Previsão de casos de AVC

Diogo Carrilho - 127534

Francisco Rodrigues – 98830

Jorge Gomes - 104558

Rafael Machado - 87341

Domínio

O AVC é a segunda causa de morte e a terceira causa de incapacidade a nível global. Em Portugal, é igualmente uma das principais causas de mortalidade e morbilidade.

O nosso dataset tem como parâmetros:

- Id
- Gender (Male, Female)
- Age
- Hypertension (no- 0; yes – 1)
- Hearth_deseases (no- 0; yes– 1)
- Ever_married (no - false; yes – true)
- Work_type (never worked, children, private, govt_job, self_employed)
- Residence_type (rural; urban)
- Avg_glucose_lvl: nível médio de glucose
- bmi: Índice de massa muscular
- Smoking_status (never smoked, unknown, formerly smoked, smokes)

Fonte de dados

A fonte de dados usada parte do dataset 'Stroke Prediction Dataset' do Kaggle, a qual representa 11 características clínicas para prever eventos de AVC.

Aquisição dos Dados (API, Scraping ou Dataset)

A fonte de dados usada foi o Kaggle, com o dataset em csv presente no link https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data?select=healthcare-dataset-stroke-data.csv

Processamento dos Dados

O processamento de dados começa pela importação do dataset em csv e identificar as colunas e linhas (12 colunas, 5110 linhas).

Após isso começamos por retirar as linhas com valores NaN, resultando numa redução de cerca de 200 linhas.

Identificamos, na coluna smoking_status, um dos dados como unknown e optamos também por remover as linhas com esse valor por reconhecermos que prejudicaria o nosso modelo. Consequentemente, houve uma redução de mais 1483 linhas

No campo gender, removemos o valor other por apenas representar 1 linha e, não tendo um AVC, ser redundante.

Segue-se a isso a redução para os atributos relevantes. Visto que o dataset incluía uma coluna de id não relevante, esta foi retirada, resultando num total de 11 colunas a utilizar no modelo preditivo.

Decidimos manter as outras colunas por terem valores relevantes e também porque seria possível observar correlações interessantes nomeadamente em relação ao tipo de trabalho ou zona de residência.

Criação do Modelo Preditivo

O modelo criado tem como objetivo prever a ocorrência de AVCs tendo como base as restantes variáveis apresentadas no dataset.

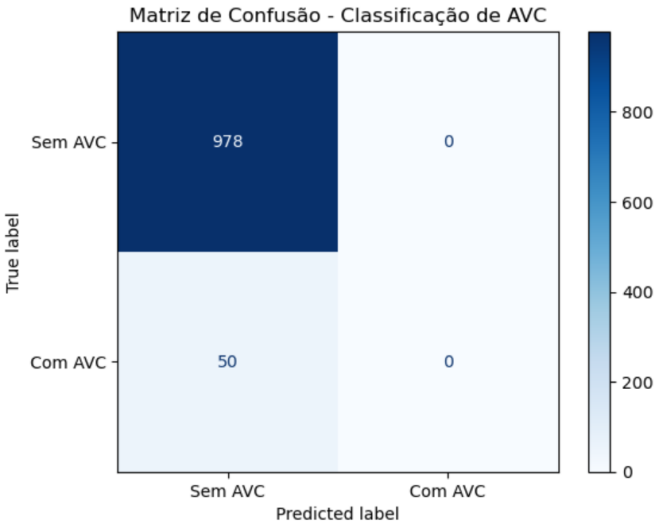
A implementação do modelo começou pelo tratamento dos dados de texto para as colunas categóricas, utilizando a função Ordinal Encoder do scikit-learn, que permite converter os dados em texto para valores numéricos.

Após isto, realizámos a divisão entre os dados de previsão (X) e os dados a prever (y), correspondentes a um DataFrame (com todas as colunas exceto ‘stroke’) e uma Series com a coluna ‘stroke’.

