UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

DETECÇÃO DE ANOMALIAS UTILIZANDO AUTOENCODER VARIACIONAL

João Pedro Marin Comini
Orientador: Prof. Dr. Kelton Augusto Pontara Costa

AGENDA

INTRODUÇÃO

FUNDAMENTAÇÃO TEÓRICA

METODOLOGIA

RESULTADOS

CONCLUSÃO

UM ATAQUE HACKER A CADA 29 SEGUNDOS

Pesquisa da Universidade de Maryland

PROTEÇÃO: SDIs, ANTIVIRUS E FIREWALL

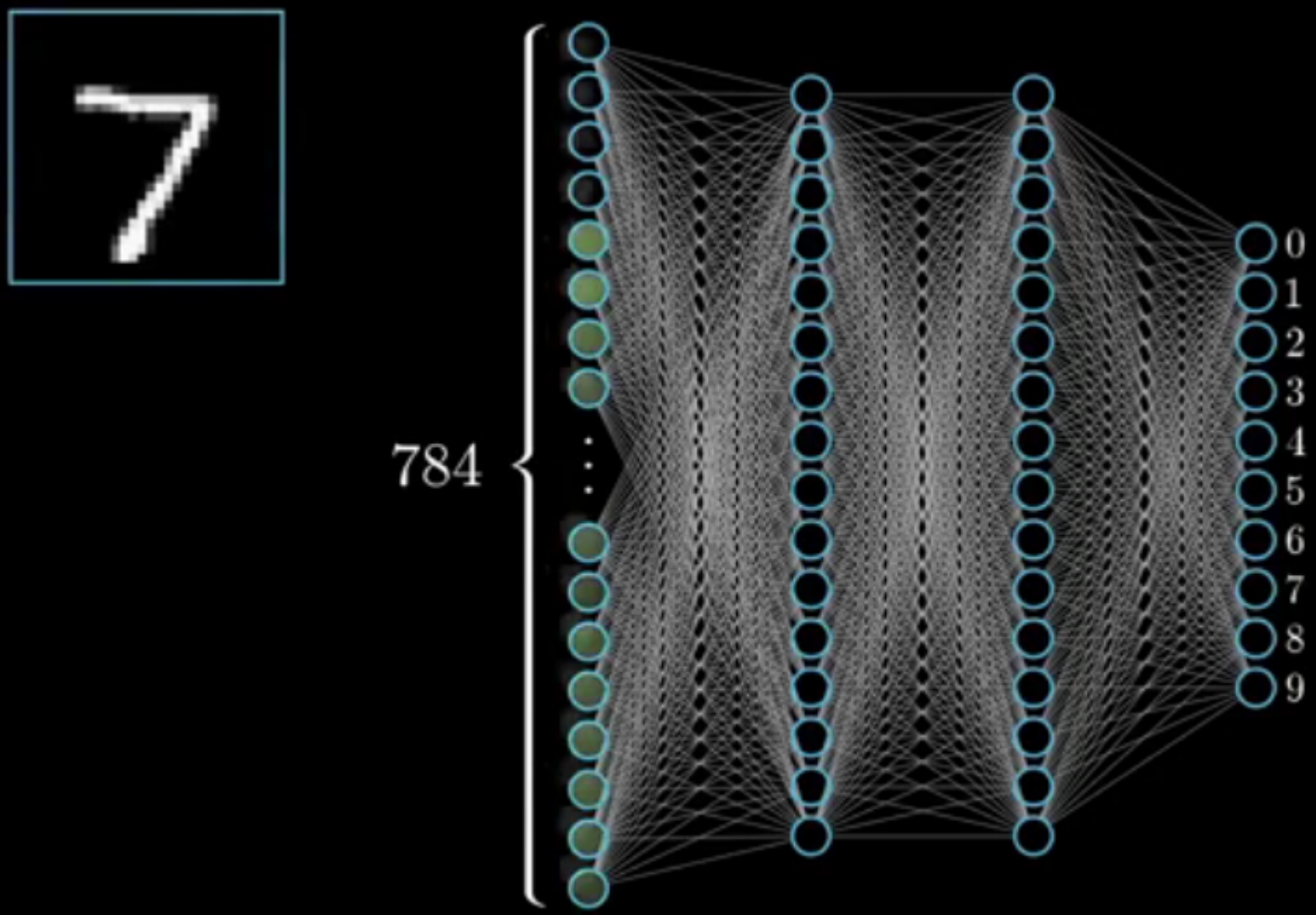
Camadas complexas para preservar os usuários e servidores.

DETECÇÃO DE ANOMALIAS

SDIs baseados em anomalias tentam encontrar conexões incomuns em seu contexto.

APRENDIZADO DE MÁQUNA

Vastamente utilizada para detecção de anomalias em diversos campos.



AUTOENCODER VARIACIONAL

Estudo sobre o modelo proposto no artigo de Kingma e Welling (2013).

