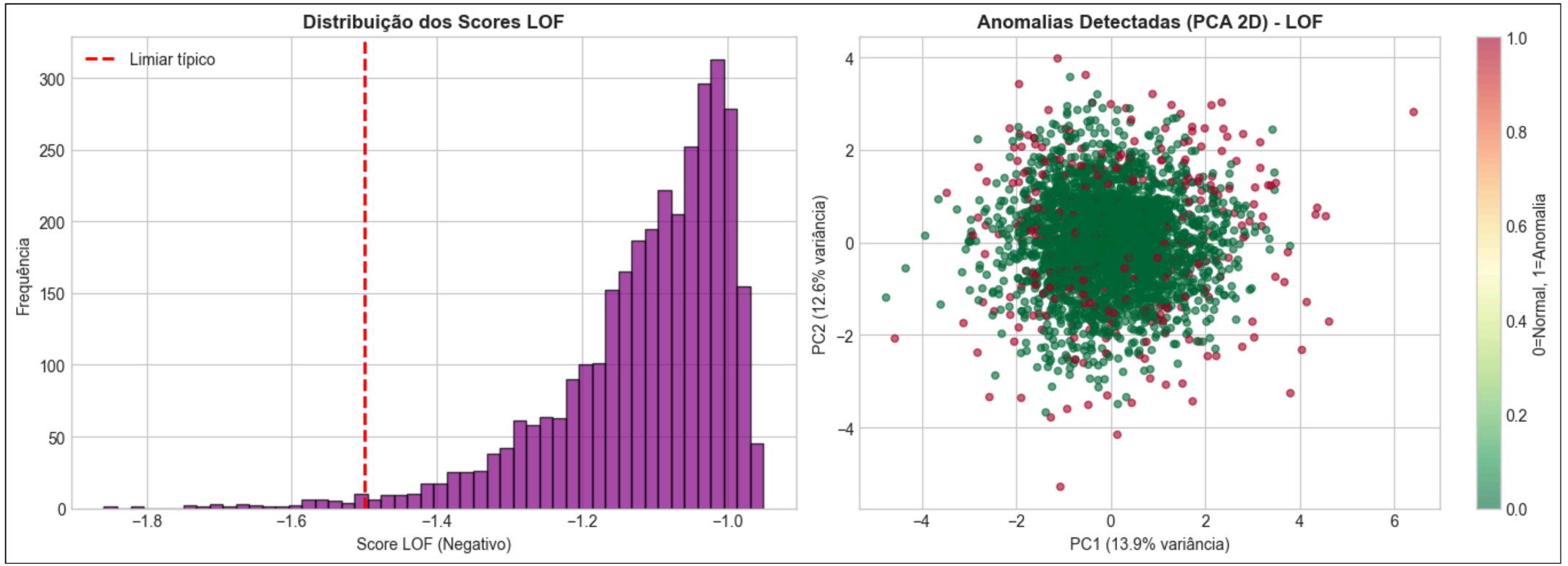
Parâmetros do modelo:

- Número de vizinhos (k): 20
- Taxa de contaminação: 10.0%

Resultados:

- Total de amostras: 3276
- Anomalias detectadas: 328 (10.01%)
- Amostras normais: 2948 (89.99%)

Local Outlier Factor (LOF)



Autoencoder Neural Network

Princípio de Funcionamento:

Encoder: Comprime os dados de alta dimensionalidade para uma representação latente.

Bottleneck: Camada central com menor dimensionalidade (gargalo)

Decoder: Reconstrói os dados originais a partir da representação latente.

Erro de Reconstrução: MSE entre entrada e saída.

