Analise e conceção de um modelo de regressão logística aplicado a uma base de dados

Aprendizagem automática

Autores:

João Sá PG41078

Tiago Silva PG41100

Conteúdo

1. Introdução

Para a realização do projeto da disciplina de Aprendizagem Automática I foi proposto encontrar e explorar um dataset. Neste é esperado fazer o tratamento de dados necessário e criar um modelo estatístico de modo a fazer previsões para novas entradas no dataset.

Decidimos utilizar o Datasse "*Pima Indian Diabetes*" que se trata de uma base de dados com informações de marcadores biológicos de pacientes femininos pertencentes a uma população indígena. O dataset foi criado com o intuito de analisar o impacto de múltiplas gravidezes no desenvolvimento dos diabetes.

Iremos começar por analisar todos os nossos atributos, bem como relações entre eles através de gráficos de modo a retirar conclusões sobre o dataset e auxiliar também o nosso tratamento de dados.

Para a realização do modelo vamos utilizar regressão logística pois a classe é uma variável categórica e foi uma das técnicas abordada nas aulas da disciplina.
Com este dataset nós temos dois objetivos:

- Prever com alguma precisão se o paciente tem ou não tem Diabetes;
- Mostrar o impacto das gravidezes para a ocorrência desta doença;

A base de dados é constituída por 768 casos onde cada caso tem oito atributos e uma classe como mostra a figura abaixo:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figura 1. Pima Indian Diabetes

2. Metodologia

Para resolução do modelo proposto iremos utilizar uma técnica estatística estudada na unidade curricular, regressão logística.

A regressão logística é utilizada para fazer previsões de classes categóricas e frequentemente binárias tal como o nosso caso de estudo. A regressão logística faz a analise de dados distribuídos binomialmente, figura 2, em que p_i é desconhecido e n_i são os números de ensaios de Bernoulli que são conhecidos.

$$Y_i \sim B(p_i, n_i), \text{ for } i = 1, \ldots, m,$$

Figura 2 - Formula de regressão logística

Esta técnica é geralmente usada em problemas logísticos nas seguintes áreas:

- Problemas médicos
- Companhias de seguros
- Instituições financeiras

3. Resultados

3.1. Análise gráfica

Antes de começarmos a construir o nosso modelo e retirarmos os atributos menos importantes decidimos analisar os nossos dados de modo a entender melhor o nosso dataset e percebermos melhor as possíveis relações entre os diferentes atributos.

O primeiro gráfico que criamos foi um gráfico de barras com a classe de modo a analisar a diferença de casos de pacientes com Diabetes e para pacientes sem Diabetes que temos disponíveis na base de dados.

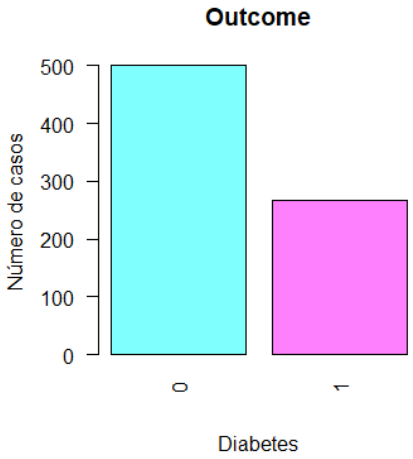

Figura 3. Número de casos com Diabetes

Conseguimos verificar que o número de casos sem Diabetes é quase o dobro em relação ao número de casos com Diabetes.

Em relação ao segundo gráfico que visualizamos foi um diagrama de caixas com todos os atributos para termos uma melhor perceção dos nossos dados.

