Universidade do Minho
Escola de Engenharia

Universidade do Minho

Aprendizagem Automática II

Stroke Prediction

Autores:
Luís Freitas, PG38347
João Silva, PG42834
Guilherme Palumbo, PG42832
Luís Maia, A84241

Grupo:
8

6 de junho de 2021

Conteúdo

Lista de Figuras

1 Introdução

1.1 Identificação do Projeto e Objetivos

No âmbito da Unidade Curricular Aprendizagem Automática II foi nos proposto a realização de um projeto, onde deveríamos selecionar um conjunto de dados com uma dimensão razoável. O principal objetivo do projeto é a partir do conjunto de dados selecionado desenvolver algumas tarefas de *Data Mining*, nomeadamente:

- carregamento e exploração inicial (incluindo sumarização, visualização e pré-processamento);
- preparação dos dados;
- seleção de atributos;
- aprendizagem não supervisionada (redução de dimensionalidade, clustering);
- aprendizagem supervisionada (usando modelos "tradicionais";
- modelos Deep Learning;
- análise de resultados (comparar os resultados entre os modelos "tradicionais"e modelos Deep Learning).

1.2 Metodologia de Trabalho

Para a realização deste trabalho, utilizamos como metodologia de trabalho, a Metodologia CRISP-DM, uma vez que é uma metodologia que se baseia na experiência prática de como as pessoas desenvolvem os projetos de *Data Mining*. Esta metodologia é constituída por 6 fases, sendo estas:

1. ***Business Understanding*** - identificar e compreender o problema que precisa de ser resolvido, ou seja, converter os objetivos e requisitos de negócio num problema de *Data Mining*;
2. ***Data Understanding*** - compreensão dos dados, onde procuramos analisar os mesmos de forma a obter o máximo de informação;
3. ***Data Preparation*** - conjunto de tarefas de inspeção e prepração dos dados com o objetivo de se obterem os dados finais, para proceder à criação e validação dos modelos (remover outliers, remover atributos, substituir valores em falta);
4. ***Modeling*** - selecionar e aplicar os modelos tendo em conta os objetivos definidos. Nesta etapa, aplicam-se os algoritmos, capazes de produzir resultados satisfatórios sobre o conjunto de dados preparado na fase anterior.

5. ***Evaluation*** - avaliação do desempenho dos modelos desenvolvidos na fase anterior, tendo em conta as métricas de avaliação pré-definidas;

6. ***Deployment*** - aplicação do(s) modelo(s) no processo da tomada de decisão.

1.3 Estrutura do Relatório

Para além deste capítulo, e tendo em conta os objetivos e a metodologia de trabalho, este relatório está organizado em capítulos, são eles:

Capitulo 2: Descrição e Análise de Dados - neste capítulo iremos fazer uma descrição detalhada do nosso conjunto de dados, onde apresentamos alguns gráficos e dados estatísticos;

Capítulo 3: Tratamento dos Dados - neste capítulo iremos preparar os dados, de forma a termos o melhor conjunto de dados para o desenvolvimento dos modelos;

Capítulo 4: Modelos não Supervisionados - neste capítulo iremos apresentar os modelos não supervisionados desenvolvido e alguns dos seus resultados;

Capítulo 5: Seleção de Atributos - neste capítulo iremos fazer alguns testes estatísticos para entender quais os atributos mais importantes para os objetivos do problema;

Capítulo 6: Modelos Supervisionado - neste capítulo iremos apresentar os modelos supervisionados desenvolvidos, nomeadamente os modelos "Tradicionais"e modelos Deep Learning;

Capítulo 7: Conclusão - neste capítulo iremos apresentar os principais resultados e tirar algumas conclusões sobre os mesmos.