SELEÇÃO DOS DADOS

Tratatamento do conjunto de dados para utilização no treinamento do modelo.

DESENVOLVIMENTO DO MODELO

Implementar abordagens para detecção de anomalias.

RESULTADOS E COMPARAÇÃO

Comparar os resultados com outros modelos já estabelecidos.

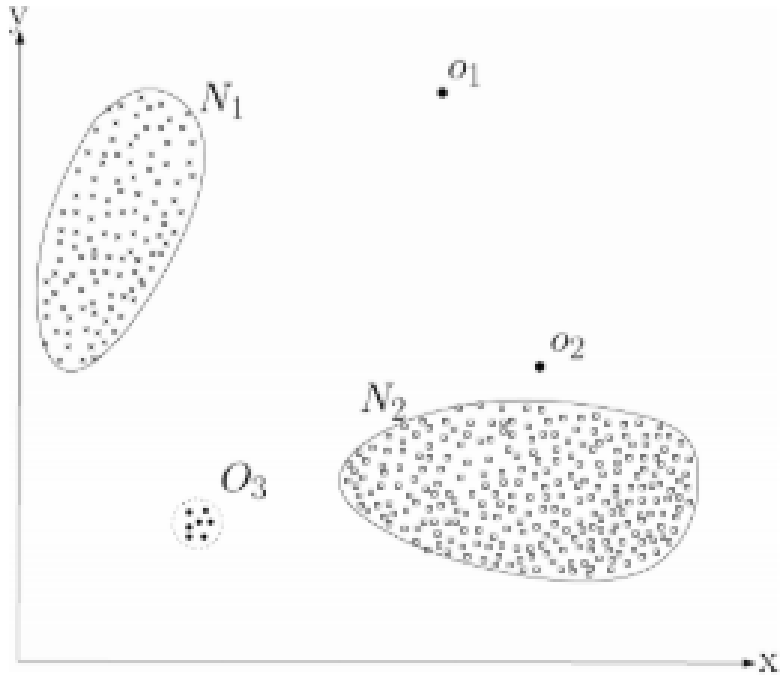
DETECÇÃO DE ANOMALIAS

Consiste em encontrar padrões extraordinários no contexto em questão.

Dados que fogem do padrão definido como "normal", são classificados como anomalias.

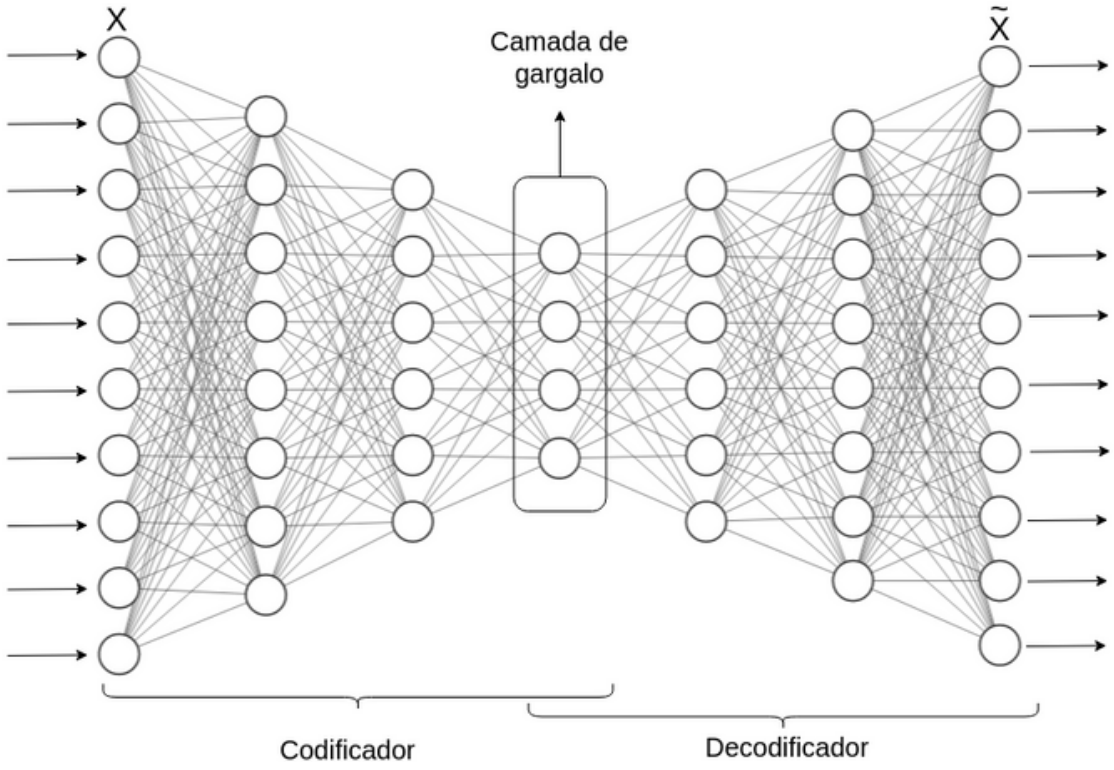
Obstáculos:

- Definir conceito de normalidade.
- Anomalias resultadas de ações maliciosas se camuflam.
- A noção de anomalia difere muito dependendo do domínio de aplicação.



