Técnicas de Inteligência Artificial para diagnóstico de Acidente Vascular Cerebral através de imagens e dados textuais sobre possíveis vítimas

Vinícius de Paula Pilan (RA:191025399)

Orientador: Prof. Dr. Clayton Reginaldo Pereira

Conteúdo da apresentação

1. Problema e Justificativa
2. Introdução
3. Ferramentas e bases de dados utilizadas
4. Desenvolvimento
5. Resultados
6. Conclusão

Referências Bibliográficas

1. Problema e Justificativa

Problema e Justificativa

Problema:

- Acidente Vascular Cerebral é uma das doenças que mais causa mortes, incapacitações e internações no mundo
- Quanto mais tardio é realizado o diagnóstico e tratamento maiores são os prejuízos e sequelas para a vítima

Problema e Justificativa

Justificativa:

- É de grande importância facilitar e agilizar o diagnóstico da doença
  - **Minimiza consequências e sequelas**, facilitando muito no processo de reabilitação
  - **Reduz quantidade de casos com maior gravidade** e até mesmo **casos de óbitos**
- É possível **desenvolver formas de auxílio** para o diagnóstico utilizando técnicas da **Inteligência Artificial** (reconhecimento de padrões)

2. Introdução

Introdução – Acidente Vascular Cerebral (AVC)

- Doença causada pela alteração do fluxo sanguíneo na região cerebral
  - **AVC Isquêmico (AVCi):** Obstrução total ou parcial de vaso sanguíneo
  - **AVC Hemorrágico (AVCh):** Rompimento de vaso sanguíneo

- Causa morte de células do cérebro
  - Falta de nutrientes e oxigênio

- Quadro pode ser identificado via Tomografia Computadorizada (TC)

Introdução – Acidente Vascular Cerebral (AVC)

Fatores de risco do AVC genéticos e fisiológicos:

- Envelhecimento
- Histórico familiar
- Sexo (masculino)

Fatores de risco do AVC relacionados a estilo de vida:

- Tabagismo
- Estresse
- Sedentarismo
- Consumo excessivo e frequente de álcool e drogas

Introdução – Acidente Vascular Cerebral (AVC)

Fatores de risco do AVC relacionados a patologias:

- Hipertensão
- Diabetes
- Obesidade
- Colesterol elevado
- Doenças cardiovasculares (principalmente as que produzem arritmia cardíaca)
- Doenças do sangue (ex: trombose)

Introdução – Aprendizado de Máquina

- Subárea da Inteligência Artificial (IA)

- Sistemas capazes de aprender comportamentos, reconhecer padrões e especular resultados
  - estimar valores
  - fazer classificações preditivas

Basicamente, três tipos de aprendizado:

1. **supervisionado**
2. não supervisionado
3. por reforço

Introdução – Aprendizado supervisionado

- Utiliza dados rotulados no treinamento do modelo

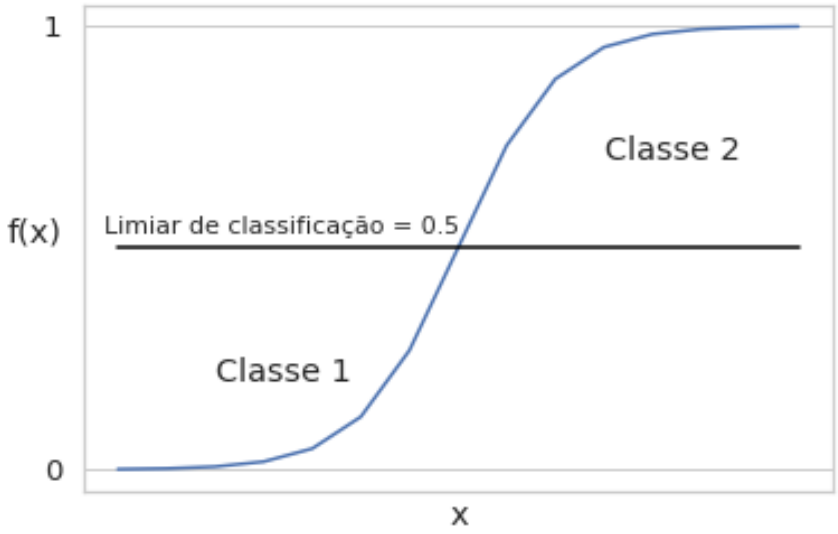
- Principais modelos:
  - regressão linear, **regressão logística**, máquina de suporte vetorial, árvores de decisão, k-vizinhos mais próximos, **floresta aleatória**, entre outros...

Introdução – Aprendizado supervisionado

Modelo de Regressão Logística:

- **Estimar probabilidade de um evento** acontecer a partir da combinação linear de variáveis independentes entre si (x * θ)