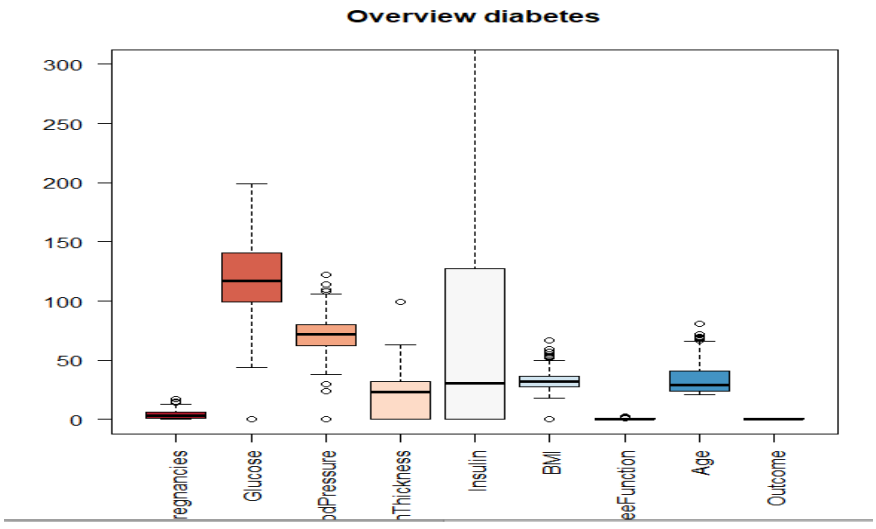

Figura 4. Overview da base de dados

Com a Figura 2 conseguimos ter uma melhor perceção da Amplitude Interquartil de cada variável bem como a sua mediana. Achamos que este gráfico ajudou para termos uma ideia melhor da dimensão dos atributos da base de dados.

Após termo analisado os atributos individualmente, analisámos os diferentes atributos com a classe para tentarmos encontrar relações entre eles.

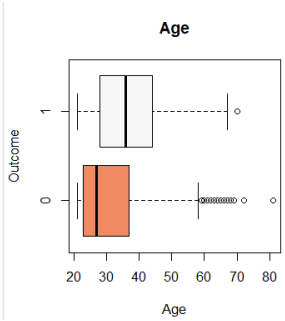

Figura 5. Relação da Idade com a Classe

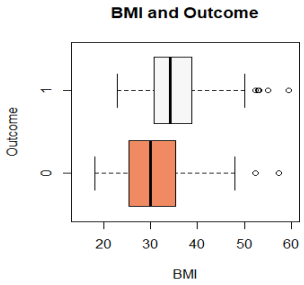

Figura 6. Relação do BMI com a classe

Em relação à Figura 3 podemos verificar uma relação direta entre a idade e o risco de desenvolver diabetes, com o diagrama de caixa podemos verificar o impacto que a idade tem para a ocorrência de diabetes, claramente que a média de idade e a variância Interquartil é superior quando a paciente tem Diabetes.

Quanto à Figura 4 podemos verificar o mesmo, com o aumento do BMI é mais provável a paciente ter Diabetes e a sua variância Interquartil também é superior.

Na próxima Figura vamos analisar a relação das Gravidezes com a Classe sendo que um dos objetivos do nosso trabalho é mostrar o impacto que as Gravidezes têm para a ocorrência de Diabetes.

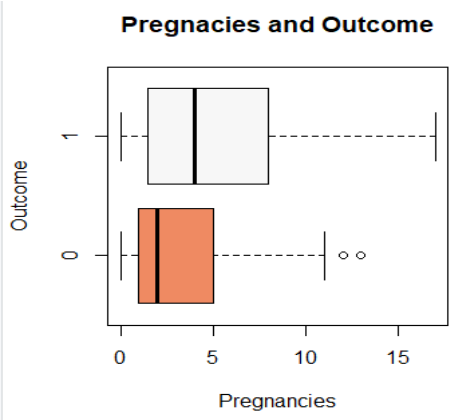

Figura 7. Relação Gravidezes com a Classe

Na figura 5 podemos analisar a variância Interquartil de gravidezes quando a paciente não tem diabetes (Outcome a 0) e quando têm (Outcome a 1). Podemos verificar que no povo Indígena existe mais gravidezes que o normal, no entanto, a diferença é evidente. Conseguimos concluir que as Gravidezes têm um grande impacto para o desenvolvimento de Diabetes na população feminina.

Após analisarmos os atributos individualmente com a classe tentamos analisar a relação entre os atributos com a classe onde foi possível obtermos resultado interessantes. Na próxima figura vamos mostrar a relação da Idade e da Glucose.