Autoencoder para Detecção de Anomalias

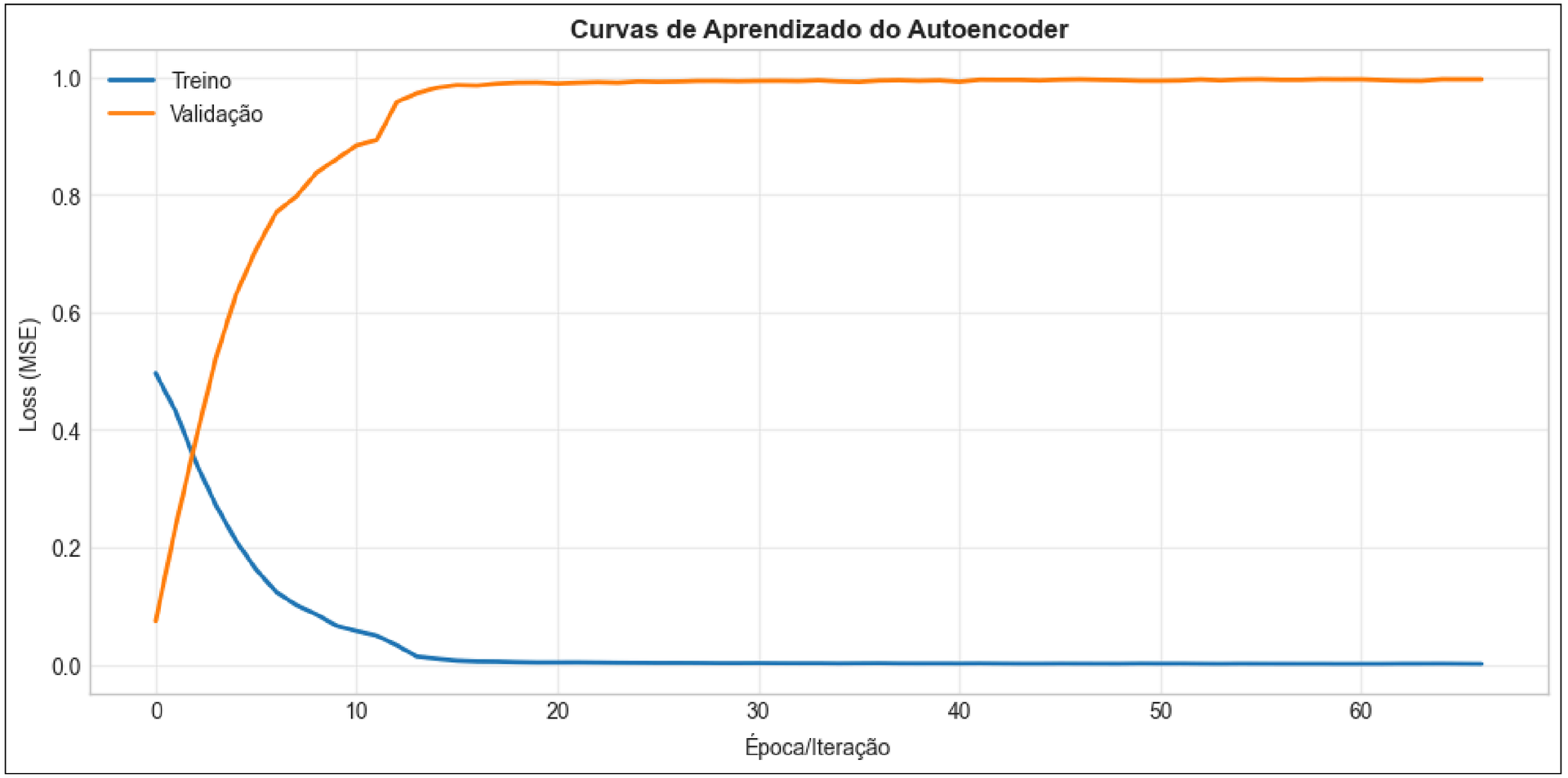
Hiperparâmetros Utilizados:

- **hidden_layer_sizes=(128, 64, 32, 16, 32, 64, 128):** Arquitetura simétrica
- **activation='relu':** Função de ativação ReLU.
- **solver='adam':** Otimizador Adam.
- **early_stopping=True:** Parada antecipada para evitar overfitting.