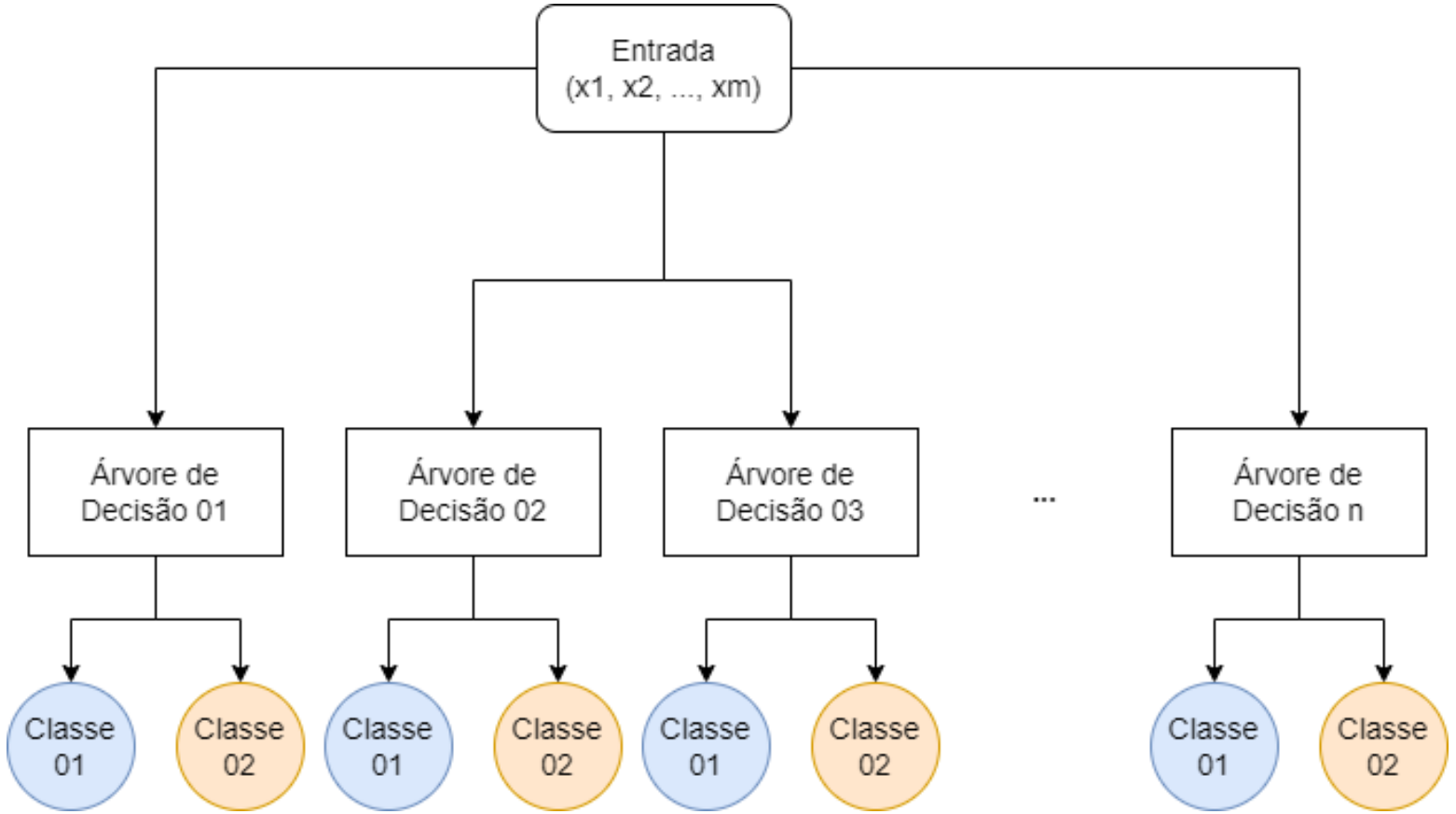
- Regressão que utiliza a função logística



Fonte: Elaborado pelo autor.

Introdução – Aprendizado supervisionado

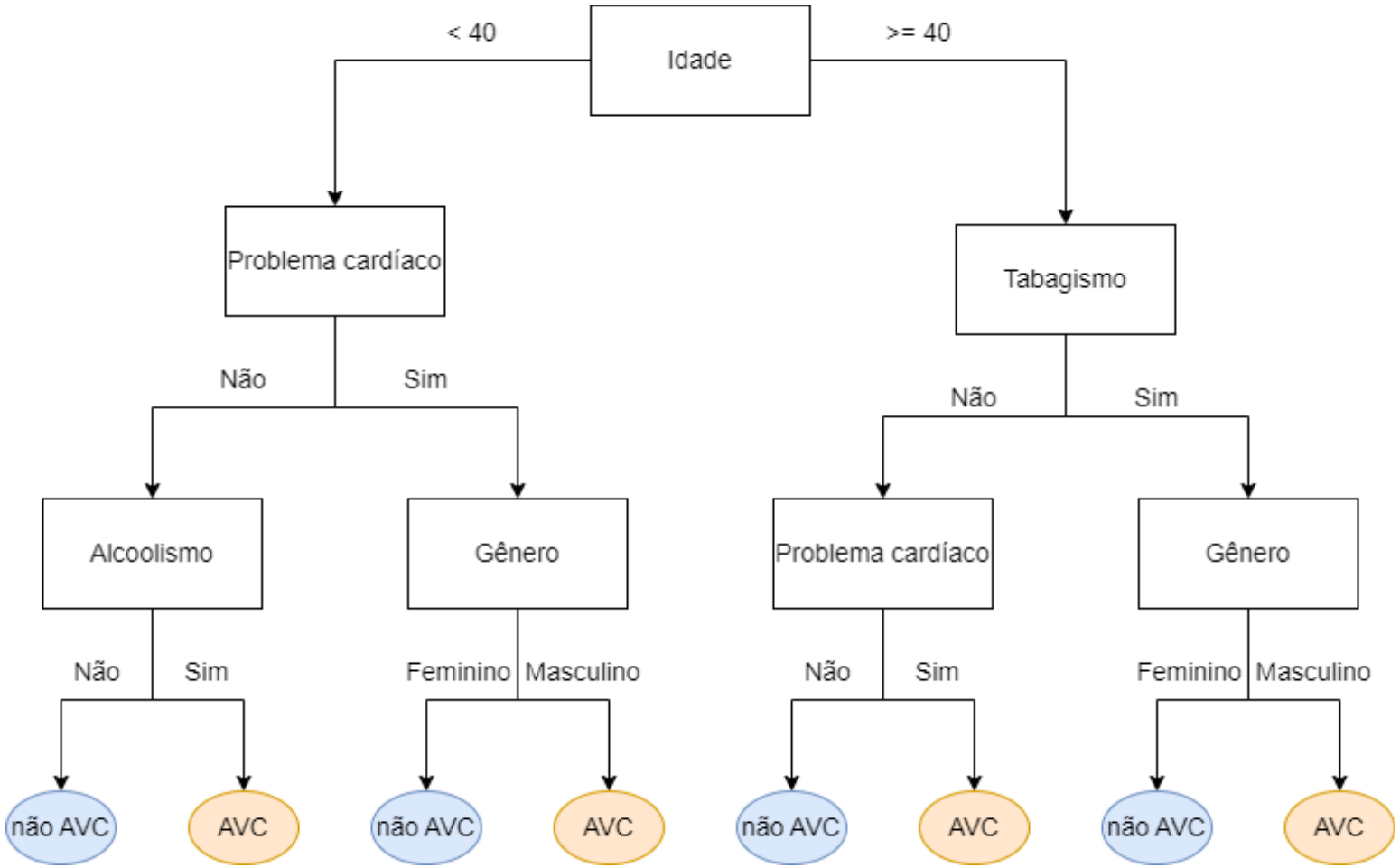
Floresta Aleatória:



Fonte: Elaborado pelo autor.

