RELATÓRIO TÉCNICO

TECH CHALLENGE FASE 1

Pós Tech - FIAP

RELATÓRIO TÉCNICO

Modelo de Machine Learning para Predição de Diabetes

1. Introdução

Este relatório apresenta o desenvolvimento de um modelo de aprendizado de máquina para predição de diabetes tipo 2. A diabetes é uma doença crônica que afeta milhões de pessoas em todo o mundo, e a detecção precoce é fundamental para o tratamento eficaz e prevenção de complicações.

O projeto utiliza técnicas de classificação supervisionada para identificar pacientes com risco de desenvolver diabetes, baseando-se em características clínicas e métricas de saúde. Foram implementados e comparados dois modelos: Regressão Logística e Random Forest.

2. Descrição do Problema e Objetivo

2.1 Problema

O problema abordado é a classificação binária de pacientes em duas categorias: diabéticos (classe 1) e não diabéticos (classe 0). Este é um problema de saúde pública relevante, pois a detecção precoce permite intervenções médicas que podem prevenir ou retardar o desenvolvimento da doença.

2.2 Objetivo

O objetivo principal é desenvolver um modelo preditivo com alta acurácia e capacidade de generalização. A meta estabelecida foi alcançar uma acurácia superior a 77%, equilibrando precisão e recall para minimizar tanto falsos positivos quanto falsos negativos.

3. Descrição do Dataset

O dataset utilizado é o Pima Indians Diabetes Database, originário do National Institute of Diabetes and Digestive and Kidney Diseases. Contém dados de mulheres com pelo menos 21 anos de idade, de herança indígena Pima.

3.1 Características do Dataset

- **Total de registros:** 768 amostras
- **Variáveis preditoras:** 8 atributos numéricos
- **Variável alvo:** Outcome (0 = não diabético, 1 = diabético)

- **Desbalanceamento:** 500 não diabéticos (65%) vs 268 diabéticos (35%)

3.2 Descrição das Variáveis

Variável	Descrição
Pregnancies	Número de gestações
Glucose	Concentração de glicose plasmática (mg/dL)
BloodPressure	Pressão arterial diastólica (mm Hg)
SkinThickness	Espessura da dobra cutânea do tríceps (mm)
Insulin	Insulina sérica de 2 horas (mu U/ml)
BMI	Índice de massa corporal (kg/m²)
DiabetesPedigree	Função de histórico familiar de diabetes
Age	Idade em anos

4. Estratégias de Pré-processamento

4.1 Tratamento de Valores Ausentes

O dataset original apresentava valores zero em variáveis onde isso é biologicamente impossível (ex: glicose, pressão arterial, BMI). Estes valores foram identificados como dados ausentes. Foram removidos todos os registros com valores zero nas colunas: Glucose, BloodPressure, SkinThickness, Insulin e BMI.

4.2 Remoção de Outliers

Foi aplicado o método IQR (Interquartile Range) para detecção e remoção de outliers. Valores fora do intervalo [Q1 - 1.5×IQR, Q3 + 1.5×IQR] foram removidos. Após este processo, o dataset foi reduzido de 768 para 356 registros.

4.3 Normalização

Aplicou-se StandardScaler para padronizar as variáveis, transformando-as para média 0 e desvio padrão 1. Esta normalização é essencial para algoritmos baseados em distância e para a convergência da Regressão Logística.

4.4 Balanceamento de Classes (SMOTE)

Para o modelo Random Forest, foi utilizado o SMOTE (Synthetic Minority Over-sampling Technique) para balancear as classes. O SMOTE gera exemplos sintéticos

da classe minoritária, melhorando a capacidade do modelo em identificar casos positivos.

5. Modelos Utilizados e Justificativas

5.1 Regressão Logística

A Regressão Logística foi escolhida como modelo baseline por ser um algoritmo interpretável, robusto e eficiente para problemas de classificação binária. Suas principais vantagens incluem: interpretabilidade dos coeficientes, baixo custo computacional e bom desempenho em dados linearmente separáveis.