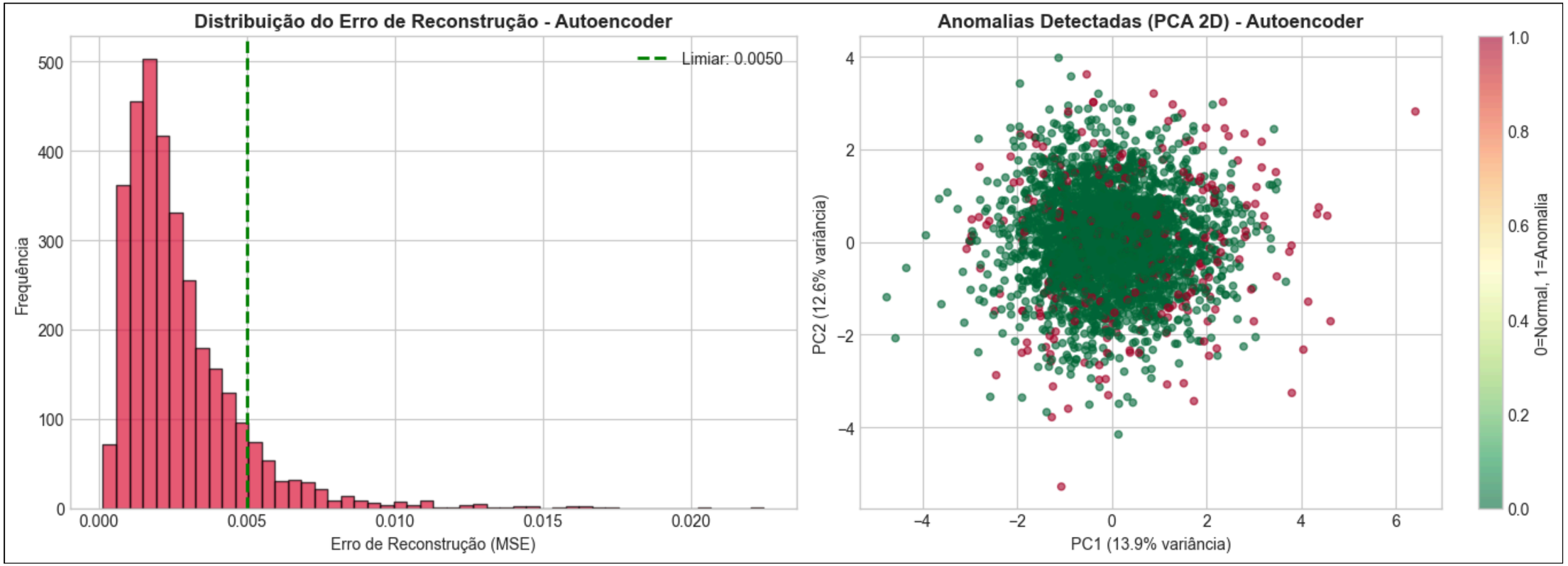
Resultados do Autoencoder:

- Limiar de erro (percentil 90): 0.0050
- Total de amostras: 3276
- Anomalias detectadas: 328 (10.01%)
- Amostras normais: 2948 (89.99%)

