Previsão da taxa de concessão de bolsas

Andrea Sampaio Elias
Departamento de Teleinformática line
Universidade Federal do Ceará
Fortaleza, Brasil
andrease@alu.ufc.br

Arnaldo Aguiar Portela Filho
Departamento de Teleinformática line
Universidade Federal do Ceará
Fortaleza, Brasil
arnaldoaguiarp@gmail.com

José Vanderley Sousa De Freitas Filho
Departamento de Teleinformática line
Universidade Federal do Ceará
Fortaleza, Brasil
vanderley_freitas@alu.ufc.br

Marcelo Colares Da Silva
Departamento de Teleinformática line
Universidade Federal do Ceará
Fortaleza, Brasil
colaresmarcelo2018@gmail.com

***Resumo* — Este artigo trata do estudo e da modelagem de dados de concessão de bolsas. Tem-se como objetivo obter a previsão de quais aplicações de pedido de bolsa são mais prováveis de obter sucesso. Assim, perde-se menos tempo na análise dos pedidos de bolsa.**

Palavras-chave – bolsas de estudo, análise, regressão logística, LDA, KNN, QDA, python, matriz de confusão, probabilidade.

I. INTRODUÇÃO

No mundo inteiro, houve redução na quantidade de fundos disponíveis para a concessão de bolsas de pesquisa, o Brasil é um dos países onde este fato vem ocorrendo. Na Austrália, as taxas de sucesso nas aplicações de pedido de bolsa diminuíram em média 75%. Com isso, muitos estudantes estão perdendo tempo com aplicações fadadas ao insucesso.

Com esse problema em mente, fez-se necessário a criação de um modelo de análise para prever o sucesso dos pedidos de subsídios. O modelo será usado pela Universidade de Melbourne para prever quais solicitações de bolsas têm probabilidade de serem bem-sucedidas, de forma que menos tempo seja desperdiçado em solicitações que provavelmente não serão. Espera-se também que se esclareçam quais fatores são importantes para determinar se uma inscrição será bem-sucedida.

O campo de estudo escolhido foi a Universidade de Melbourne, que é uma universidade pública cujo principal campus fica em Melbourne, Victoria, na Austrália. Segunda universidade australiana mais antiga, seu principal campus é em Parkville, um subúrbio ao norte do centro financeiro da cidade.

II. MÉTODOS

Diversos métodos de classificação foram utilizados na modelagem do conjunto de dados em análise. No presente estudo, analisando-se dados relacionados à concessão de bolsas de estudo, com o objetivo de conseguir prever quando as aplicações são bem-sucedidas, usou-se inicialmente o método conhecido como LDA (Linear Discriminant Analysis), em tradução, a análise discriminante linear.

A análise discriminante linear é um método muito comumente usado como técnica de redução de dimensionalidade na etapa de pré-processamento de dados para a classificação de padrões e em aplicativos de aprendizado de máquina. O objetivo do método é projetar um conjunto de dados em um espaço de dimensão inferior com boa separabilidade de classes para evitar o overfitting e também reduzir os custos computacionais.

Inicialmente separou-se o conjunto de dados entre X's e y's de treino e de teste. Os preditores usados como saída (y) são justamente a classificação das aplicações de pedido de bolsa, se bem-sucedidas ou malsucedidas.

	Class
0	successful
1	successful
2	successful
3	successful
4	unsuccessful
...	...
8185	unsuccessful
8186	unsuccessful
8187	successful
8188	successful
8189	successful

8190 rows × 1 columns

	Class
0	unsuccessful
1	unsuccessful
2	successful
3	successful
4	successful
...	...
513	unsuccessful
514	successful
515	successful
516	unsuccessful
517	successful

518 rows × 1 columns

Fig. 1. – À esquerda tabela apresentando o conjunto de dados de y de treino. À direita tabela apresentando o conjunto de dados de y de teste.

Observa-se que os valores dessas tabelas são valores não numéricos e, como tais, não podem ser modelados tão facilmente. Como estratégia foi feita a substituição destes por valores numéricos. Os dados da coluna de classificação ‘unsucessful’ (malsucedido) e ‘sucessful’ (bem-sucedido) foram substituídos pelos valores, 0 e 1, respectivamente, como se demonstra:

	Class
0	1
1	1
2	1
3	1
4	0
...	...

	Class
0	0
1	0
2	1
3	1
4	1
...	...

8185	0
8186	0
8187	1
8188	1
8189	1

8190 rows × 1 columns

513	0
514	1
515	1
516	0
517	1

518 rows × 1 columns

Fig. 2. – À esquerda tabela apresentando o conjunto de dados de y de treino com valores numéricos. À direita tabela apresentando o conjunto de dados de y de teste com valores numéricos.

