Projeto de Machine Learning

META 1

Rodrigo Costa | 2020133365
Vasco Oliveira | 2019119011

Conteúdo

Introdução

Neste projeto iremos explorar um Dataset de histórico médico para previsão de ocorrências de aneurismas. O dataset em estudo [1] é utilizado para prever se um paciente tem possibilidade de ter um aneurisma baseado em parâmetros como género, idade, várias doenças e se fuma. Cada linha fornece informação relevante de um paciente que foi recolhida de forma anónima nos Estados Unidos.

Dataset Features

Neste capítulo temos as features que iremos utilizar como valores de entrada, assim como uma resumida explicação de cada uma delas.

- **id**: id de paciente.
- **gender**: "Male", "Female" ou "Other" para distinguir sexo do paciente.
- **age**: idade do paciente.
- **hypertension**: 0 se o paciente não tem hipertensão, 1 se o tem.
- **heart_disease**: 0 se o paciente não possuir doenças de coração, 1 se o paciente tiver alguma.
- **ever_married**: ‘Yes’ para se já esteve casado, ‘No’ para não.
- **work_type**: "children", "Govt_jov", "Never_worked", "Private" ou "Self-employed" para determinar o estado de trabalho do paciente.
- **Residence_type**: "Rural" ou "Urban" para determinar em que tipo de ambiente o paciente vive.
- **avg_glucose_level**: media de glucose presente no sangue do paciente.
- **bmi**: índice de massa corporal do paciente.
- **smoking_status**: "formerly smoked", "never smoked", "smokes" ou "Unknown"* para determinar se o paciente fuma e com que regularidade.
- **stroke**: 1 se o paciente teve um aneurisma ou 0 se não.

Medidas Tomadas Na Limpeza do Dataset

Neste capítulo, iremos descrever passo a passo as medidas tomadas ao longo da limpeza do dataset, assim como a lógica por trás de cada um destes passos quando necessário.

1. Carregamento do dataset ‘healthcare-dataset-stroke-data.csv’ para dataframe utilizando pandas.
2. Análise inicial da estrutura do dataset através de data.info() e data.count(). O dataset possui 5110 linhas inicialmente.
3. Remoção das colunas sem valor, no que notamos que falta bmi em 201 linhas, dado o tamanho do dataset, a perda destes valores equivale a 3.93% da dataframe inicial, pelo que é justificável a remoção das linhas com valores em falta.

4. Verificação de valores repetidos através de data.duplicated().any(). Não existe nenhum valor repetido.
5. Capitalização dos nomes das features e normalização de valores em string para letra minúscula.
6. Remoção da feature ‘id’ pois não contribui em nada para o objetivo de detetar possibilidade de aneurisma, dado que é só um índice do dataset.
7. Remoção de linhas cujo ‘Smoking_status’ que são ‘unknown’. Esta é uma remoção muito significativa, pois equivale a 29.02% das linhas do dataset. No entanto, as restantes 3426 linhas chegam para treinar e testar o nosso algoritmo. A decisão da remoção destes dados atribui-se aos vários estudos que comprovam uma forte correlação entre fumar e aneurismas, podendo uma pessoa que fume ter um risco cerca de seis vezes maior de ter um aneurisma que uma pessoa que não fume [2]. Isto ocorre devido ao dano cardiovascular que fumar causa no sistema assim como deposito de resíduos nas artérias [3].
8. Remoção de género ‘other’ dado que é apenas uma instância no dataset e é necessária normalização dentro das features.
9. Encoding das features categóricas e ordinais para valores numéricos. Sendo ‘Smoking_status’ uma feature ordinal, 0 para ‘never smoked’, 1 para ‘formely_smoked’ e 2 para ‘smokes’. A feature ‘Gender’ é categórica logo assumimos 0 para ‘male’ e 1 para ‘female’, assim como ‘Ever_married’ 0 para ‘no’ e 1 para ‘yes’.
10. Foi utilizado *One Hot Encoding* às features 'Residence_type' e 'Work_type', ficando respeticavaente para profissão ‘Work_type_private','Work_type_self-employed','Work_type_govt_job','Work_type_children','Work_type_never_worked' e para o área de residencia 'Residence_type_rural' e 'Residence_type_urban'.
11. Foi feita uma pipeline com os encoders criados no passo anterior de forma a aplicar ao dataframe. Valores de *one hot encoding* estavam em formato de double, tendo sido convertidos para inteiros dado de forma a normalizar os dados. Depois deste passo, foram feitas várias visualizações dos processamentos efetuados de forma a verificar os resultados.

Visualização do Dataframe

Na Figura 1 podemos observar vários detalhes das features depois de serem limpas, só que nos consegue dar uma visão geral dos dados de cada feature de forma individual.

Podemos observar pela Figura 2 nos dados existem mais homens com doenças cardíacas do que mulheres, no entanto há mais mulheres com hipertensão, dois dos contribuidores no que toca a aneurismas [4].