Introdução – Aprendizado supervisionado

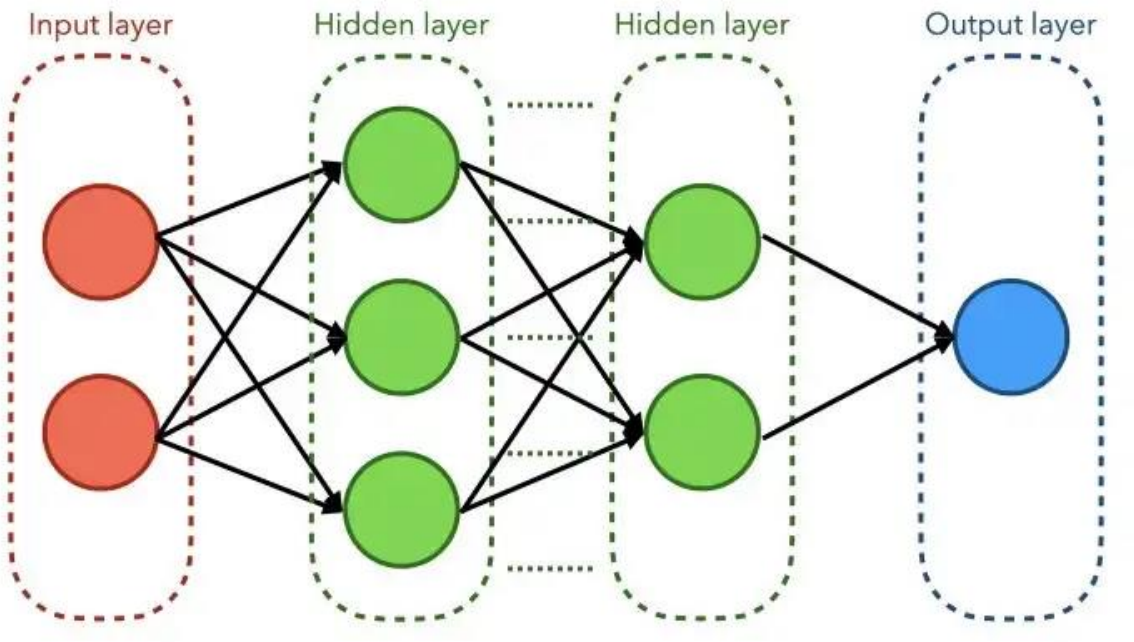
Árvore de Decisão:



Fonte: Elaborado pelo autor.

Introdução – Aprendizagem profunda

Rede Neural Artificial (*Artificial Neural Network* - ANN):



Fonte: (NURFIKRI, 2020).

Introdução – Aprendizagem profunda

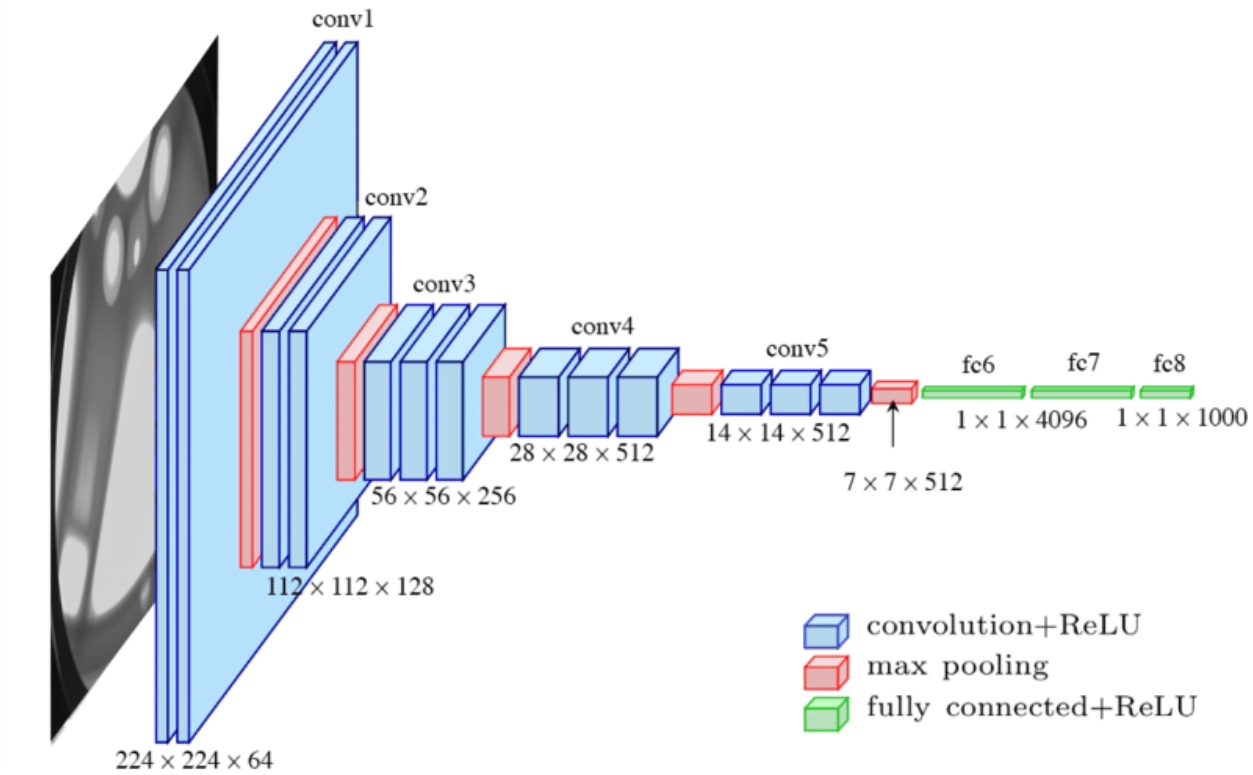
Redes Neurais Convolucionais (*Convolutional Neural Network* - CNN):