AUTOENCODER

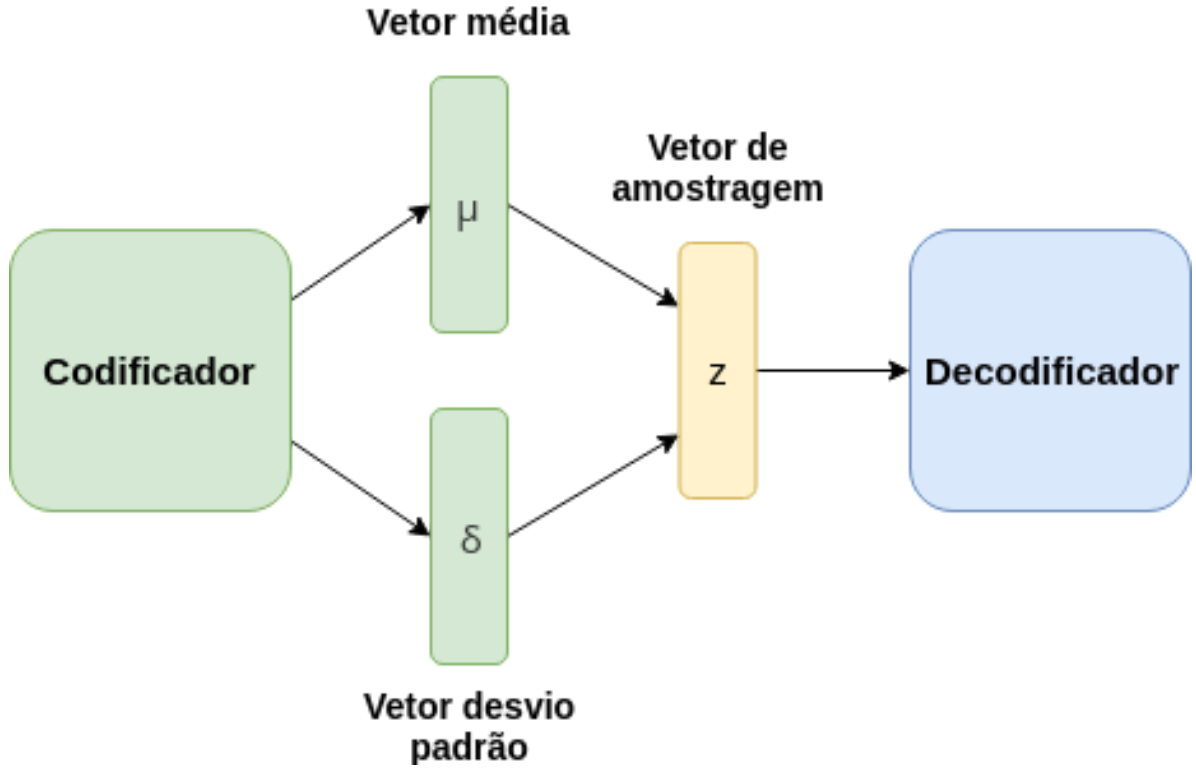
- É um modelo de rede neural em aprendizado de máquina que objetiva reconstruir o conjunto de dados.
- Dada uma entrada X, espera-se como resultado uma saida $Y \simeq X$, ou seja $f(X) \simeq X$.