Autoencoder Neural Network



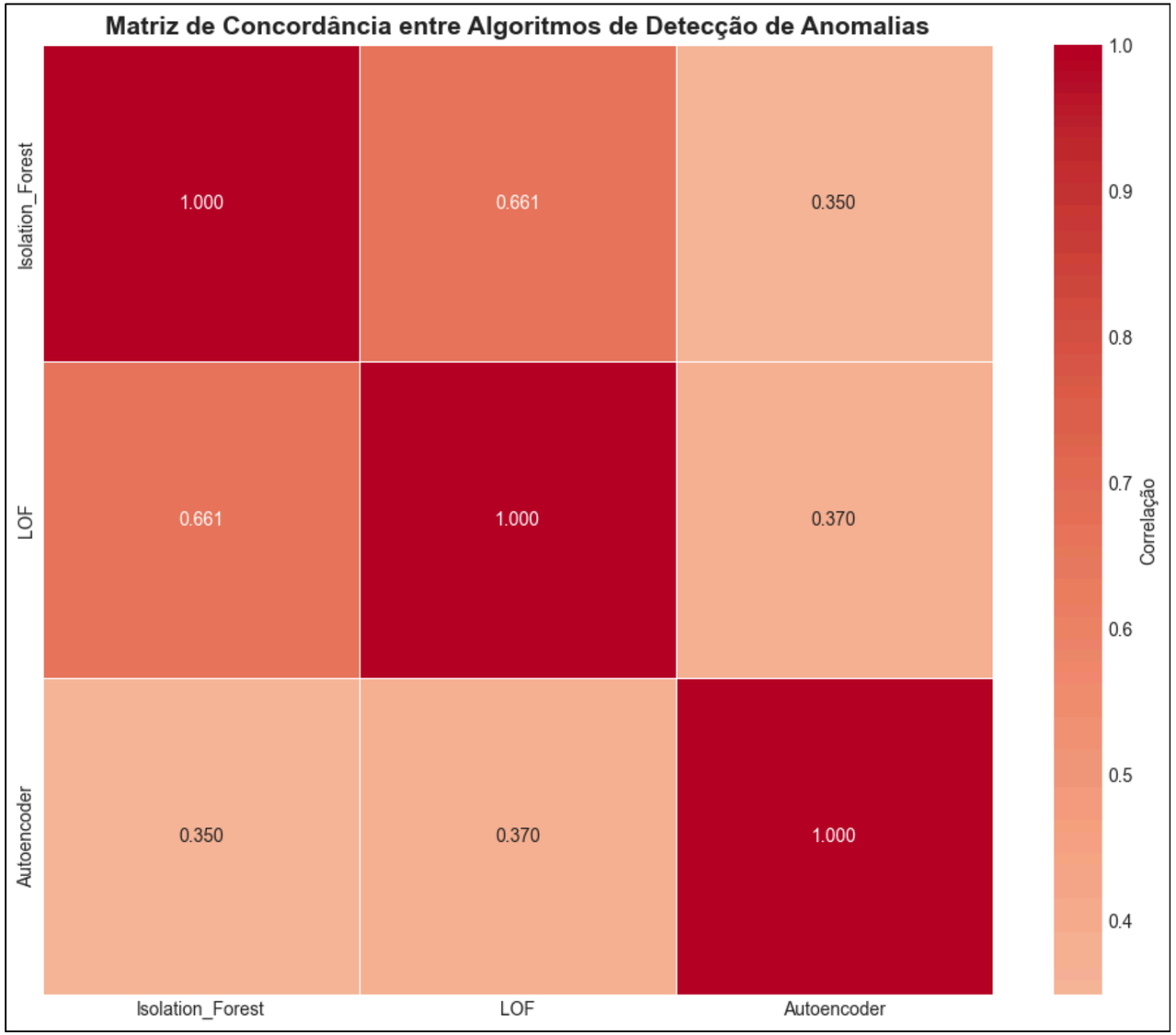
Autoencoder Neural Network