2 Descrição e Análise de Dados

O conjunto de dados que selecionamos para atingir os objetivos deste projeto foi o *Stroke Prediction Dataset*, retirado de: `https://www.kaggle.com/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-strokedata.csv&fbclid=IwAR363kTHoqeXDIqAZ7tR7OJuneTCUyPp2emTecW75eOtICDN0HqcomgOpQo`. O principal objetivo deste conjunto de dados é prever se um paciente teve um AVC. Para isso temos 5110 registos caracterizados por 12 atributos, sendo estes:

1. ***id:*** identificador único;
2. ***Gender:*** género do paciente - *Male* (Masculino), *Female* (Feminino) ou *Other* (outro);
3. ***Age:*** idade do paciente
4. ***Hypertension:*** 0 se o paciente não tem hipertensão, 1 se o paciente tem hipertensão;
5. ***Heart_ Disease:*** 0 se o paciente não tem nenhuma doença cardíaca, 1 se o paciente tem alguma doença cardíaca;
6. ***Ever_ married:*** *No* se não for casada ou *Yes* se for casada;
7. ***Work_ type:*** tipo de trabalho do paciente - *children* (criança), *Govtjov* (trabalhador do governo), *Never_ worked* (nunca trabalhou), *Private* (trabalhador do privado) ou *Self_ employed* (trabalhador por conta própria);
8. ***Resident_ type:*** tipo de residência - *Rural* ou *Urban*;
9. ***Avg_ Glucose_ Level:*** nível médio de glucose no sangue;
10. ***BMI:*** índice de massa corporal;
11. ***Smoking_ status:*** nível de tabaco - *formerly smoked* (antigo fumador), *never smoked* (nunca fumou), *smokes* (fumador) ou *Unknown* (desconhecido);
12. ***Stroke:*** 0 se o paciente não teve AVC ou 1 se o paciente teve AVC.

Uma vez apresentada uma breve descrição do conjunto de dados, de seguida efetuamos uma análise detalhada ao nosso conjunto de dados. Começamos por verificar se existiam valores duplicados e valores em falta. Posteriormente, para os atributos numéricos, apresentamos alguns dados estatísticos (como a média, mínimo, máximo) e ainda os seus *outliers*. Seguidamente, para cada variável independente fizemos uma análise individual, onde contamos o número de ocorrências para cada instância do atributo. Para além disso, relacionamos todas as variáveis independentes com a variável dependente *stroke*, de forma a perceber a quantidade/percentagem de indivíduos que

têm stroke tendo em conta as instâncias de cada variável independente. De seguida, e tendo em conta apenas os atributos numéricos, fizemos uma matriz de correlação usando o método de *Pearson*, de forma a entender como os atributos se correlacionam entre si. Adicionalmente criamos não só *scatterplots* como também *kdeplots* para tentar compreender a distribuição e relação de todos os dados das variáveis numéricas.

Por último, através de vários *groupby*, realizamos uma análise "composta", ou seja, pegamos em mais do que duas variáveis (na maioria das vezes 2 variáveis independentes e a variável dependente) e contamos quantas ocorrências haviam tendo em conta esses atributos. Por exemplo, quantos registos tinham AVC tendo hipertensão e não tendo doenças cardíacas.

Após a análise detalhada do nosso conjunto de dados podemos tirar muita informação relevante, nomeadamente que:

- não temos valores duplicados, mas temos 201 valores em falta no atributo *bmi*;
- temos pacientes com menos de 1 ano e pacientes com mais de 80 anos;
- tanto o atributo *avg_glucose_level* como o atributo *bmi* têm alguns outliers;
- a variável dependente está muito desequilibrada, uma vez que temos apenas 249 pacientes com *stroke* contra 4861 pacientes sem *stroke*;
- há mais paciente do sexo feminino (2994) e menos pacientes do sexo masculino (2115), no entanto há mais homens do que mulheres com *stroke*;
- quanto maior a idade maior é o risco de *stroke*;
- quanto maior for o *bmi* maior é o risco de *stroke*;
- há muito mais pacientes com sem hipertensão e doenças cardíacas do que pacientes com hipertensão e doenças cardíacas;
- 13.25% dos pacientes que têm hipertensão têm *smote* e apenas 3.97% dos pacientes que não têm hipertensão têm *smote*;
- 17.03% dos pacientes que têm doeças cardíacas têm *smote* e apenas 4.18% dos pacientes que não têm doenças cardíacas têm *smote*;
- os pacientes em que o tipo de trabalho é *self_employed* são o que têm mais *stroke* e de seguida são os pacientes do setor *private* e *govt_job*;
- os pacientes que tiveram *stroke* têm entre 50 a 80 anos;
- o facto de um paciente ter hipertensão e doenças cardíacas não é muito determinante para o paciente ter *stroke*.