Passou-se então à modelagem dos dados com a LDA. Foi feito o treinamento do modelo com os dados de treino e então o modelo foi testado com o conjunto de dados de teste. Com os resultados obtidos foi criada uma matriz de confusão para melhor compreensão do desempenho do modelo. A classificação é feita com base no discriminante linear, que será responsável por criar uma fronteira entre os dados:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

Como passo seguinte faz-se o uso de modelos não lineares afim de se observar sua performance em relação aos dados em estudo. O primeiro método não linear utilizado foi o 'KNN' (k-Nearest Neighbors).

O algoritmo k-Nearest Neighbors (KNN), ou método dos vizinhos mais próximos, foi utilizado para analisar o conjunto de dados do treinamento. O KNN realiza previsões usando o conjunto de dados de treinamento diretamente. As previsões são feitas para uma nova instância (x) pesquisando todo o conjunto de treinamento para as K instâncias mais semelhantes (os vizinhos) e resumindo a variável de saída para essas instâncias de K. Para a regressão, essa pode ser a variável de saída média; na classificação, esse pode ser o valor de classe do modo (ou mais comum).

O critério de decisão é feito com base na distância entre as amostras, onde uma amostra qualquer é atribuída a uma determinada classe à qual ela possui uma menor distância (*e*):

$$e = \sqrt{\sum_{j=0}^{n} (X_{i,j} - X_{i+n,j})^2}$$

Esse método calcula a distância entre dois vetores, em nossa situação foram representados como sendo linhas de uma matriz que é sua representação mais comum em aplicações reais.

O KNN irá encontrar os K vizinhos mais próximos de um dado ponto no espaço, ele então procura pelo seguinte conjunto de vetores.

$$\underset{X_{i,j} \subset X}{argmin} \sum_{i=0}^{K} \sqrt{\sum_{j=o}^{N} (X_{i,j} - t_i)^2}$$

Onde $X_{i,j}$ são linhas da matriz X, K e N são os números de vizinhos e número de elementos do vetor e é a instância que será classificada.

Isso retornará um conjunto de vetores $k_n = \{x_1, x_2, x_3\}$, onde k_n terá um total de 16 valores de x, no qual soma das distâncias entre eles e a instância t que será classificada seja a menor possível.

Feito isso é preciso saber qual a classe dominante nesses vetores, cada vetor do conjunto está associado a uma classe. Para descobrir a qual classe a instância t pertence é feito uma votação para saber em qual classe ela se encaixa. Isso é definimos a instância em questão como pertencendo a classe que tenha maior ocorrência no conjunto que retornar da iteração.

O segundo método não linear utilizado para classificação foi a Análise Discriminante Quadrática (QDA), nele os resultados da classificação são feitos primeiramente assumindo que as observações de cada classe são desenhadas a partir de uma Distribuição Gaussiana e cada um tem a sua Matriz de Covariância. Logo cada classe terá respectivamente uma média e uma Matriz de Covariância "representando - a"; com estes valores podemos aplicar o Bayes' classifier para minimizar a probabilidade de classificações erradas.

O método QDA tem seu valor mostrado quando a amostra de treino é muito grande de tal forma que a variância não seja um atributo de grande prioridade ou pelo simples fato de que não possível o conceito de uma Matriz de Covariância comum para todas as classes. Logo, nos casos que a amostragem de treino é pequena ou quando a matriz é comum a todas as classes, o QDA não é recomendado. A classificação é feita com base no discriminante linear, que será responsável por criar uma fronteira entre os dados, mas diferentemente do LDA, o QDA apresenta uma fronteira não linear:

$$\begin{aligned} \delta_k(x) &= -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k \\ &= -\frac{1}{2} x^T \Sigma_k^{-1} x + x^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k + \log \pi_k \end{aligned}$$

Uma matriz de confusão é uma tabela que permite a visualização do desempenho de um modelo de um método de classificação. É, portanto, uma tabela que mostra as frequências de classificação para resultados do modelo. Ela vai nos mostrar as quantidades:

- Verdadeiro positivo (true positive - TP): ocorre quando no conjunto real, a classe positiva foi prevista **corretamente**. No mundo real o resultado foi "bem-sucedido" e o modelo previu corretamente que foi "bem-sucedido".
- Falso positivo (false positive — FP): ocorre quando no conjunto real, a classe positiva foi prevista **incorretamente**. No mundo real o resultado foi "malsucedido" e o modelo previu incorretamente que foi "bem-sucedido".
- Verdadeiro negativo (true negative — TN): ocorre quando no conjunto real, a classe negativa foi prevista **corretamente**. No mundo real o resultado foi "malsucedido" e o modelo previu corretamente que foi "malsucedido".
- Falso negativo (false negative — FN): ocorre quando no conjunto real, a classe negativa foi prevista **incorretamente**. No mundo real o resultado foi "bem-sucedido" e o modelo previu incorretamente que foi "malsucedido".