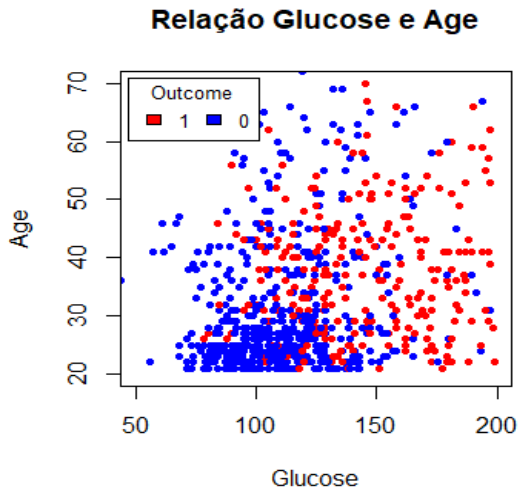

Figura 8. Relação entre a Glucose a Idade

Conseguimos visualizar uma relação entre a Glucose e a Idade, conforme a Idade sobe, a Glucose tem tendência a subir e sendo este um dos fatores de risco para o desenvolvimento de diabetes percebe-se por esta análise de dados que a idade é um fator de risco no desenvolvimento de diabetes.

Em seguida vamos mostrar a relação entre a Glucose e a pressão arterial.

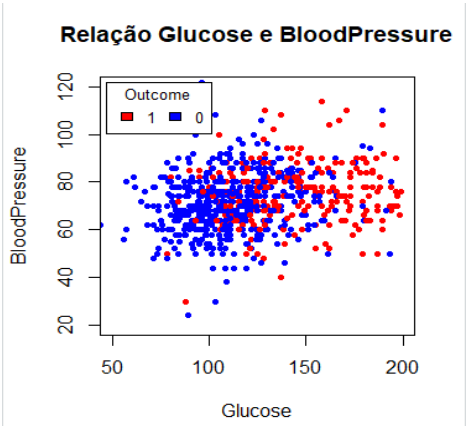

Figura 9. Relação entre Blood Pressure e Glucose

Na Figura 7 conseguimos verificar que a relação entre ambos os atributos não é muito evidente comparado com a Figura 6. Podemos analisar que com a Glucose baixa os pacientes raramente têm Diabetes sendo que o valor de Pressão Sanguínea não tem tanto impacto pois, com valores de Glucose inferiores a 120 são raros os casos de pacientes com Diabetes e com Glucose superior a 150 quase todos os pacientes têm Diabetes.

Na próxima Figura iremos analisar a relação entre os atributos BMI e Glucose.

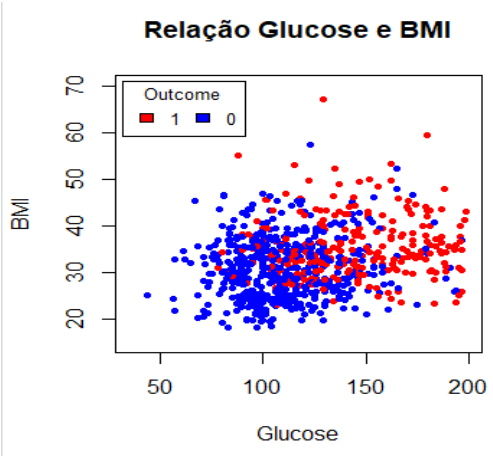

Figura 10 - Relação entre glucose e BMI

Na Figura 8 é possível visualizar uma relação entre ambos os atributos. Com um BMI inferior a 30 e Glucose inferior a 130 quase todos os pacientes não têm Diabetes e, quando aumentamos ambos os atributos a frequência de casos com Diabetes é quase 100%.

Após termos analisado as relações entre os diferentes atributos e mostrado o impacto de cada um para a determinação do diagnostico de diabetes decidimos fazer uma Matriz de Correlação onde resumimos a relação entre os atributos.

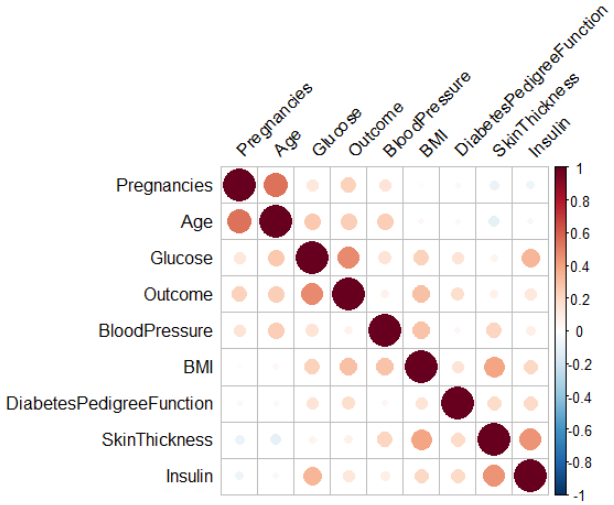

Figura 11. Matriz de correlação

Na Figura 9 conseguimos analisar a Matriz de Correlação onde podemos obter as relações entre os atributos, uma das maiores relações que podemos verificar é entre a Idade e as Gravidezes o

que é por demais evidente e decidimos que não era necessário fazer um gráfico para explicar essa relação.