3 Tratamento dos Dados

Uma vez feita a descrição e análise dos dados, e depois de termos descoberto algumas informações relevantes, iniciamos o tratamento/preparação dos dados de forma a garantir o melhor conjunto de dados possível para a realização dos modelos. Começamos por substituir os valores em falta do atributo *bmi*, sendo que substituímos pela média dos valores arredondado a uma casa decimal. Optamos pela média dos valores, porque a média do *bmi* dos indivíduos que tinham AVC e a média do *bmi* dos indivíduos que não tinham AVC era muito próxima da média do *bmi*.

Posteriormente, e como vimos na análise de dados, a variável dependente *stroke* apresenta um desequilíbrio que pode prejudicar/enviesar o desempenho dos modelos, o que nos levou a aplicar o *smote*. O *smote* é uma técnica que permite gerar dados sintéticos a partir do conjunto de dados original, e assim equilibrar os dados. Uma vez que a maioria dos registos eram de pacientes que não tiveram *stroke*, ao aplicar o *smote* tivemos apenas em consideração os registos em que os pacientes tinham *smote* (ou seja tinham AVC = 0), uma vez que é a instância menos representativa. Para além disso, ao aplicar o smote utilizamos uma metodologia híbrida, onde acrescentamos dados sintéticos à classe com menor frequência e removemos dados da classe mais representativa (com maior frequência). Como resultado final do *smote*, temos um conjunto de dados onde temos o mesmo número de registos para ambas as classes da variável dependente (*Figura X*).

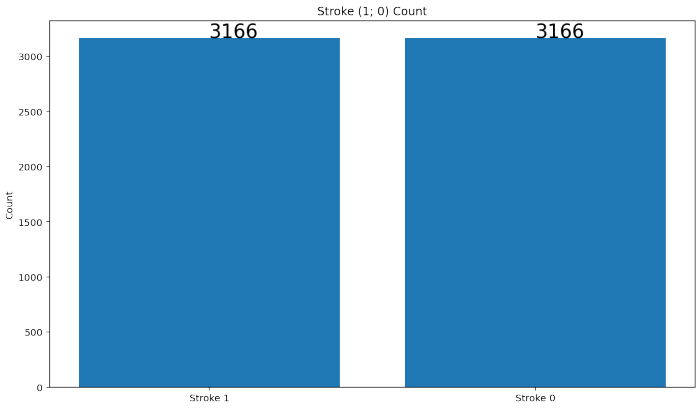

Figura 1: Conjunto de Dados após o *smote*.

4 Modelos Não Supervisionados

Os modelos não-supervisionados que utilizados, foi a redução de dimensionalidade através da análise de componentes principais (PCA), e o modelo K-Means.

Na utilização do PCA, reduzimos os dados a 2 componentes principais e através da distribuição dos dados no gráfico conseguimos perceber que a redução de dimensionalidade não nos iria trazer umas distribuição favorável para conseguirmos perceber o comportamento distinto entre as observações com stroke e sem stroke.

De qualquer forma conseguimos perceber que no caso da distribuição dos dados, as observações com stroke localizam-se mais do lado esquerdo do gráfico, podendo ser o fator idade a influenciar esta distribuição (com base na análise de dados).

