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

Base de dados para criação do classificador

- 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

- Criar um classificador que indique caso de AVC ou não a partir de dados sobre indivíduos

Fases da criação:

1. Preparação dos dados
2. Modelagem
3. Avaliação dos resultados

Preparação dos dados

Balanceamento

- Distribuição original da variável alvo:
  - 249 casos para ocorrência de AVC (5%)
  - 4861 casos de não ocorrência de AVC (95%)
- **Problema:** sem balanceamento, classificador fica tendencioso

Balanceamento

1. Separação do conjunto total a partir das duas classes: *AVC* (249 elementos) e *não AVC* (4861 elementos)
2. Subamostragem do conjunto de dados da classe *não AVC (4861 → 251)*
3. Junção do conjunto de dados da classe *AVC* com o novo conjunto de dados da classe *não AVC* (total: 5110 → 500 elementos)

Correção de formato

- Maioria dos algoritmos de Machine Learning conseguem trabalhar apenas com atributos de formato numérico

- Necessário conversão dos dados para este formato

Correção de formato

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

Gênero
Masculino
Masculino
Feminino
Feminino
Feminino
Masculino
Feminino

Masculino
1
1
0
0
0
1
0

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”.

Correção de formato

- Correção para variáveis de texto com vários possíveis valores:

Tipo de emprego
Privado
Privado
Autônomo
Privado
Criança
Cargo público
Autônomo

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

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.

Tratamento de dados nulos

- Dados nulos não trazem informação
- Classificador não consegue interpretá-los
- **Correção:** substituição pela média, mediana... ou eliminação

Tratamento de dados nulos

- Única coluna com dados nulos foi *bmi*:

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

- Correção feita: substituição pela mediana

Normalização

- Mudar os valores das colunas numéricas para usar uma escala comum sem distorcer as diferenças nos intervalos de valores

- Necessária para alguns algoritmos modelarem os dados corretamente

Normalização

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

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Modelagem

Algoritmos utilizados

- Algoritmos de aprendizado supervisionado:
  - ✔ Máquina de vetor de suporte (SVM)
  - ✔ Floresta aleatória

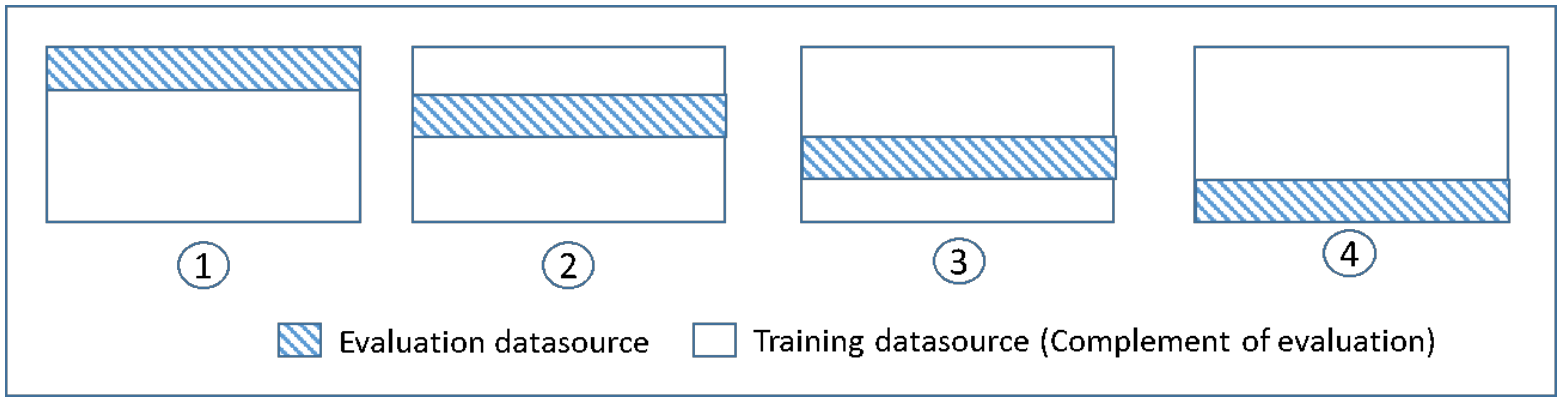
- Treinamentos feitos para cada um desses dois com intuito de se escolher o que melhor soluciona o problema

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

Separação treino e teste

- Exemplo de validação cruzada com **quatro dobras**:



Fonte: https://docs.aws.amazon.com/pt_br/machine-learning/latest/dg/cross-validation.html

Avaliação dos resultados

Métricas escolhidas

- Métricas para avaliar classificação:
  - ✔ Precision
  - ✔ Recall
  - ✔ F1-score
  - ✔ AUC ROC score

- Taxa de falso positivo
- Taxa de falso negativo

Obrigado pela atenção!