- Utilizadas para identificação de objetos, reconhecimento de características e classificação de imagens
- Valorizam detalhes da imagem relevantes para a classificação
- Três tipos principais de camadas em sua estrutura:
  - **convolução (*convolution*):** realce das características via aplicação do kernel
  - **agrupamento (*pooling*):** redução do número de valores recebidos mantendo as características realçadas pelos filtros da convolução
  - **totalmente conectadas (*fully connecteds*):** neurônios totalmente conectados que aplicam alguma função de ativação nos valores recebidos

Introdução – Aprendizagem profunda

Redes Neurais Convolucionais (*Convolutional Neural Network* - CNN)

VGG-16:



Fonte: (FERGUSON et al., 2017).

3. Ferramentas e bases de dados utilizadas

Ferramentas

- Python:
  - linguagem de programação de alto nível com alta disponibilidade de bibliotecas e recursos para Ciência de Dados e Aprendizado de Máquina

- Google Colab

Ferramentas

- Bibliotecas e Frameworks Python:
  - **Numpy e Pandas:** análise e gerenciamento de valores (vetores, matrizes, etc.) e conjuntos de dados (*datasets*)
  - **Matplotlib e Seaborn:** visualização gráfica de dados
  - **Scikit-learn:** biblioteca de implementações de modelos de Aprendizado de Máquina e métricas de avaliação de desempenho
  - **Tensor Flow:** plataforma de código aberta que facilita a criação de modelos de Aprendizagem de Máquina e Aprendizagem profunda

Base de dados fatores de risco

- Conjuntos de dados disponibilizados para propósitos educacionais da plataforma *Kaggle*

- 5110 registros rotulados de vítimas de AVC ou de condição normal
  - AVC: 249 elementos
  - Não AVC: 4861 elementos

Base de dados fatores de risco - Variáveis categóricas

1. **Sexo**
2. **Hipertensão**
3. **Doença cardíaca**
4. **É / já foi casado** (pode estar ligada a estilo de vida e estresse)
5. **Tipo de residência:** Urbana ou rural (ligada a estilo de vida)
6. **Tipo de emprego:** privado, funcionário público (governo), empreendedor ou nenhum
7. **Status com relação a tabagismo:** indivíduo nunca fumou, fuma formalmente, regularmente ou situação desconhecida
8. **Vítima de AVC** (variável alvo do trabalho)

Base de dados fatores de risco - Variáveis quantitativas

1. **Idade**
2. **Nível de glicose**
3. **BMI:** Índice de massa corpórea - IMC
4. **Id:** Identificação de cada registro (variável desconsiderada por não ter impacto na variável alvo)

Base de dados fatores de risco

Exemplos retirados da base de dados utilizada

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
6	53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
7	10434	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked	1
8	27419	Female	59.0	0	0	Yes	Private	Rural	76.15	NaN	Unknown	1
9	60491	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1

Fonte: Elaborado pelo autor.

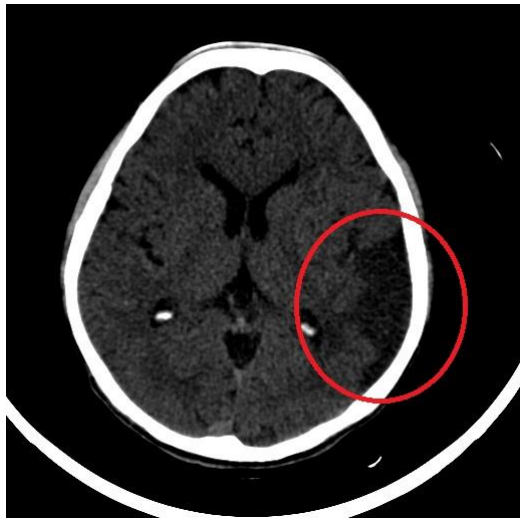
Base de imagens de Tomografia Computadorizada

- Conjunto de dados com 473 imagens, divididas em três classes diferentes:
  - **AVC Isquêmico**
  - **AVC Hemorrágico**
  - **Condição normal (não AVC)**

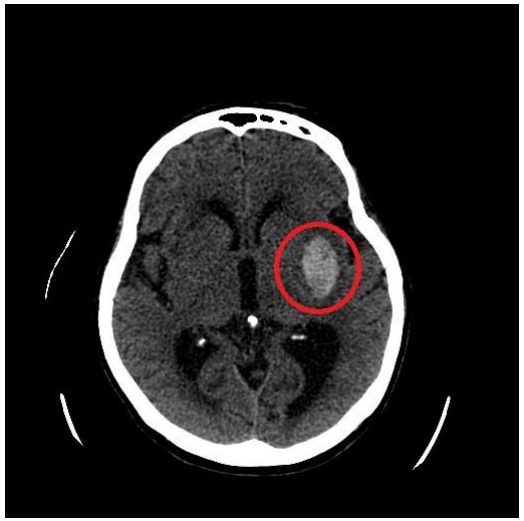
- No presente trabalho, conjunto foi redistribuído para apenas duas classes:
  - **AVC (AVCi + AVCh):** 299 imagens
  - **Condição normal (não AVC):** 174 imagens

Base de imagens de Tomografia Computadorizada

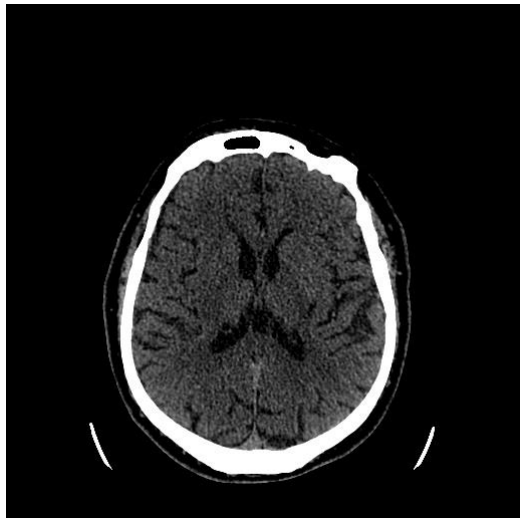
Exemplos retirados da base de dados utilizada



AVCi



AVCh



Condição normal

Fonte: Elaborado pelo autor.