Com isto terminamos a nossa visualização dos dados que mostrou ser de grande importância para a perceção da Base de dados, após analisar os dados conseguimos entender melhor todos os atributos.

4. Missing Values

Os Missing Values verificam-se nos atributos Insulin e SkinThickness porque é do senso comum que a Insulin e SkinThickness nunca podem ter valores igual a zero.

Perante esta situação nós encontramos três soluções:

- Eliminar os Missing Values;
- Eliminar os dois atributos;

Decidimos eliminar ambos os atributos no nosso modelo, se eliminarmos os Missing Values a base de dados ficava com 394 casos, ou seja, perdíamos 374 casos sendo que, mesmo após eliminar os Missing Values, os *p-value* dos atributos continuavam longe de zero e por isso não tinha qualquer sentido reduzir o Dataset por causa de atributos que não mostram importância para a previsão.

5. Modelo

5.1. Dataset de treino e teste

De modo a treinar e avaliar corretamente o nosso modelo, foi feita uma divisão do dataset em dois, 80% para treino e 20% para teste. Esta divisão é um passo essencial para o treino de modelos e analise da sua precisão pois permite que não haja *overfitting* no modelo.

5.2. Modelo inicial

Inicialmente, inclui-se todas as variáveis disponíveis de modo a perceber quais as variáveis significativas para o problema. Definiu-se que para uma variável ser considerada significativa definiu-se um "*P-Value*" (teste de hipótese nula) máximo de 0.05, eliminando todas as variáveis que fossem superiores.