A acurácia do modelo pode ser medida através dos dados da matriz de confusão, pela seguinte fórmula:

$$acurácia = \frac{TP + TN}{TP + TN + FP + FN}$$

Utilizou-se como forma de avaliação do modelo, a curva 'ROC'. A Curva Característica de Operação do Receptor (curva ROC) consiste em uma representação gráfica que ilustra a performance de um sistema classificador binário à medida que o seu limiar de discriminação muda. Ela também é conhecida como curva de característica de operação relativa, uma vez que o critério de mudança é resultado da operação de duas características, falsos positivos e falsos negativos. A curva ROC é obtida pela representação da razão RPV = Verdadeiros Positivos / Verdadeiros Totais versus a razão RPF = Falsos Positivos / Falsos Totais, para vários valores do limiar de classificação. O RPV também pode ser interpretado como a sensibilidade, taxa de verdadeiros positivos, e RPF = 1-especificidade ou taxa de falsos positivos. Seu estudo possibilita selecionar modelos possivelmente ideais (modelos ótimos) e descartar modelos não tão ótimos, independentemente do contexto de custos ou da distribuição de classe. Uma medida padrão de comparação entre sistemas é a área sob a curva (AUC).

III. Resultados

Os resultados obtidos com a LDA, primeiro método usado no estudo, e a modelagem do conjunto de dados são a seguir demonstrados.

A matriz de confusão obtida com os dados do modelo é a seguinte:

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo 284	Falso Negativo 33
	Não	Falso Positivo 45	Verdadeiro Negativo 156

Fig. 3. – Matriz de confusão com os dados obtidos com o modelo LDA, valores verdadeiro e falso positivos e verdadeiro e falso negativos.

Os valores obtidos na matriz de confusão foram: Verdadeiro Positivo (TP): 284, Verdadeiro Negativo (TN): 156, Falso Positivo (FP): 33, Falso Negativo (FN): 45. Com esses valores pode-se medir a acurácia do modelo. Como demonstrado a seguir, a acurácia obtida foi de aproximadamente 84,94%.

$$acurácia = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{284 + 156}{284 + 156 + 45 + 33} = 0{,}894208$$

Fez-se ainda, testes para saber com o uso de quantos preditores o modelo obtém melhor desempenho. Para tanto testou-se o modelo 252 vezes, ou seja, fez-se testes usando de 1 a 252 preditores (no gráfico demonstrado como entre 0 e 251) e obteve-se o desempenho do modelo para cada um deles. A seguir plotou-se o gráfico mostrando acurácia *versus* número de preditores usados, como demonstrado a seguir:

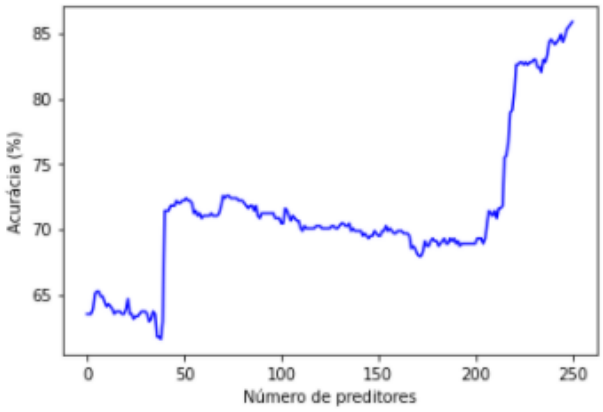

Fig. 4. – Gráfico mostrando acurácia *versus* número de preditores usados nos testes.

A partir da análise do gráfico acima, pôde-se concluir que o nível de acurácia é aumentado conforme se aumenta o número de preditores, dessa forma, torna-se necessária a utilização dos 252 preditores para se obter o melhor resultado possível, com a acurácia chegando a 85.91%. O modelo obtido no primeiro procedimento realizado é, portanto, capaz de prever o resultado das aplicações dos pedidos de bolsa quase que 86% corretamente.