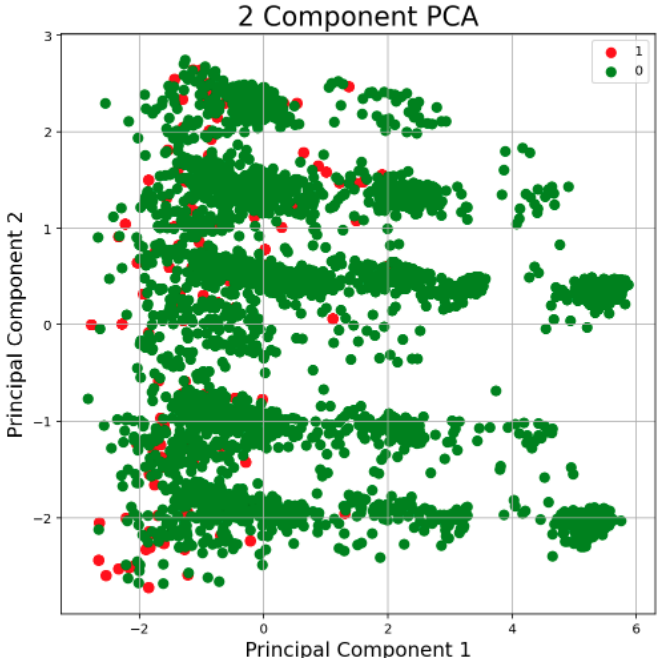

Figura 2: Análise de Componentes Principais

Com o Modelo K-Means tentamos dividir os dados em 2 clusters, na esperança em que cada um se adequa-se a cada uma das classes, tornando este problema semi-supervisionado.

De facto este tipo de modelos não é o mais adequado no contexto destes dados, podemos observar no gráfico em baixo que a performance do K-Means não foi a melhor na utilização dos dois clusters para dividir a distribuição, sendo que dados de ambas as classes ficaram distribuídos em ambos.

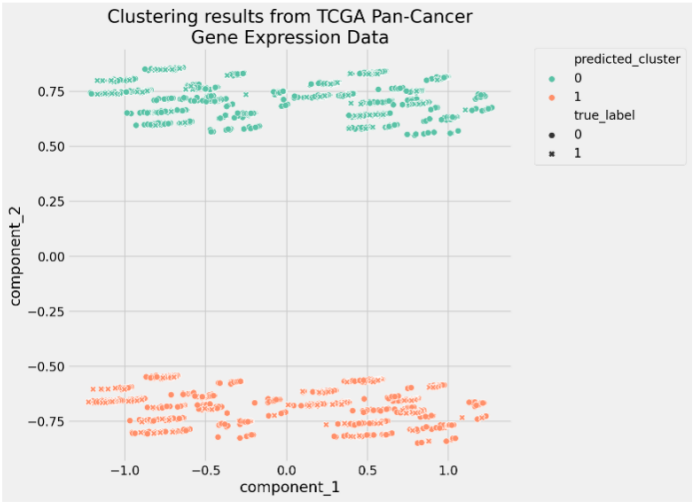

Figura 3: K-Means com PCA

5 Seleção de Atributos

Foram utilizados métodos de selecção de atributos com o objetivo de perceber que variáveis independentes é que são mais importantes na explicação da variavel dependente.

Os métodos utilizados foram :

1. ***Chi-Squared Test***
2. ***Extra-Tree Classifier***
3. ***Recursive Feature Elemination with Random Forest***

A principal conclusão que o grupo tirou na utilização destas três técnias é que os atributos Age, AVG_glucose_level e bmi são as consideradas mais influentes nos resultados dos modelos que vamos utilizar.

Foi usado também um teste VIF(Variance Inflation Factor) para percebermos se existia problemas de multicolineariedade entre as variáveis numéricas. Com base neste teste e na análise de dados, o grupo percebeu que as variáveis independentes não sofrem de problemas de multicolineariedade sendo que, apesar de umas serem mais importantes que outras, não faz sentido excluirmos nenhuma para os modelos de previsão supervisionados.

6 Modelos Supervisionados

É importante percebermos que a variavél dependente (Stroke) representa um problema de saúde, sendo que durante o desenvolvimento dos modelos é mais importante estes conseguirem prever os indivíduos que vão ter este problema do que os indivíduos que não o vão ter, isto é, vamos dar mais valor aos modelos que conseguem prever mais casos com stroke. Em todos os modelos, menos na regressão logística, foram utilizados métodos para a optimização de hiperparametros (RandomSearchCV). O treino dos modelos era com vista a melhorar as medidas F1-Score, Recall e AUC, pois se utilizássemos a accuracy, os resultados no teste não seriam os melhores sendo que os dados são desbalanceados e há mais importância na classe com menos frequência.

O primeiro modelo a ser utilizado foi a regressão logística. Foram testados os dados antes do SMOTE(desbalanceados) e depois do SMOTE(Balanceados).