INFERÊNCIA VARIACIONAL + AUTOENCODER

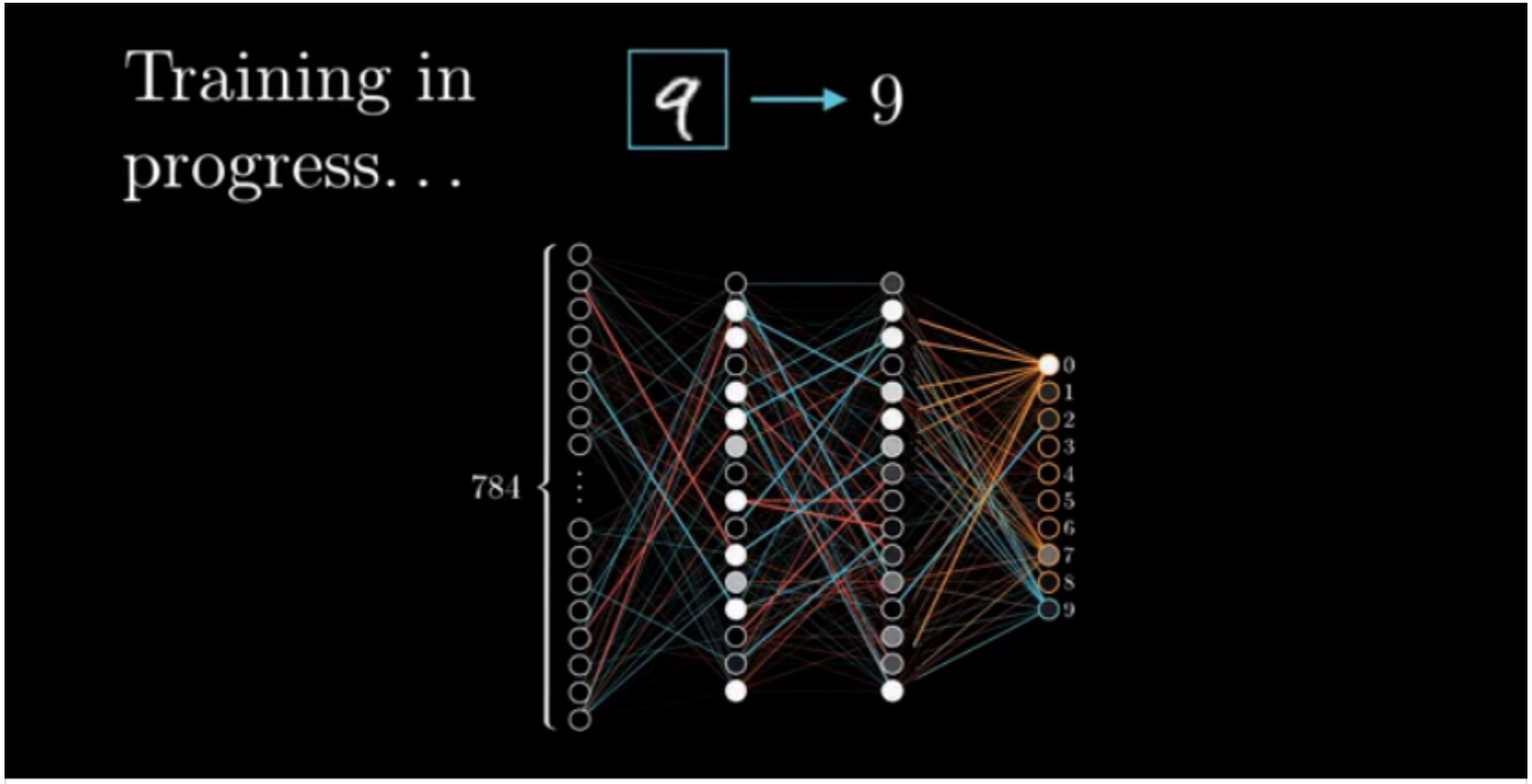
O autoencoder variacional possui modificações em relação ao modelo anterior:

- O codificador e o decodificador são substituidos por modelos probabilísticos $q(z|x)$ e $p(x|z)$, respectivamente.
- A camada de gargalo é composta de três camadas de dimensões iguais: camada de média, camada de desvio padrão e camada de amostragem.



OTIMIZAÇÃO

- A rede é otimizada utilizando o algoritmo de *backpropagation*.



OTIMIZAÇÃO

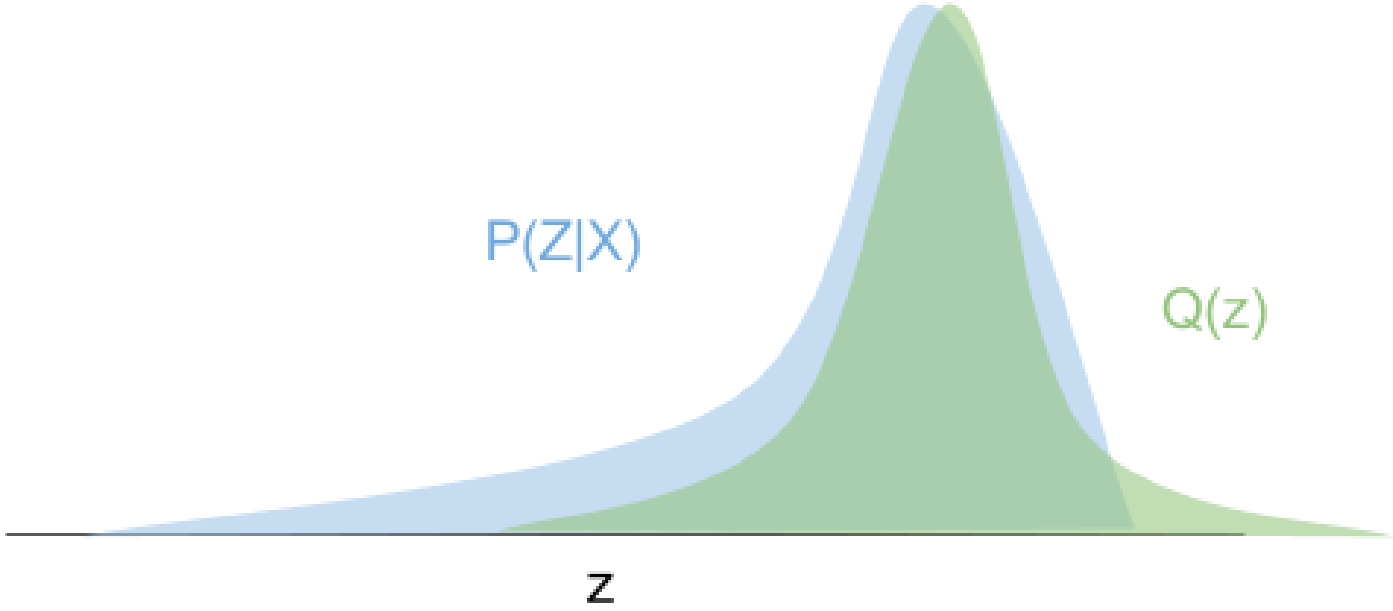
- Autoencoders comuns possuem apenas o erro de reconstrução na função custo. Uma função custo muito comum é a função de Erro Quadratico Médio (EQM ou SME).