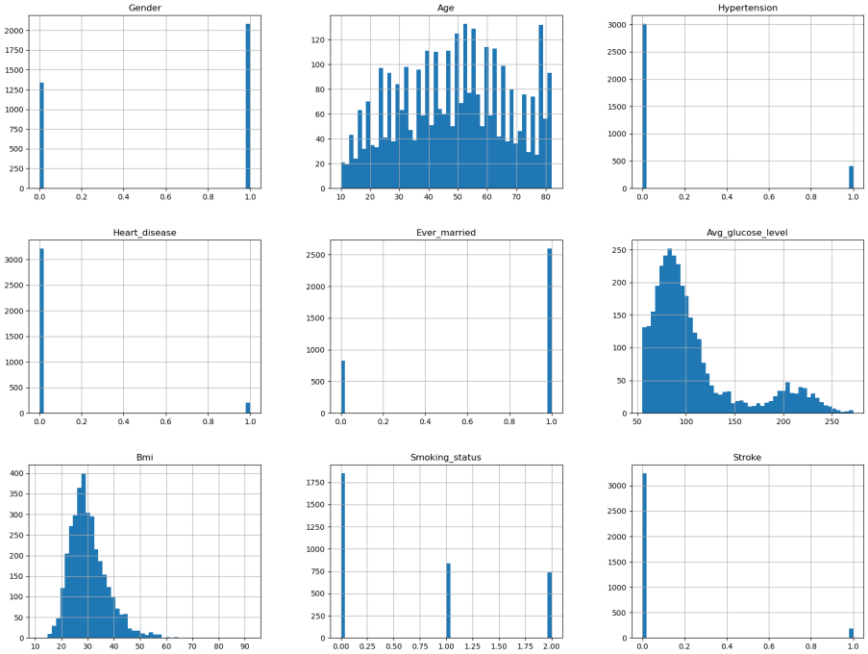

Figura 1 Gráfico de features tratadas

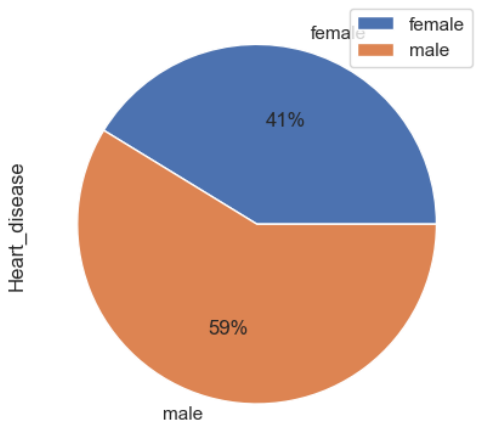

Figura 2 Doença cardiovascular por género

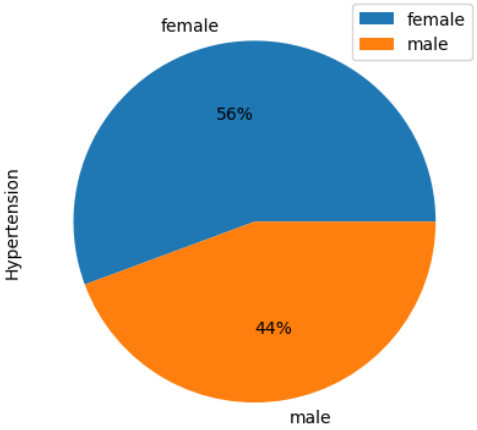

Figura 3 Hipertensão por género

Na Figura 4 podemos observar que nos dados que temos, pessoas que fumaram ou costumavam fumar tem uma maior frequência de doenças cardiovasculares.

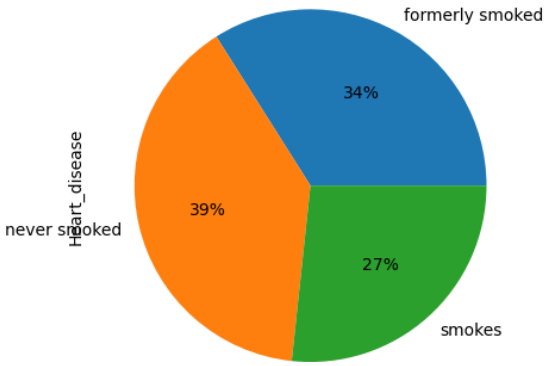

Figura 4 Doença cardiovascular por smoking_status

Na Figura 5 podemos ver a frequência de aneurismas conforme a idade e o género, podendo notar que o pico da incidência de aneurismas devido á idade avançada se encontra entre os 70-80 anos.