4. Desenvolvimento

Classificador fatores de risco – pré-processamento

- Balanceamento do conjunto de dados (para 525 registros no total):
  - AVC: 249 → 249 elementos
  - Não AVC: 4861 → 276 elementos (subamostragem aleatória)

Classificador fatores de risco – pré-processamento

Atributo	Distribuição dos valores
Gênero	Masculino: 41.71%
	Feminino: 58.29%
Hipertensão	0: 82.48%
	1: 17.52%
Doença cardíaca	0: 88.76%
	1: 11.24%
É / já foi casado	Sim: 80%
	Não: 20%

Fonte: Elaborado pelo autor.

Classificador fatores de risco – pré-processamento

Atributo	Distribuição dos valores
Tipo de emprego	Setor privado: 62.86%
	Autônomo: 22.09%
	Cargo público: 14.29%
	Nunca trabalhou: 0.76%
Tipo de residência	Urbano: 50.86%
	Rural: 49.14%
Tabagismo	Nunca fumou: 37.33%
	Fuma formalmente: 24.57%
	Fuma constantemente: 22.48%
	Situação desconhecida: 15.62%
AVC	0: 52.57%
	1: 47.43%

Fonte: Elaborado pelo autor.

Classificador fatores de risco – pré-processamento

Correções formato das variáveis

Até dois possíveis valores – substituição das classes por 0 e 1

Gênero
Masculino
Masculino
Feminino
Feminino
Feminino
Masculino
Feminino

→

Masculino
1
1
0
0
0
1
0

Fonte: Elaborado pelo autor.

Tipo de emprego
Privado
Privado
Autônomo
Privado
Criança
Cargo público
Autônomo

One-Hot Encoding →

Privado	Autônomo	Cargo público	Criança
1	0	0	0
1	0	0	0
0	1	0	0
1	0	0	0
0	0	0	1
0	0	1	0
0	1	0	0

Fonte: Elaborado pelo autor.

Mais de dois possíveis valores – criação de colunas binárias para cada valor

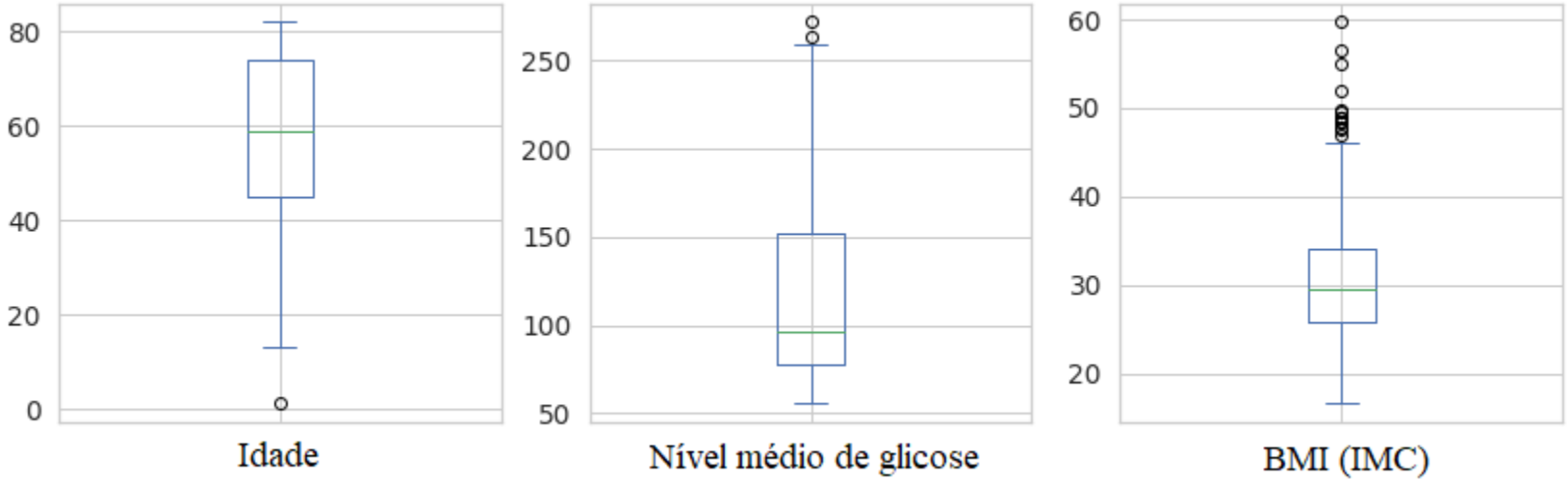
Classificador fatores de risco – pré-processamento

Dados nulos

- Única coluna com eles foi BMI: 7.6% (40 elementos em um total de 525 registros)
- Correção: substituição desses valores pela mediana da coluna
- A mediana foi escolhida pois gerou melhores resultados em comparação com a substituição pela média da coluna

