Técnicas de Inteligência Artificial para diagnóstico de acidente vascular cerebral através de imagens e dados textuais sobre possíveis vítimas

Nome: Vinícius de Paula Pilan

RA: 191025399

Resumo – sobre o projeto

- **Problema abordado:** quanto mais tardio é o diagnóstico de Acidente vascular cerebral (AVC), pior são os prejuízos para a vítima
- Criar um classificador de dados e de imagens sobre AVC com intuito de agilizar diagnósticos da doença
- Ao total serão desenvolvidos dois modelos:
 - **1. Classificador de dados:** recebe informações sobre determinado indivíduo e o classifica como possível vítima ou não
 - **2. Classificador de imagens (rede neural):** recebe imagens de radiografia sobre um indivíduo e o classifica como possível vítima ou não

Cronograma atualizado

Desenvolvimento até o momento

Base de dados com informações sobre vítimas

- Stroke Prediction Dataset
 - 12 diferentes características e 5110 entradas
- Informações presentes no conjunto:
 - 1. id: identificador único
 - 2. gender: sexo
 - 3. age: idade
 - 4. hypertension: indica se o paciente tem hipertensão
 - 5. heart_disease: indica se o paciente tem alguma doença cardíaca
 - 6. ever married: indica se o paciente é casado
 - 7. work_type: indica se o paciente trabalha e, se sim, qual o tipo de emprego
 - 8. Residence_type: tipo de residencia, rural ou urbana
 - 9. avg_glucose_level: media do nível de glicose no sangue do paciente
 - **10. bmi:** índice de massa corporal (padrão americano)
 - 11. smoking_status: situação do paciente com relação a fumar
 - 12. stroke: indica se o paciente teve ou não avc

Classificador de dados

Fases da criação:

- 1. Preparação da base de dados
- 2. Modelagem
- 3. Avaliação dos resultados

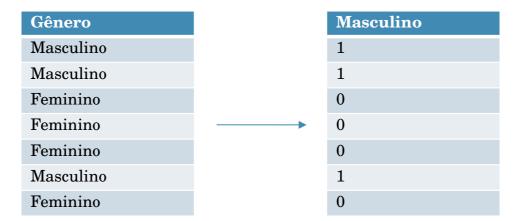
Preparação da base de dados – Balanceamento

- Distribuição original da variável alvo:
 - 249 casos para ocorrência de AVC (5%)
 - 4861 casos de não ocorrência de AVC (95%)

- Subamostragem do conjunto de dados da classe *não AVC (4861 → 251)*
 - total: $5110 \rightarrow 500$ elementos

Preparação da base de dados – Correção de formato

• Correção para variáveis de texto com apenas dois possíveis valores:



Nesses casos, para corrigir o formato dessas colunas para um formato numérico pode-se substituir um desses valores pelo dígito "1" e o outro pelo "0".

Preparação da base de dados – Correção de formato

• Correção para variáveis de texto com vários possíveis valores:

Tipo de emprego	Privado	Autônomo	Cargo público	Criança
Privado	1	0	0	0
Privado	1	0	0	0
Autônomo	0	1	0	0
Privado	 1	0	0	0
Criança	0	0	0	1
Cargo público	0	0	1	0
Autônomo	0	1	0	0

Nesses casos, para corrigir o formato dessas colunas para um formato numérico cria-se novas colunas binárias para cada um dos possíveis valores da coluna original.

Preparação da base de dados – Dados nulos

• Única coluna com dados nulos foi *bmi*:

Distribuição da variável BMI com relação a dados nulos					
	249 casos de AVC	209 valores não nulos (84%)			
	249 Casos de AvO	40 valores nulos (16%)			
Conjunto de dados total	251 casos de não AVC	245 valores não nulos (98%)			
	251 casos de nao AvC	6 valores nulos (2%)			

• Correção feita: substituição pela mediana

Preparação da base de dados – Normalização

- Normalização escolhida: *min-max*
 - redimensiona para o intervalo [0,1] ou [-1, 1]
 - lida melhor com dados de distribuição não normal

$$x_{scaled} = rac{x - x_{min}}{x_{max} - x_{min}}$$

Modelagem – Algoritmos utilizados

- Algoritmos de aprendizado supervisionado:
 - ✓ Máquina de vetor de suporte (SVM)
 - ✔ Floresta aleatória
- Treinamentos feitos para cada um desses dois com intuito de se escolher o que melhor soluciona o problema

Modelagem – Conjunto para treino e para teste

- Validação cruzada com **cinco** dobras diferentes:
 - 500 elementos totais → 100 elementos por dobra (escolhidos aleatoriamente)
- Uma dobra para teste e as demais para treino
 - 100 elementos para teste (20% dos dados totais)
 - 400 elementos para treino (80% dos dados totais)
- Cinco possibilidades de treinamentos e testagens diferentes

Avaliação dos resultados – Métricas escolhidas

- Métricas para avaliar classificação:
 - Precision
 - ✓ Recall
 - ✓ F1-score
 - ✓ AUC ROC score
- Taxa de falso positivo
- Taxa de falso negativo

O que falta ser feito

O que falta ser feito

- Métricas para avaliar classificação:
 - Precision
 - ✓ Recall
 - ✓ F1-score
 - ✓ AUC ROC score
- Taxa de falso positivo
- Taxa de falso negativo

Obrigado pela atenção!