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(X_i - \tilde{X}_i)^2$$

- Em autoencoder variacional, mais um termo é adicionado à funcão custo, chamado de Divergência de Kullback-Leibler.

$$D_{KL}(p||q) = \mathbb{E}_{x\sim p}[\log p(x) - \log q(x)]$$

- A Divergência de Kullback-Leibler é o termo de regularização da função custo de um Autoencoder Variacional.
- A sua adição permite aproximar a distribuição intratável P($z|x$) para uma distribuição tratável e conhecida Q($z|x$).



TRUQUE DE REPARAMETRIZAÇÃO

Devido à operação de amostragem em um autoencoder variacional ser descontínua, uma reparametrização é feita para que seus parâmetros possam ser otimizados através do algoritmo de *backpropagation*.

$$z \sim q(z|x) = \mathcal{N}(\mu, \sigma^2)$$

$$z = \mu + \sigma \odot \epsilon \,|\, \epsilon \sim \mathcal{N}(0, I)$$

CONJUNTO NSL-KDD

- Variação do famoso conjunto de dados KDDCup99.
- Corrige algumas das falhas do conjunto original.
- 40 classes: a classe normal e outras 39 classes maliciosas que podem ser divididas em 4 categorias de ataque: DoS, Probing, U2R, R2L.

DURAÇÃO	PROTOCOLO	SERVIÇO	FLAG	SRC_BYTES	DEST_BYTES	...
0	tcp	ftp_data	0	492	0	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
2	tcp	http	1	92	0	...

CONJUNTO NSL-KDD

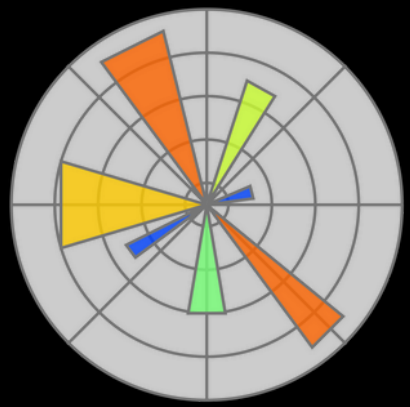
	Originais	Distintos
Ataques	3925650	262178
Normais	972781	812814
Total	4898431	1074992

Como citado anteriormente, o conjunto NSL-KDD corrige falhas encontradas no conjunto KDDCup99. A principal mudança é a remoção de registros redundantes.

- Redução de 78.05% na quantidade de registros.

METODOLOGIA

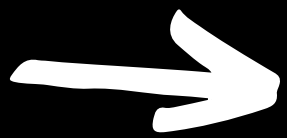
FERRAMENTAS



TRATAMENTO DO CONJUNTO DE DADOS

- One-Hot Encoding com Scikit-Learn:

PROTOCOLO
udp
tcp
icmp



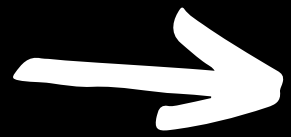
udp	tcp	icmp
1	0	0
0	1	0
0	0	1

Transformação de variáveis qualitativas em quantitativas

TRATAMENTO DO CONJUNTO DE DADOS

- Normalização com Scikit-Learn:

DURAÇÃO	SRC_BYTES
0	7
2	56
30	32



DURAÇÃO	SRC_BYTES
0	0.125
0.0357	1
0.5357	0.5714

Transforma a escala numérica dos dados em um intervalo conhecido.