Classificador fatores de risco – pré-processamento

Distribuição dos valores quantitativos e anômalos



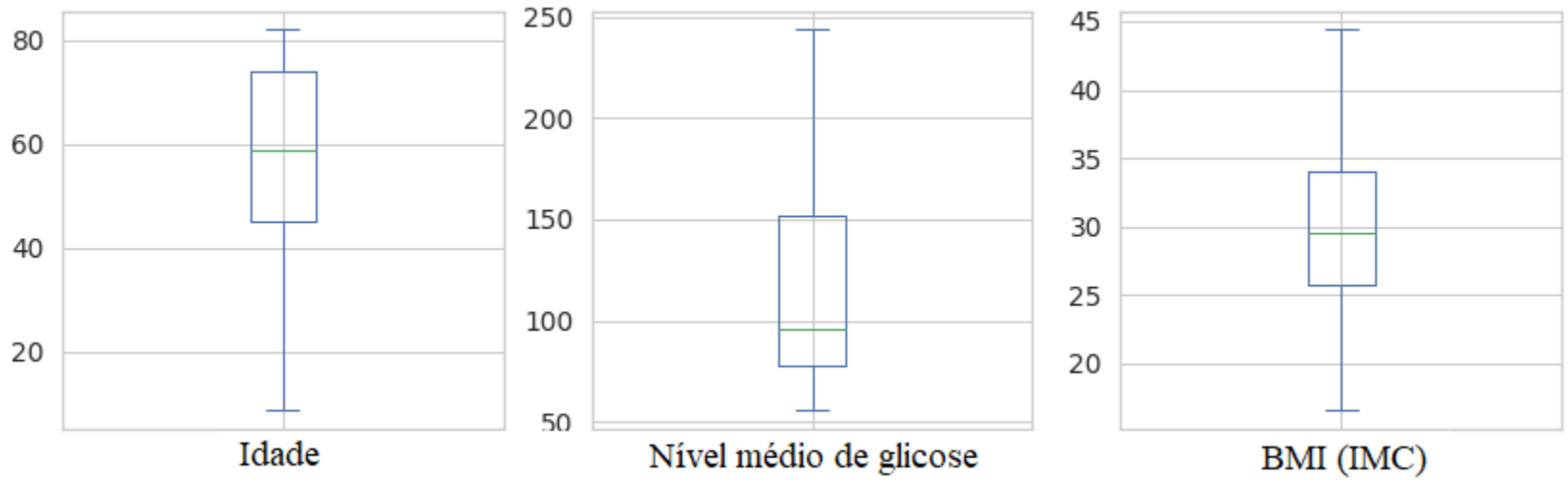
Fonte: Elaborado pelo autor.

Classificador fatores de risco – pré-processamento

- **IQR (*Interquartile range*):** avaliar o grau de dispersão dos valores em torno da medida de centralidade do conjunto
  - IQR = 3° quartil – 1° quartil
  - Limite superior = Mediana + 1.5 x IQR
  - Limite inferior = Mediana - 1.5 x IQR

Classificador fatores de risco – pré-processamento

Distribuição dos valores quantitativos corrigidos



Fonte: Elaborado pelo autor.

Classificador fatores de risco – pré-processamento

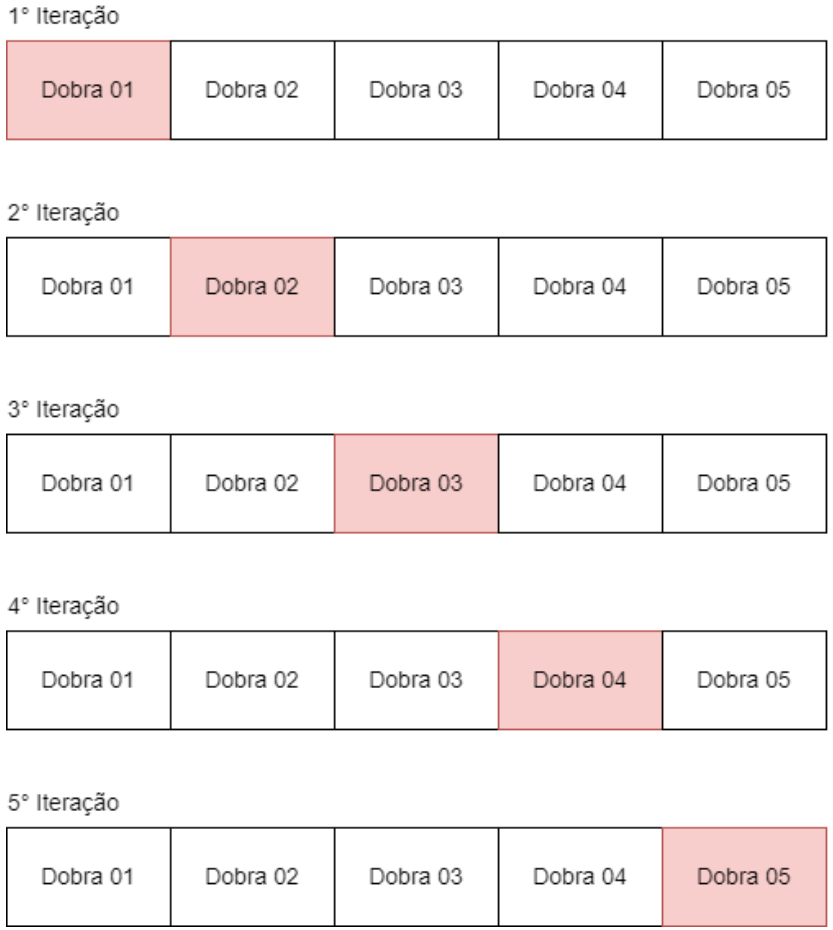
Normalização dos dados

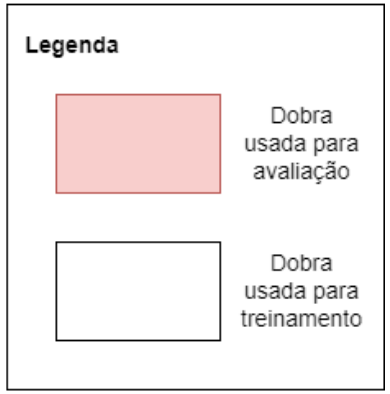
- sklearn.preprocessing.MinMaxScaler()
- padrão de escala ficou entre 0 e 1

Classificador fatores de risco – modelagem

Criação do classificador

- Validação cruzada: 05 dobras





Fonte: Elaborado pelo autor.