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -7.9362649  0.7794558 -10.182  < 2e-16 ***
Pregnancies               0.1262019  0.0356066   3.544 0.000394 ***
Glucose                   0.0333615  0.0040609   8.215  < 2e-16 ***
BloodPressure            -0.0144375  0.0058654  -2.461 0.013836 *
SkinThickness             0.0021461  0.0075941   0.283 0.777477
Insulin                  -0.0016035  0.0009778  -1.640 0.101033
BMI                       0.0853748  0.0161826   5.276 1.32e-07 ***
DiabetesPedigreeFunction  0.8789220  0.3278797   2.681 0.007349 **
Age                       0.0141239  0.0101652   1.389 0.164698
```

Figura 12 - Sumário do modelo

Após a análise á figura 10, considera-se os Pr(>|Z|) < 0.05 como não relevantes para o modelo como definido anteriormente, eliminando-se os seguintes atributos:

- Age, este atributo tal como foi analisado anteriormente fazia correlação com o número de gravidezes, dai ter um *P-Value* elevado.
- Insulin, é importante notar que embora com muitos valores em falta neste atributo, mesmo com estes removidos, a variável continuava com um *P-Value* muito alto
- SkinThickness, tem o P-Value mais alto e, portanto, é a primeira a ser removida dado que não tem valor para o nosso modelo.

O modelo atingiu uma precisão de 79% nos dados de teste, que embora seja uma precisão relativamente alta é preciso analisar a matriz de confusão, figura 11, de modo a perceber se o modelo está a fazer boas previsões para os dois resultados possíveis (diabética ou não diabética).

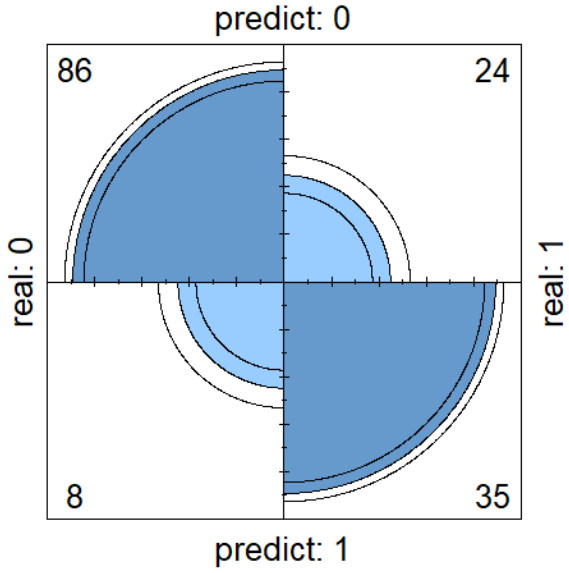

Figura 13 - Matriz confusão do modelo

Analisando a matriz de confusão, figura 11, consegue-se perceber que o pior caso, prever a paciente como não diabética e ser diabética aconteceu 24 vezes. É importante melhorar o modelo de modo a evitar os falsos negativos.

5.3. Modelo final

Após a análise do modelo inicial, foram removidas uma a uma as variáveis consideradas estatisticamente irrelevantes para o modelo. Desta forma é esperado que além de mais simplicidade no modelo, atinja-se melhores resultados.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -7.746244   0.747246 -10.366  < 2e-16 ***
Pregnancies              0.144359   0.030811   4.685 2.79e-06 ***
Glucose                  0.032543   0.003713   8.764  < 2e-16 ***
BloodPressure           -0.013248   0.005513  -2.403  0.01626 *
BMI                      0.091435   0.015878   5.759 8.48e-09 ***
DiabetesPedigreeFunction 0.866943   0.332191   2.610  0.00906 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 14 – Sumário do modelo Final

Removidas as três variáveis consideradas, obtemos o modelo da fig12. Todas as variáveis apresentam um *p-value abaixo dos 0.05 e considera-se este o modelo final. Analisando em específico o coeficiente da variável ‘Pregnancies’, é percetível o impacto no desenvolvimento de diabetes.*

O modelo atingiu uma precisão de 81% nos dados de teste, no entanto é preciso analisar a matriz de confuso de modo a perceber a qualidade do modelo, figura 13.

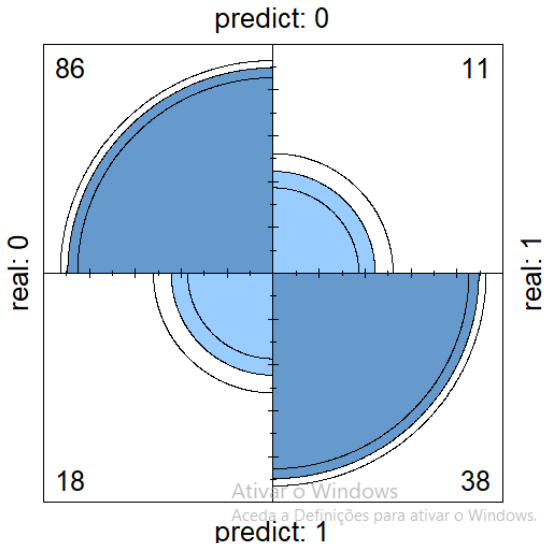

Figura 15 - Matriz confusão do modelo final

Analisando a matriz de confusão consegue-se perceber que os falsos negativos reduziram consideravelmente, no entanto os falsos positivos subiram o que também é problemático quando se trata de diagnósticos de problemas metabólicos. Considerando estes fatores e que, no geral a precisão aumentou consideramos este modelo melhor e mais eficiente no diagnóstico de diabetes.

6. Conclusão

Em suma, a analise aos dois modelos mostra que embora se consiga fazer previsões corretas quanto ao diagnostico de diabetes em pacientes, o número de falsos positivos e falsos negativos implicam que este modelo não possa ser usado como preditor da condição, podendo ser apenas usado como auxílio ao diagnostico.

De modo a melhorar o modelo seria necessário uma base de dados com mais pacientes, os valores de *insulina* e *skinthickness* corretamente fornecidos e só assim há expectativa de atingir valores acima de 90% de precisão nas previsões.

No entanto, consideramos que cumprimos com a proposta inicial e desenvolvemos uma análise de dados ao problema eficaz e metódica, além de conseguir a implementação de um modelo que consegue fazer previsões corretamente.

7. Bibliografia

[1]. Inês Sousa, Slides Aprendizagem Automática, Universidade do Minho, (2019)

[2] Bittencourt, H.R.. (2012). Regressão logística politômica: revisão teórica e aplicações. 5.