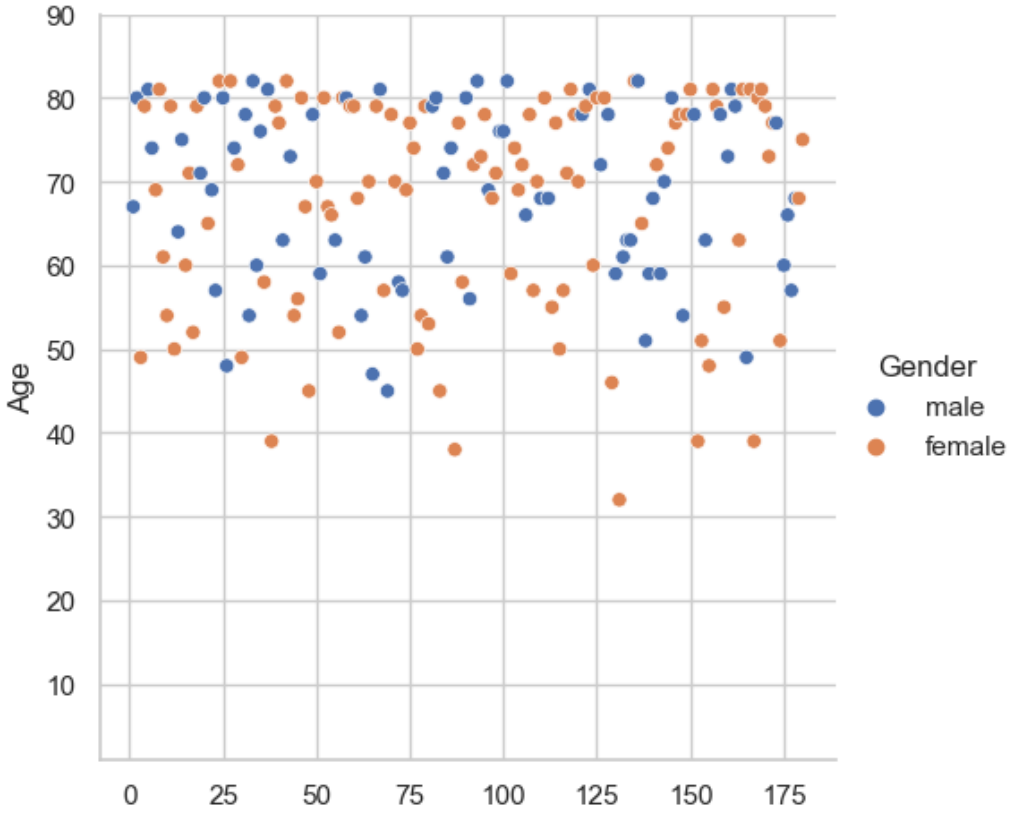

Figura 5 Instâncias de Aneurisma por idade e género

Na Figura 6 podemos observar que ocorrem mais aneurismas em pessoas com o bmi que indica obesidade (25-30), isto devido aos problemas cardiovasculares associados com excesso de peso e falta de exercício físico.

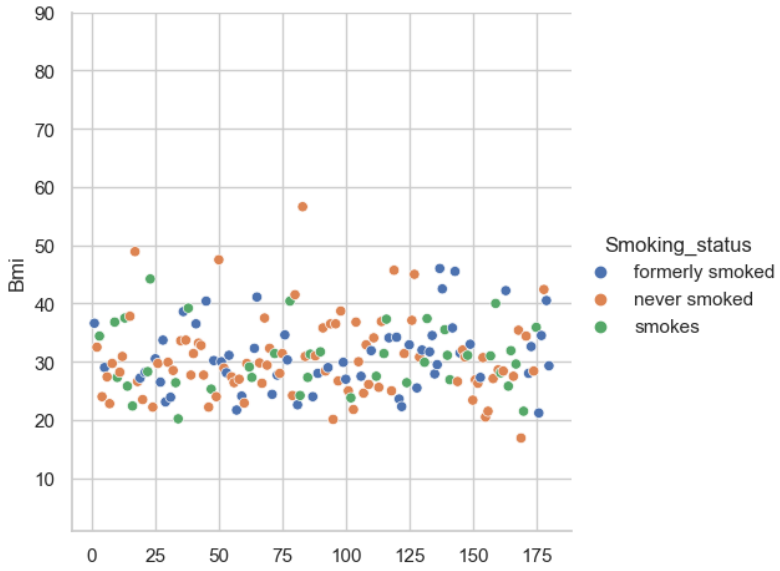

Figura 6 Instâncias de aneurisma por Bmi e Smoking_status

Conclusão

Ao fazer a análise dos dados que temos foi necessário estar a procurar por informação relativa ao caso de estudo do dataset. Desta forma abrangendo o trabalho de uma forma mais contextualizada, ao ter de analisar cada feature de forma critica de modo a saber o seu futuro valor para a futura implementação num algoritmo de classificação binária.

Foi necessário fazer alguns ‘cortes’ no que toca á informação inicial, perdendo uma percentagem significativa de dados. No entanto sentimos que continuamos com dados os suficientes e agora tratados, com o suficiente valor para facilitar os passos futuros deste projeto.

Referencias

[1]https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[2]https://www.medicalnewstoday.com/articles/can-smoking-cause-a-stroke

[3]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6708836/

[4]https://www.msdmanuals.com/pt-pt/profissional/doen%C3%A7as-cardiovasculares/doen%C3%A7as-da-aorta-e-seus-ramos/aneurismas-da-aorta-abdominal-aaas