Classificador fatores de risco – modelagem

Criação do classificador

- Modelo com Regressão Logística (*seed* = 7)
- Modelo com Floresta Aleatória (*seed* = 7, n_arvores=200)

Rede neural classificadora de imagens de TC

Redimensionamento dos valores

- Camada inicial na rede do tipo *Rescaling* com valor para divisão de 255
- Redimensionamento do intervalo dos valores de cada canal de cor: de 0 a 255 para valores de 0 a 1

Rede neural classificadora de imagens de TC

Aumento dos dados do conjunto de dados original

- técnicas de aumento dos dados (data augmentation)

Alteração aplicada	Fator	Método utilizado
Inversão (*flip*)	Horizontal/ vertical	*tensorflow.keras.layers.RandomFlip*("horizontal_and_vertical", seed=10)
Rotação	0.2	*tensorflow.tf.keras.layers.RandomRotation*(0.2, seed=7)
Contraste	Valor entre [0.3, 0.5]	*tensorflow.keras.layers.RandomContrast*(factor=(0.3, 0.5), seed=(7,17))

Fonte: Elaborado pelo autor.

Rede neural classificadora de imagens de TC

Estrutura da rede neural criada

	Componente da rede	Função	Tamanho da saída gerada
1	Camada de entrada e redimensionamento	Redimensionar os canais de cores de 0 a 255 para 0 a 1	(512, 512, 3)
2	Camada de inversão	Inversão aleatória (horizontal/vertical)	(512, 512, 3)
3	Camada de rotação	Rotação de 0.2 com sentido aleatório	(512, 512, 3)
4	Camada de contraste	Contraste aleatório com fator entre 0.3 a 0.5	(512, 512, 3)

Fonte: Elaborado pelo autor.

Rede neural classificadora de imagens de TC

Estrutura da rede neural criada

	Componente da rede	Função	Tamanho da saída gerada
5	Camadas de convolução e agrupamento VGG-16	Transferência de aprendizado de uma CNN VGG (pesos fixos gerados a partir do conjunto *imagenet*)	(16, 16, 512)
6	Camada de achatamento	Redimensiona a saída para apenas um valor	(131072)
7	Camada densa de neurônios	64 neurônios com função de ativação ReLU	(64)
8	Camada densa de neurônios	32 neurônios com função de ativação ReLU	(32)
9	Camada densa de neurônios	Apenas 1 neurônio com função de ativação sigmoid	(1)

Fonte: Elaborado pelo autor.

Rede neural classificadora de imagens de TC

Treinamento da rede estruturada

- Compilação do modelo: cálculo do erro como entropia cruzada binária e taxa de aprendizado de 0.001
- Função *call-back* para interromper o treinamento se acurácia de validação >= 95% e AUC ROC de validação >= 0.9
- Utilizando o conjunto de dados de treino (65% do conjunto – 307 imagens)

Rede neural classificadora de imagens de TC

Treinamento da rede estruturada

- 12 épocas (iterações que percorrem todo o conjunto de dados)
- Tamanho do lote (batch size) de 32 imagens por iteração do treinamento de cada época
- As camadas da rede VGG-16 não tiveram seus pesos alterados durante o treinamento feito

5. Resultados

Classificador fatores de Risco

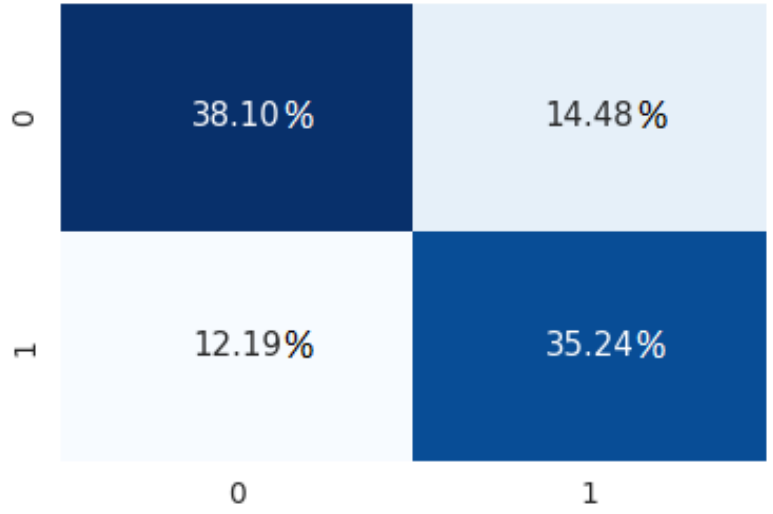
Acurácia:

Acurácia do classificador	Regressão Logística	Média: 73.33% Desvio padrão: 2.48%
	Floresta Aleatória	Média: 73.14% Desvio padrão: 1.40%

Fonte: Elaborado pelo autor.

Classificador fatores de Risco

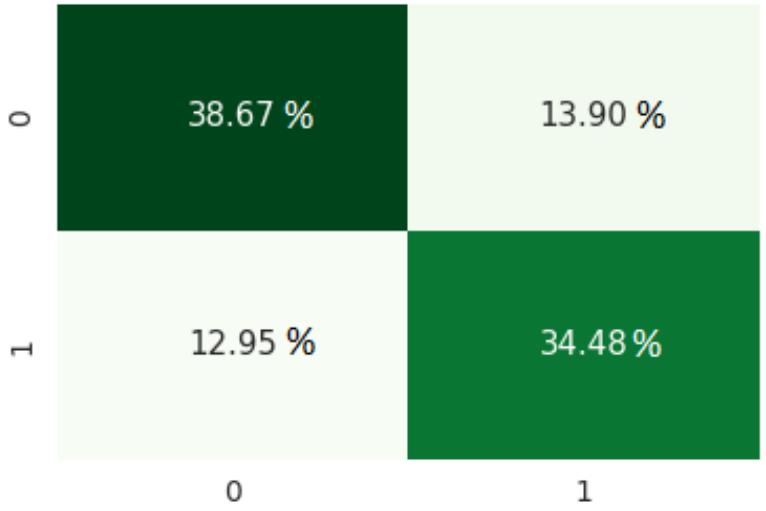
Matriz de Confusão - modelo de Regressão Logística:



Fonte: Elaborado pelo autor.