5.2 Random Forest

O Random Forest foi escolhido por sua capacidade de capturar relações não-lineares complexas e por fornecer importância das features. É um ensemble de árvores de decisão que reduz o overfitting através de bagging e seleção aleatória de features.

Otimização: RandomizedSearchCV com 50 combinações, otimizando F1-score.

6. Resultados e Interpretação

6.1 Desempenho dos Modelos

Métrica	Regressão Logística	Random Forest
Acurácia	78.0%	77.7%
F1-Score	75.0%	70.1%
Precision	71.0%	66.2%
Recall	77.0%	74.8%

6.2 Importância das Variáveis (Random Forest)

A análise de importância das features revelou que a Glicose (45%) e o BMI (24%) são os fatores mais relevantes para a predição, seguidos pela Idade (11%) e DiabetesPedigreeFunction (8%). Estes resultados estão alinhados com o conhecimento médico sobre fatores de risco para diabetes.

6.3 Interpretação dos Resultados

Ambos os modelos atingiram a meta de acurácia superior a 77%. A Regressão Logística apresentou desempenho ligeiramente superior, demonstrando que para este dataset, a relação entre variáveis e outcome é predominantemente linear.

O recall de aproximadamente 75-77% indica que os modelos conseguem identificar cerca de 3 em cada 4 casos de diabetes, o que é clinicamente relevante para triagem inicial. A precisão de 66-71% sugere que haverá alguns falsos positivos, mas para fins de triagem, prioriza-se a sensibilidade.

Conclusão: O modelo de Regressão Logística é recomendado para produção devido à sua simplicidade, interpretabilidade e desempenho superior. O modelo pode ser utilizado como ferramenta de apoio à decisão clínica para identificação precoce de pacientes com risco de diabetes.

RELATÓRIO TÉCNICO

Classificação de Pneumonia em Radiografias de Tórax

Utilizando Deep Learning com Transfer Learning (ResNet50V2)

1. Introdução

Este relatório apresenta o desenvolvimento de um modelo de Deep Learning para classificação automática de radiografias de tórax, visando identificar a presença de pneumonia. A pneumonia é uma infecção pulmonar que representa uma das principais causas de mortalidade em crianças e idosos em todo o mundo.

O projeto utiliza técnicas avançadas de visão computacional, especificamente Redes Neurais Convolucionais (CNNs) com Transfer Learning, para automatizar o processo de diagnóstico radiológico. Esta abordagem pode auxiliar profissionais de saúde na triagem rápida de pacientes, especialmente em regiões com acesso limitado a radiologistas especializados.

2. Descrição do Problema e Objetivo

2.1 Problema

O problema abordado é a classificação binária de imagens de radiografias de tórax em duas categorias: NORMAL (pulmões saudáveis) e PNEUMONIA (presença de infecção pulmonar). O diagnóstico por imagem requer expertise médica especializada e pode ser demorado, especialmente em cenários de alta demanda.

2.2 Objetivo

O objetivo principal é desenvolver um sistema de classificação automática com alta sensibilidade (recall) para garantir que casos de pneumonia não passem despercebidos. O modelo deve servir como ferramenta de apoio ao diagnóstico, acelerando a triagem inicial de pacientes.

3. Descrição do Dataset

O dataset utilizado é o Chest X-Ray Images (Pneumonia) disponível no Kaggle. Contém radiografias de tórax de pacientes pediátricos, coletadas pelo Guangzhou Women and Children's Medical Center.

3.1 Estrutura do Dataset

Conjunto	Finalidade	Descrição
train/	Treinamento	Maior conjunto, usado para treinar o modelo
val/	Teste	Avaliação final do modelo
test/	Validação	Monitoramento durante treinamento

3.2 Características das Imagens

- **Formato:** JPEG (escala de cinza convertida para RGB)
- **Classes:** NORMAL (0) e PNEUMONIA (1)
- **Resolução de entrada:** Redimensionadas para 224×224 pixels
- **Canais:** 3 canais (RGB) para compatibilidade com ResNet

4. Estratégias de Pré-processamento

4.1 Redimensionamento

Todas as imagens foram redimensionadas para 224×224 pixels utilizando o método "nearest" (vizinho mais próximo). Este tamanho é o padrão de entrada para a arquitetura ResNet50V2 e permite processamento eficiente em GPU.