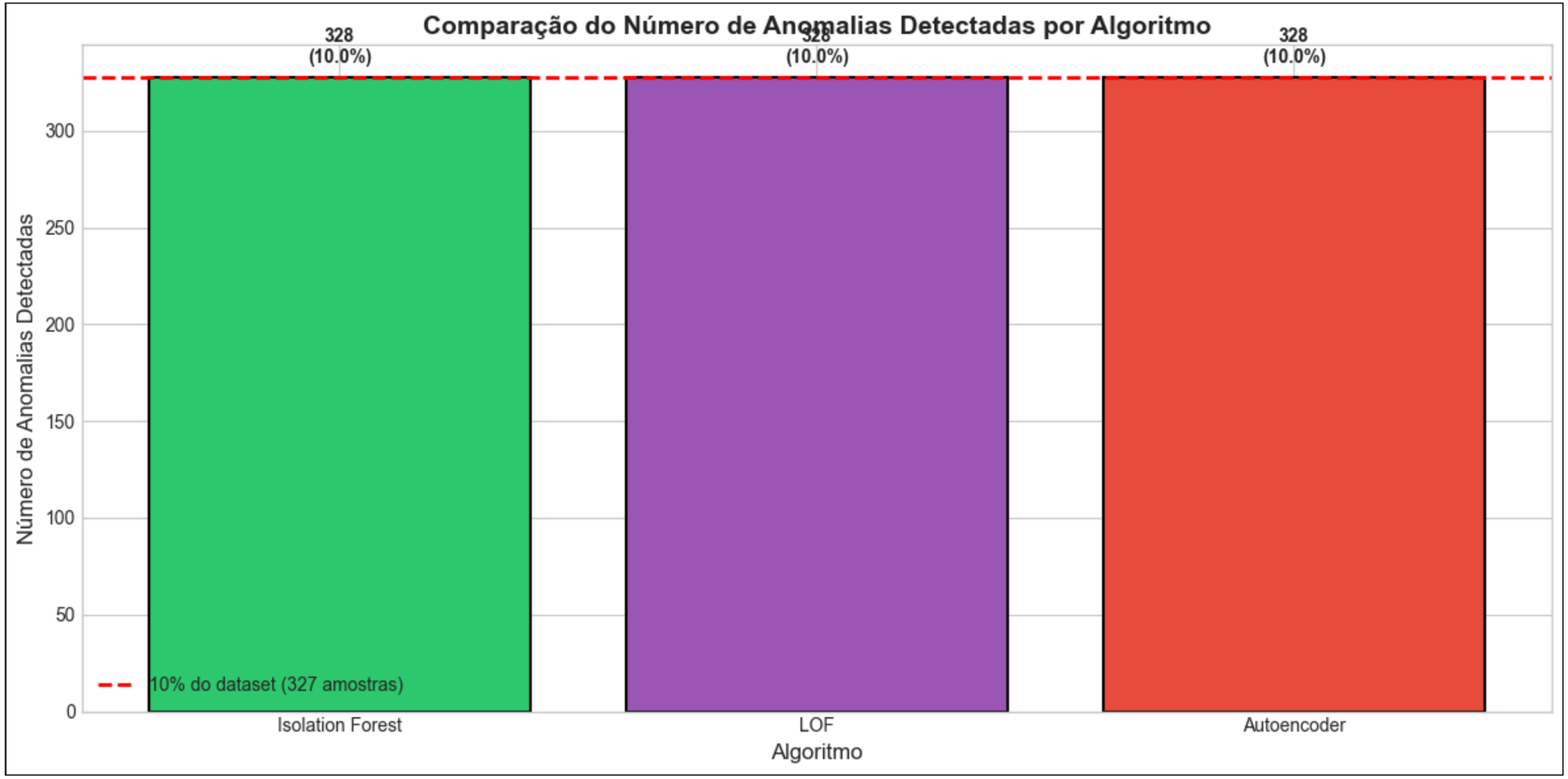
Comparação e Avaliação dos Modelos

Interpretação:

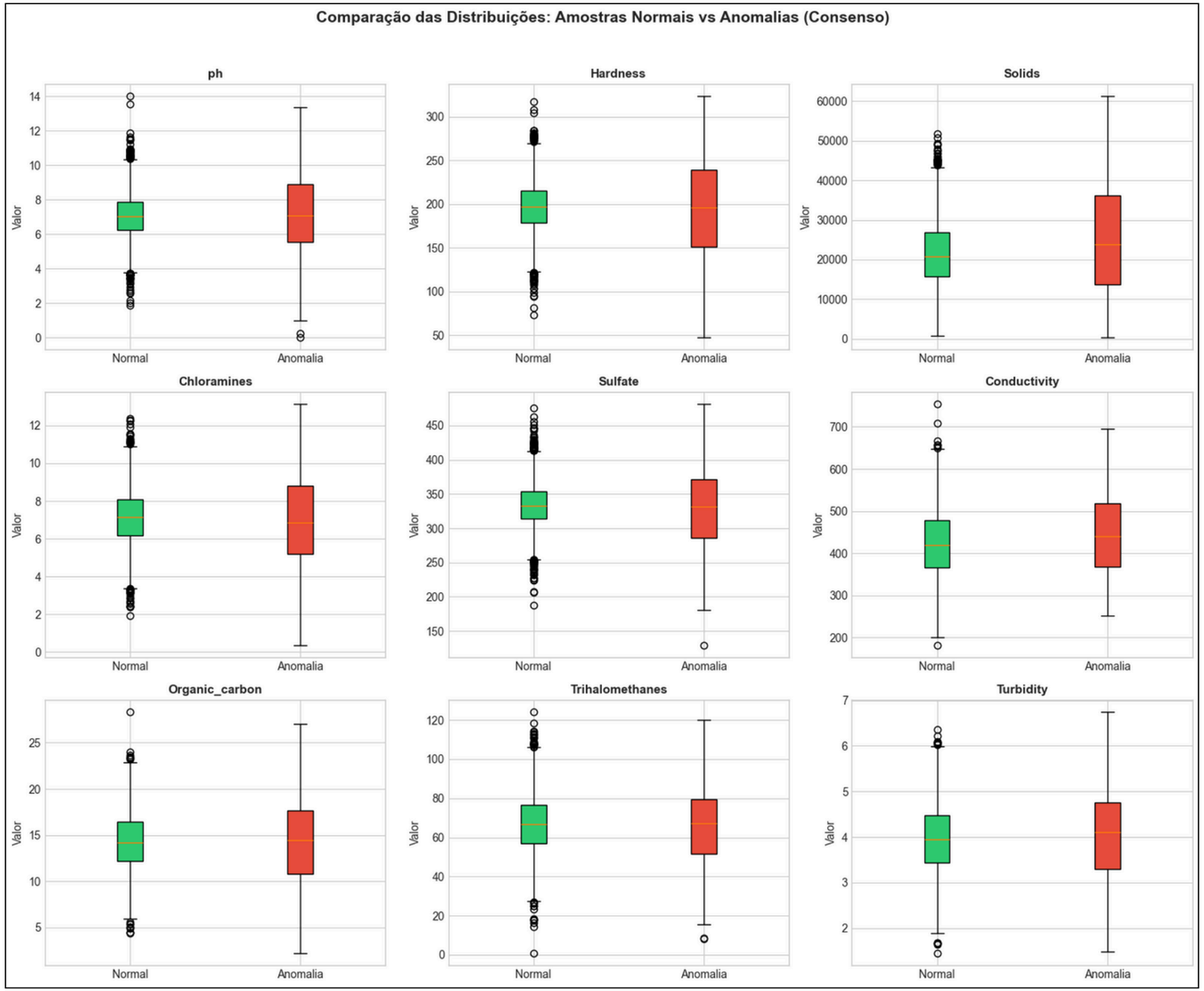
- Valores próximos de 1 indicam alta concordância entre os métodos
- Valores próximos de 0 indicam baixa concordância
- Métodos com alta concordância tendem a identificar as mesmas anomalias.



Comparação e Avaliação dos Modelos



Análise Detalhada das Anomalias Detectadas



Análise Detalhada das Anomalias Detectadas