Classificador fatores de Risco

Matriz de Confusão - modelo de Floresta Aleatória:



Fonte: Elaborado pelo autor.

Classificador fatores de Risco

Taxa de Verdadeiro Positivo (*True Positive Rate* – TPR):

TPR	Regressão Logística	Média: 74.53% Desvio padrão: 4.10%
	Floresta Aleatória	Média: 72.95% Desvio padrão: 4.11%

Fonte: Elaborado pelo autor.

Classificador fatores de Risco

Taxa de Falso Positivo (*False Positive Rate* – FPR):

FPR	Regressão Logística	Média: 27.28% Desvio padrão: 5.27%
	Floresta Aleatória	Média: 26.23% Desvio padrão: 3.99%

Fonte: Elaborado pelo autor.

Classificador fatores de Risco

Área da Curva de Característica de Operação do Receptor (curva ROC):

AUC ROC	Regressão Logística	Média: 0.8029 Desvio padrão: 0.0506
	Floresta Aleatória	Média: 0.7982 Desvio padrão: 0.0349

Fonte: Elaborado pelo autor.

Rede neural classificadora de imagens de TC

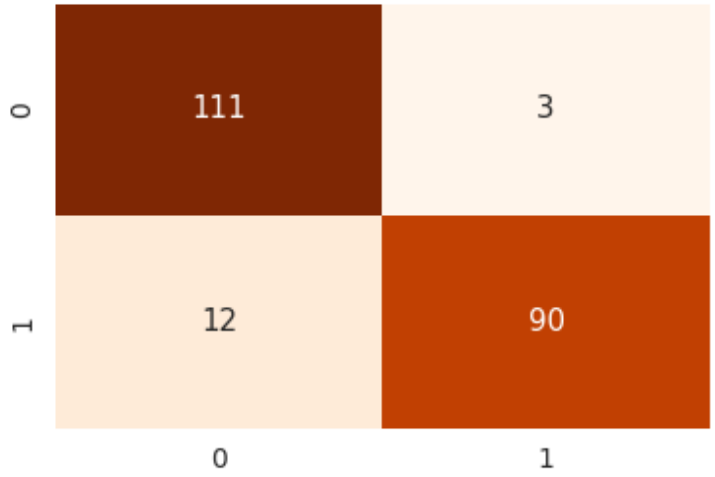
Acurácia:

Acurácia da Rede	Dados de treino	93.06%
	Dados de validação	97.39%

Fonte: Elaborado pelo autor.

Rede neural classificadora de imagens de TC

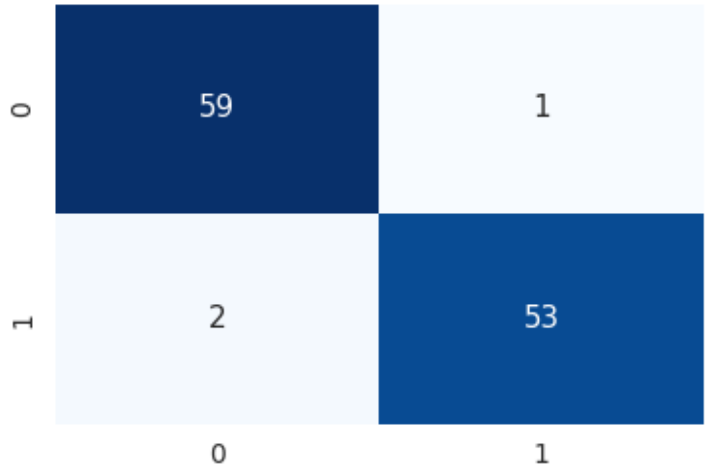
Matriz de confusão para dados de treino:



Fonte: Elaborado pelo autor.

Rede neural classificadora de imagens de TC

Matriz de confusão para dados de teste:



Fonte: Elaborado pelo autor.

Rede neural classificadora de imagens de TC

Taxa de Verdadeiro Positivo (*True Positive Rate* – TPR):

TPR	Dados de treino	88.24%
	Dados de validação	96.36%

Fonte: Elaborado pelo autor.

Rede neural classificadora de imagens de TC

Taxa de Falso Positivo (*False Positive Rate* – FPR):

FPR	Dados de treino	2.63%
	Dados de validação	1.67%

Fonte: Elaborado pelo autor.

Rede neural classificadora de imagens de TC

Área da Curva de Característica de Operação do Receptor (curva ROC):

AUC ROC	Dados de treino	0.9802
	Dados de validação	0.9944

Fonte: Elaborado pelo autor.

6. Conclusão

Conclusão

- Técnicas da **Inteligência Artificial** foram utilizadas (das subáreas **Aprendizagem de Máquina** e **Aprendizagem Profunda**)

- Modelos tiveram desempenhos satisfatórios

Conclusão

- Os classificadores desenvolvidos **são apenas demonstrações** de formas de diagnóstico da doença abordada
  - Fins demonstrativos e não soluções do problema em si

- AVC é ligado diretamente a questões de medicina e saúde
  - Os classificadores criados **não são soluções para autodiagnostico** da doença e nem são substitutos de profissionais da saúde
  - Em caso de qualquer suspeita procurar assistência médica e não negligenciar a situação para minimizar problemas e prejuízos

Referências Bibliográficas

Referências Bibliográficas

NURFIKRI, F. An Illustrated Guide to Artificial Neural Networks. 2020. Disponível em: https://towardsdatascience.com/an-illustrated-guide-to-artificial-neural-networks-f149a549ba74. Acesso em: 28 novembro 2022.

FERGUSON, M.; AK, R.; LEE, Y.-T.; LAW, K. Automatic localization of casting defects with convolutional neural networks. In: . [S.l.: s.n.], 2017. p. 1726–1735.

Obrigado pela atenção!

Vinícius de Paula Pilan (RA: 191025399)

Orientador: Prof. Dr. Clayton Reginaldo Pereira