A accuracy com os resultados desbalanceados é muito elevada mas o modelo nao é capaz de prever nenhuma observação com stroke.

```
Accuracy of logistic regression classifier on test set: 0.95
[[1695    0]
 [  94    0]]
              precision    recall  f1-score   support

           0       0.95      1.00      0.97      1695
           1       0.00      0.00      0.00        94

    accuracy                           0.95      1789
   macro avg       0.47      0.50      0.49      1789
weighted avg       0.90      0.95      0.92      1789
```

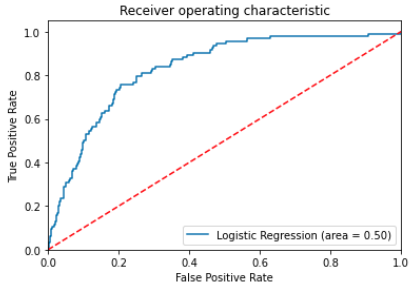

Figura 4: Resultados da Regressão Logistica com dados desbalanceados

Os resultados melhoram bastantes com os dados balanceados, o modelo piora a sua accuracy mas melhora na previsão das observações com stroke = 1.

O modelo Random Forest apresentou resultados interessantes, sendo que a accuracy é ainda mais baixa que a do ultimo modelo da regressão logística, mas existe uma boa performance nas previsões aos casos em que o stroke é positivo.

A rede Neuronal densa é constituída por 2 camadas densas intermédias, cada uma com 20 neurónios sendo que a melhor performance foi obtida com 20 épocas e o batch size de 63. Os resultados foram interessantes como os do random forest sendo que o modelo tem pior performance na previsão dos individuos com stroke e melhor performance das previsões gerais.

```
Accuracy of logistic regression classifier on test set: 0.72
[[1217  478]
 [  19   75]]
              precision    recall  f1-score   support

           0       0.98      0.72      0.83      1695
           1       0.14      0.80      0.23        94

    accuracy                           0.72      1789
   macro avg       0.56      0.76      0.53      1789
weighted avg       0.94      0.72      0.80      1789
```

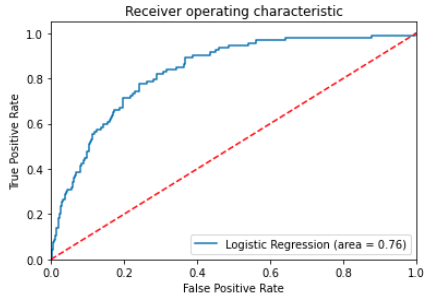

Figura 5: Resultados da Regressão Logistica com dados balanceados

```
[[1127  568]
 [  12   82]]
              precision    recall  f1-score   support

           0       0.99      0.66      0.80      1695
           1       0.13      0.87      0.22        94

    accuracy                           0.68      1789
   macro avg       0.56      0.77      0.51      1789
weighted avg       0.94      0.68      0.77      1789
```

Figura 6: Resultados do Random Forest

```
[[1237  458]
 [  19   75]]

Accuracy:
0.7333705980994969

Precision:
0.14071294559099437

Recall:
0.7978723404255319

F1 Score:
0.23923444976076558
```

Figura 7: Resultados da Rede Neuronal Densa

7 Conclusão

Este Relatório aborda algumas partes fundamentais do todo que foi o desenvolvimento do projeto.

O grupo de trabalho tentou abordar um pouco de todo o conhecimento que foi ganho durante as aulas de Aprendizagem Automática II e com a ajuda da técnica SMOTE conseguimos chegar a resultados interessantes nos modelos supervisionados.

Fica aqui o pensamento futurista de que seria possível, com mais tempo, obter ainda melhores resultados nos modelos. O grupo ganhou bastante conhecimento e experiência para desenvolvimento de modelos e análise de dados dando ênfase ao tratamento de dados com datasets desbalanceados e o conhecimento de que nem todos os problemas podem ser solucionados usando várias metodologias, como este caso em que os modelos não-supervisionados não forneceram boa performance nos resultados pretendidos, sendo que é importante perceber sempre qual será a melhor solução para cada tipo de problema.

Em conclusão, os melhores resultados foram obtidos com o modelo random forest e uma rede neuronal densa.

O trabalho futuro proposto é, para além da melhoria dos resultados, o desenvolvimento de um front-end em que o utilizador seja capaz de inserir os seus dados num formulário e receber como resposta a probabilidade de stroke.