Após isto realizámos a divisão de dados em conjunto de teste e treino, com uma divisão de 30% teste e 70% treino, uma proporção comum na utilização de modelos de previsão, fazendo depois o fit com x_train e y_train e o predict com o x_test.

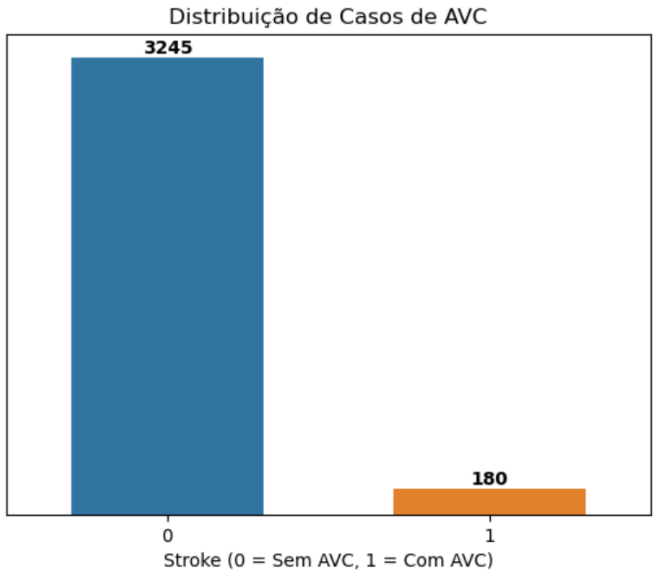
De forma a comparar a usabilidade dos dados calculámos a accuracy entre os resultados y obtidos do x_test com o y_test, com uma accuracy de 95%.

No entanto, esta accuracy não resulta numa previsão realmente correta, como podemos ver ao realizar a matriz de confusão.

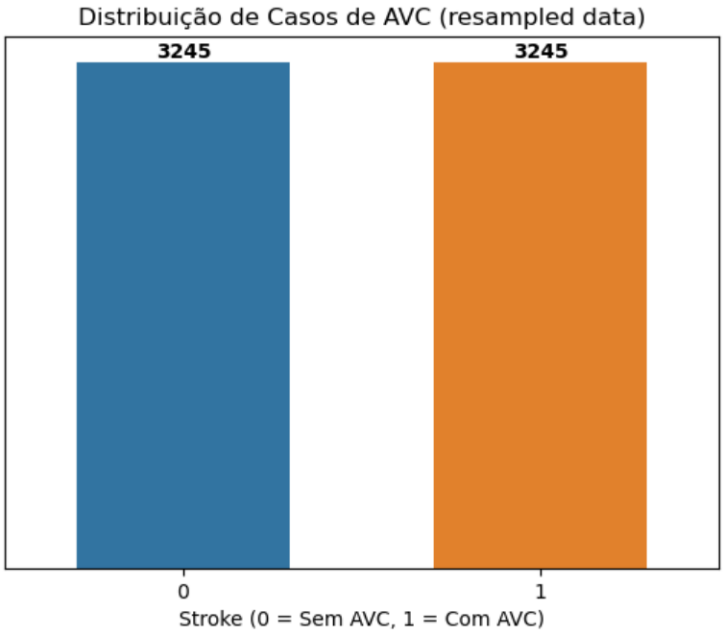


Ao realizarmos o modelo, o modelo estava erradamente a prever para todos os pacientes, que não existiria AVC, no entanto estávamos com um erro de falso negativo em 50 dos casos.

Este erro decorreu principalmente da discrepância de volume de dados entre os strokes = 0 e strokes = 1 dos dados originais.