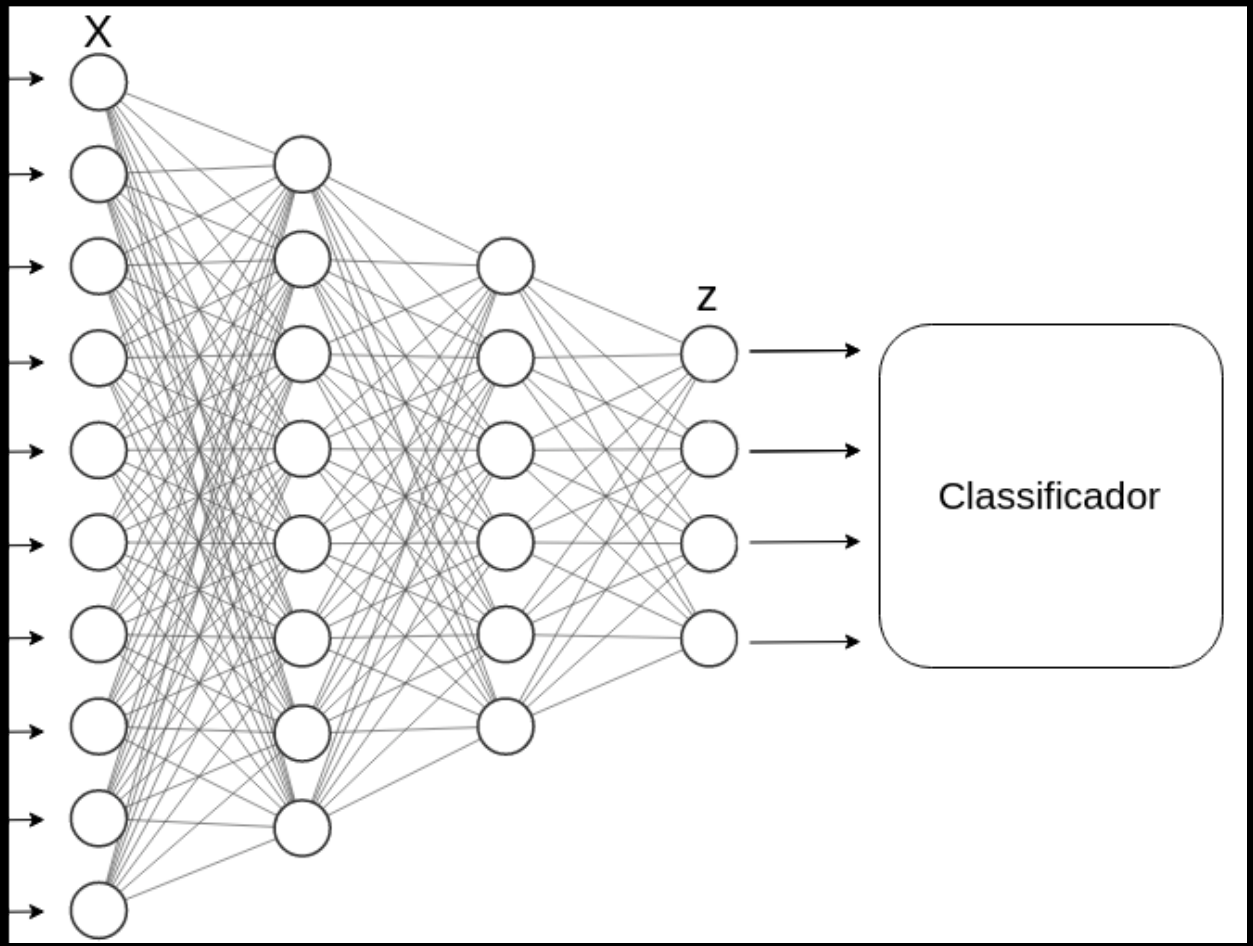
$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

DESENVOLVIMENTO DO MODELO

- Arquitetura do modelo possui 7 camadas com as seguintes dimensões: 96, 64, 32, 16, 32, 64 e 96 neurônios.
- Uso das bibliotecas TensorFlow e Keras API no treinamento devido a facilidade e fácil customização.
- Conjunto de dados dividido em:
  - 60% para treino;
  - 40% para teste.

ABORDAGENS UTILIZANDO O MODELO

- Detectando anomalias através da representação codificada:



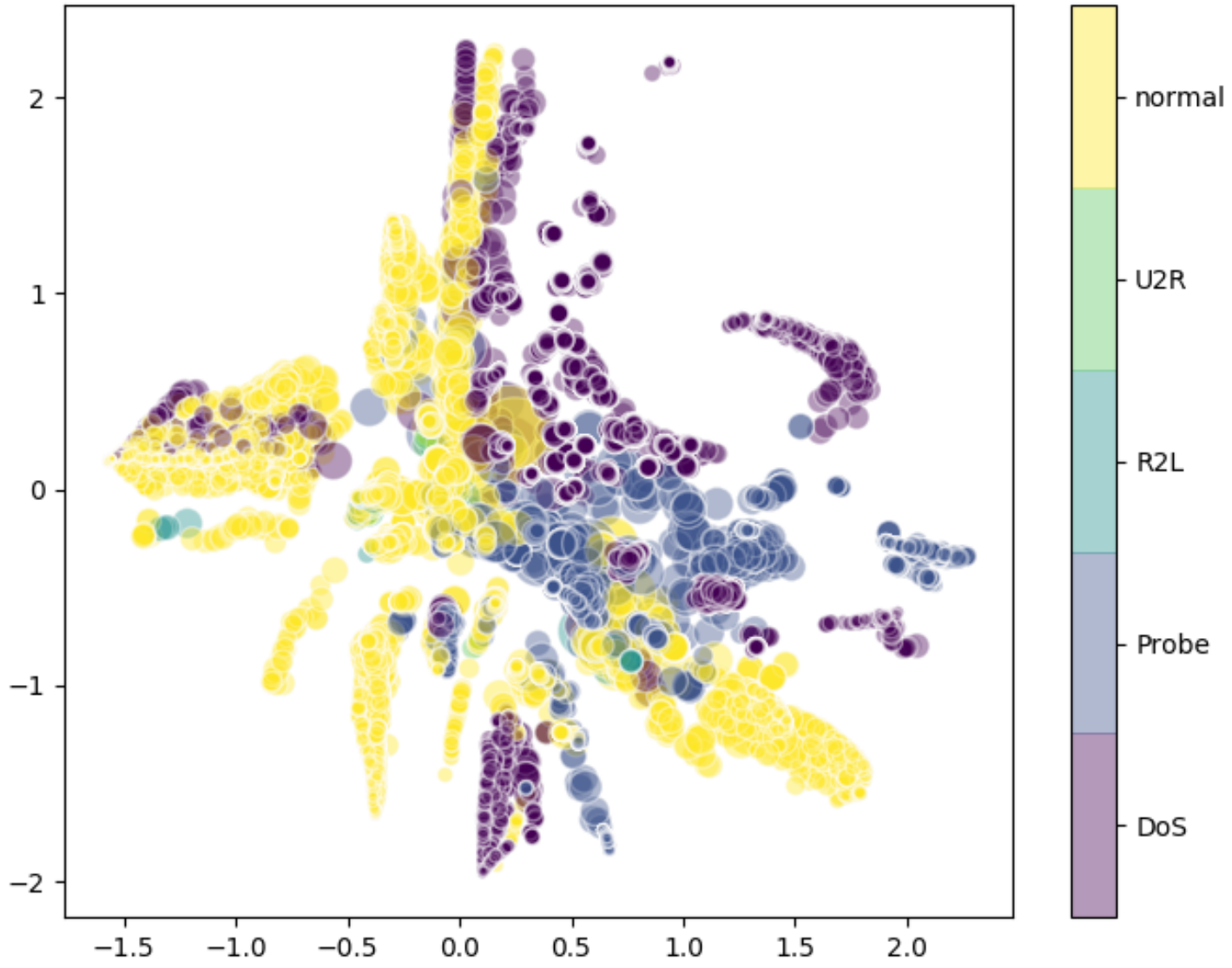
ABORDAGENS UTILIZANDO O MODELO

- Detectanto anomalias através do erro de reconstrução probabilístico:
  - Treinamento do modelo apenas com dados normais.
  - Reconstruir os dados com o modelo e calcular seus erros de reconstrução.
  - Classificar como anômalos os dados com um erro ϵ maior que o limite L estabelecido.
- Esta abordagem parte do princípio de que o modelo aprendeu apenas a reconstuir dados normais, portanto o erro de reconstrução dos dados anormais é maior.

RESULTADOS

CLASSIFICANDO ATRAVÉS DA REPRESENTAÇÃO CODIFICADA

Espaço latente do autoencoder após o treinamento:



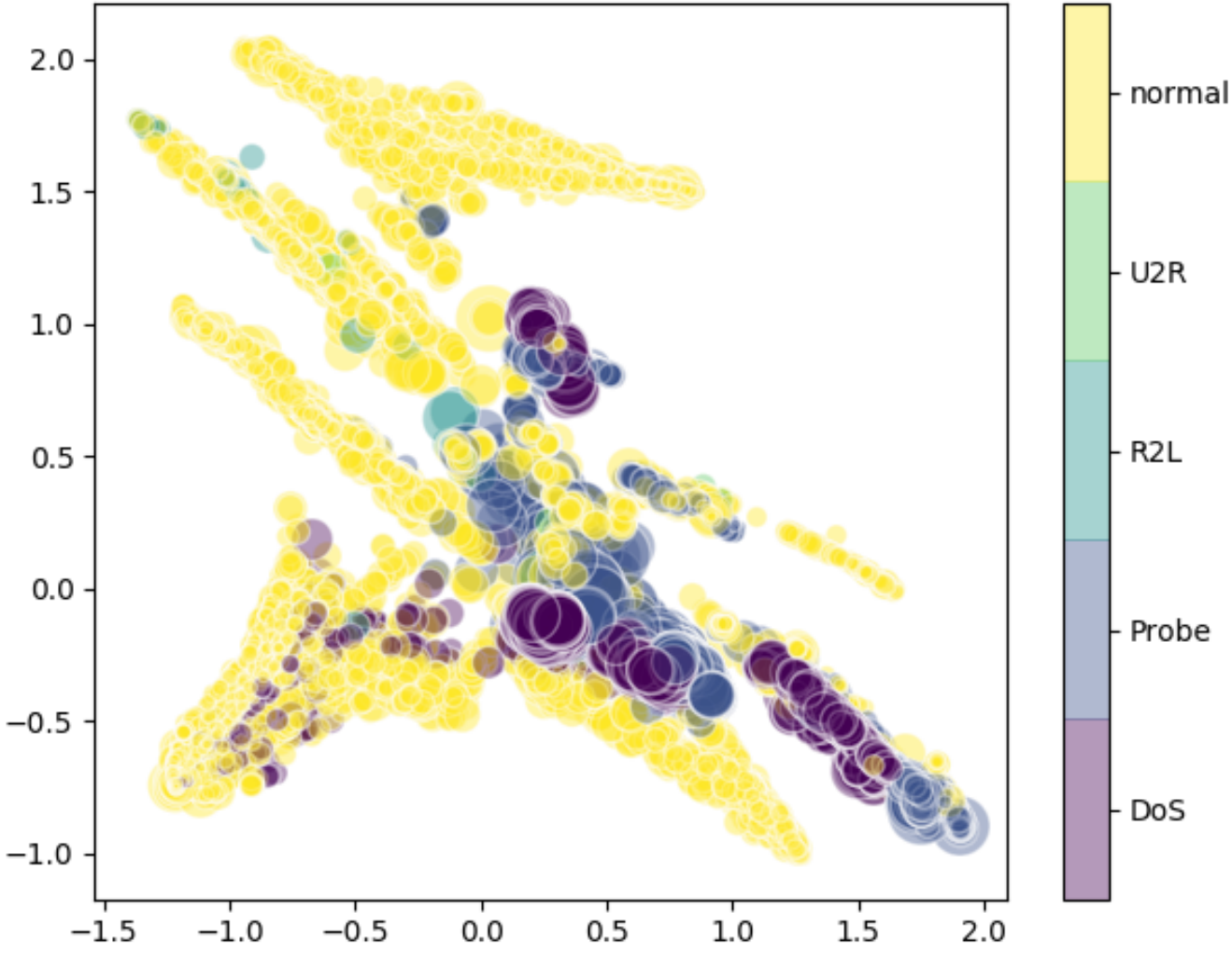
CLASSIFICANDO ATRAVÉS DA REPRESENTAÇÃO CODIFICADA

Comparação do treinamento de classificadores treinados com os dados codificados e com os dados originais.

	dados originais	dados codificados
Rede Neural	99.62%	97.63%
Naive Bayes	85.90%	87.21%
SVM	98.49%	96.60%
RFC	99.89%	99.74%

CLASSIFICANDO ATRAVÉS DO ERRO DE RECONSTRUÇÃO

Espaço latente do autoencoder após o treinamento com apenas dados normais:



CLASSIFICANDO ATRAVÉS DO ERRO DE RECONSTRUÇÃO

Percebe-se que, com este método, o autoencoder possui a capacidade de reconstruir com mais precisão apenas os dados considerados normais.

Neste método, classificam-se como anômalos os dados que possuirem um error de reconstrução maior que um limite estabelecido.

Com esta simples abordagem, foi possível classificar corretamente 93,51% dos dados. Isso demonstra a capacidade do modelo de extrair informações relevantes do contexto.

CONCLUSÃO

Após o desenvolvimento do projeto e através dos resultados obtidos é possível concluir:

- Os objetivos do trabalho foram cumpridos;
- As técnicas apresentadas podem ser generalizadas para qualquer área;
- Resultados satisfatórios foram obtidos;
- Ainda há espaço para melhorias no modelo.

REFERÊNCIAS

CUKIER, M. Study: Hackers Attack Every 39 Seconds. 2007. Disponível em: <https://eng.umd.edu/news/story/study-hackers-attack-every-39-seconds>. Acesso em: 29 Set. 2019.

YOUSEFI-AZAR, M.; VARADHARAJAN, V.; HAMEY, L.; TUPAKULA, U. Autoencoder-based feature learning for cyber security applications. In: IEEE. 2017 International joint conference on neural networks (IJCNN). [S.l.], 2017. p. 3854–3861.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. ACM computing surveys (CSUR), ACM, v. 41, n. 3, p. 15, 2009.

DHANABAL, L.; SHANTHARAJAH, S. A study on nsl-kdd dataset for intrusion detection system based on classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering, v. 4, 6 2015.

KINGMA, D. P.; WELLING, M. Auto-encoding variational bayes. 2013.

CANADIAN INSTITUTE FOR CYBERSECURITY. NSL-KDD dataset. 2009. Disponível em: <https://www.unb.ca/cic/datasets/nsl.html>. Acesso em: 01 Out. 2019.

MUITO OBRIGADO!