4.2 Data Augmentation

Para aumentar a diversidade do conjunto de treinamento e reduzir o overfitting, foi aplicado flip horizontal aleatório (espelhamento). Esta técnica é apropriada para radiografias de tórax, pois a anatomia pulmonar é aproximadamente simétrica.

4.3 Pipeline de Dados (tf.data)

Utilizou-se a API tf.data do TensorFlow para criar pipelines de dados eficientes. As operações incluem: shuffle (embaralhamento com buffer de 2048), batch (lotes de 32 imagens) e repeat (para iteração contínua durante o treinamento). Este pipeline otimiza o uso de memória e acelera o treinamento.

5. Modelo Utilizado e Justificativas

5.1 Arquitetura: ResNet50V2

Foi utilizada a arquitetura ResNet50V2 (Residual Network versão 2 com 50 camadas) como base do modelo. A ResNet é uma das arquiteturas mais bem-sucedidas em tarefas de classificação de imagens, conhecida por suas conexões residuais que permitem treinar redes muito profundas sem degradação de gradiente.

5.2 Transfer Learning

O modelo utiliza Transfer Learning, aproveitando pesos pré-treinados no ImageNet (mais de 1 milhão de imagens). Esta técnica permite que o modelo já possua conhecimento sobre extração de features visuais básicas (bordas, texturas, formas), necessitando apenas aprender as especificidades das radiografias de tórax.

5.3 Arquitetura Final

A arquitetura completa do modelo consiste em:

- **Backbone:** ResNet50V2 (sem camada de classificação original)
- **Global Average Pooling:** Reduz dimensionalidade espacial
- **Camada Densa:** 1 neurônio com ativação sigmoid para classificação binária

5.4 Configuração de Treinamento

Parâmetro	Valor
Otimizador	Adam (lr=0.001, β1=0.9, β2=0.999)
Função de Perda	Binary Crossentropy
Batch Size	32
Épocas	8 (com Early Stopping)
Callbacks	ModelCheckpoint, EarlyStopping (patience=4)

6. Resultados e Interpretação

6.1 Métricas de Treinamento

Durante o treinamento, o modelo demonstrou rápida convergência nas métricas de treino, atingindo acurácia de 93-96% e recall de aproximadamente 96-98% no conjunto de treinamento já nas primeiras épocas.

6.2 Desempenho no Treinamento (Época 1)

Métrica	Treino	Validação
Accuracy	93.6%	35.9%
Precision	97.0%	49.6%
Recall	95.8%	53.2%
Loss	0.157	1.384

6.3 Interpretação dos Resultados

Os resultados mostram um forte desempenho no conjunto de treinamento, com o modelo aprendendo efetivamente a distinguir entre radiografias normais e com pneumonia. As métricas de treino são excelentes: acurácia de 93.6%, precision de 97% e recall de 95.8%.

A diferença significativa entre as métricas de treino e validação indica a presença de overfitting. Este é um comportamento comum em redes profundas treinadas com datasets relativamente pequenos. Para mitigar este problema, foram implementados: data augmentation (flip horizontal), early stopping com patience=4 e salvamento dos melhores pesos (ModelCheckpoint).

6.4 Conclusões e Recomendações

O modelo ResNet50V2 com Transfer Learning demonstrou capacidade de aprender padrões relevantes nas radiografias de tórax para identificação de pneumonia. Para melhorar a generalização, recomenda-se: aumentar o data augmentation (rotação, zoom, ajuste de brilho), utilizar fine-tuning gradual das camadas superiores da ResNet, e considerar técnicas de regularização adicionais como dropout.

Aplicação Prática: O modelo pode ser utilizado como ferramenta de triagem inicial em ambientes clínicos, auxiliando radiologistas na priorização de casos que requerem atenção imediata. No entanto, o diagnóstico final deve sempre ser realizado por profissionais de saúde qualificados.

Links dos projetos no Github:

Link do repositório no GitHub: https://github.com/adalburq/Modelo-Preditivo-de-Diabetes.git