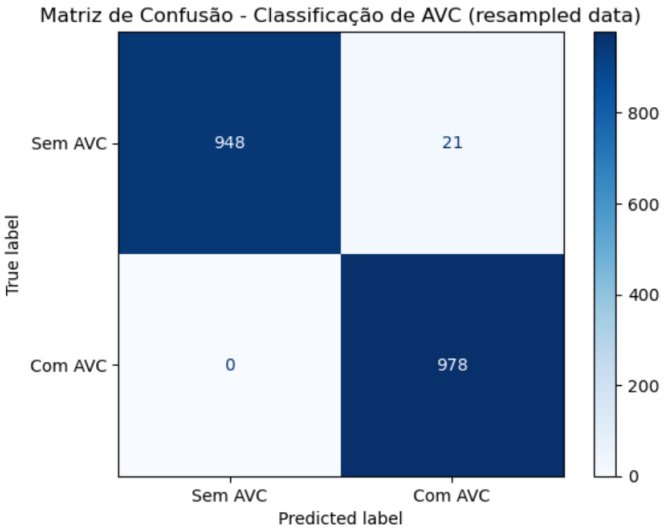


Para tentarmos obter um modelo mais útil, recorremos ao oversampling de casos de AVC. Optamos pela duplicação pura das linhas ao invés da técnica SMOTE por termos colunas como work_type ou smoking_status que, fazendo aproximações, prejudicava a sua interpretação e correlação com os casos de AVC. Para o oversampling utilizámos a função resample do scikit-learn, que permitiu garantir uma divisão igual entre linhas correspondentes a stroke = 0 e stroke = 1.

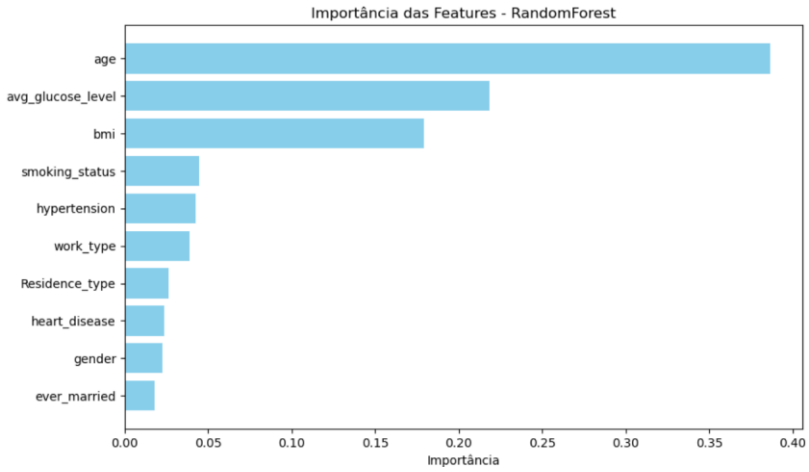


Visualização dos Dados

Após o resampling dos dados realizámos novamente uma divisão entre dados de teste e treino, corremos o modelo e obtivemos uma maior accuracy, de 98,92% e uma matriz de confusão a indicar a inexistência de falsos negativos, apenas falsos positivos, que seria o melhor dos dois cenários de erro (tendo em conta que pretendemos prever a classe menor (stroke = 1). A matriz de confusão torna-se assim uma das visualizações mais úteis para compreender em que tipos de casos o nosso modelo erra (falsos negativos ou falsos positivos).



Realizámos ainda uma visualização das features mais relevantes para a previsão de stroke, que incluía em primeiro lugar age, segundo lugar avg_glucose_level e terceiro o bmi.



Conclusão

Escolhemos o *dataset 'Stroke Prediction Dataset'* do Kaggle porque pareceu-nos um bom equilíbrio entre simples e desafiante.
No processamento de dados encontramos alguns valores a NaN e unknown pelo que decidimos simplesmente removê-las. Removemos também a coluna Id por não ser relevante.
Na construção do modelo preditivo:

- Tivemos um problema relativamente à accuracy do modelo preditivo onde previa corretamente quem não tinha avc mas incorretamente quem tinha. Para tentar corrigir este problema, decidimos aumentar o número de casos de avc, copiando essas mesmas linhas x vezes até obtermos o mesmo número de casos de stroke e sem stroke.