O gráfico 'ROC' (Receiver Operation Characteristic Curve) obtido com o modelo LDA foi o seguinte:

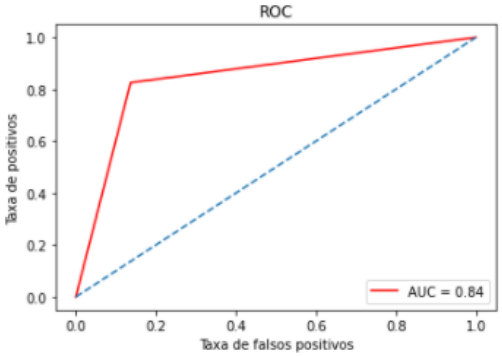

Fig. 5 – Gráfico 'ROC', com Taxa de positivos *versus* Taxa de falsos positivos.

Com o gráfico obtido pôde-se concluir que o modelo teve um bom desempenho em termos de taxas de positivos e falsos positivos.

No método KNN aplicou-se uma divisão em 16 instâncias para analisar o conjunto de treino e, com o resultado obtido, criou-se a matriz de confusão com os dados exibidos na tabela abaixo.

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo 274	Falso Negativo 55
	Não	Falso Positivo 89	Verdadeiro Negativo 100

Fig 6. - Matriz de confusão com os dados obtidos com o modelo KNN, valores verdadeiro e falso positivos e verdadeiro e falso negativos.

Com base na tabela de confusão encontrada, calculou-se a acurácia do modelo e obteve-se o valor de 72.20%, como demonstrado:

$$acurácia = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{274 + 100}{274 + 100 + 89 + 55} = 0{,}722007$$

Para confirmar os dados observados acima, plotou-se um gráfico do valor da acurácia comparando seu valor com a utilização de diferentes valores de K e, com sua análise, concluiu-se que o K com valor 16 obteve o melhor resultado para o conjunto de dados.

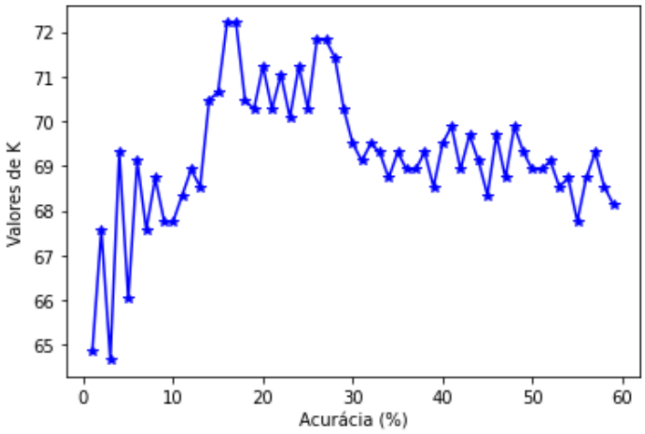

Fig. 7 – Gráfico de Valores de K *versus* Acurácia.

O gráfico 'ROC' (Receiver Operation Characteristic Curve) obtido com o modelo KNN foi o seguinte:

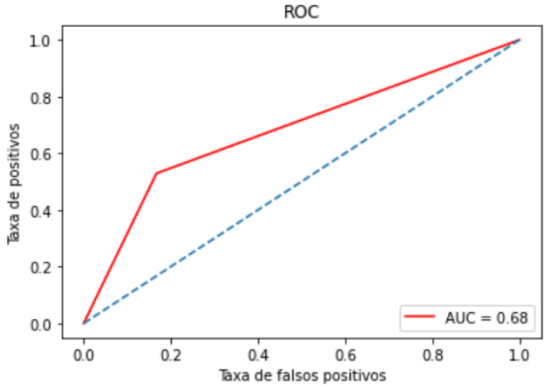

Fig 8. – Gráfico 'ROC', com Taxa de positivos *versus* Taxa de falsos positivos.

Com o gráfico obtido pôde-se concluir que o modelo teve um bom desempenho em termos de taxas de positivos e falsos positivos, apesar de ter tido uma menor performance em comparação ao obtido com o LDA.

No método QDA, baseando-se nos resultados obtidos foi possível calcular a média de acurácia, que resultou em 74,517%.

Com os dados obtidos, criou-se a matriz de confusão apresentada abaixo:

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo 264	Falso Negativo 67
	Não	Falso Positivo 65	Verdadeiro Negativo 122

Fig. 9 - Matriz de confusão com os dados obtidos com o modelo QDA, valores verdadeiro e falso positivos e verdadeiro e falso negativos.

Com os dados da matriz de confusão demonstra-se o cálculo da acurácia:

$$acurácia = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{264 + 122}{264 + 122 + 65 + 67} = 0{,}745174$$

O gráfico 'ROC' (Receiver Operation Characteristic Curve) obtido com o modelo QDA foi o seguinte:

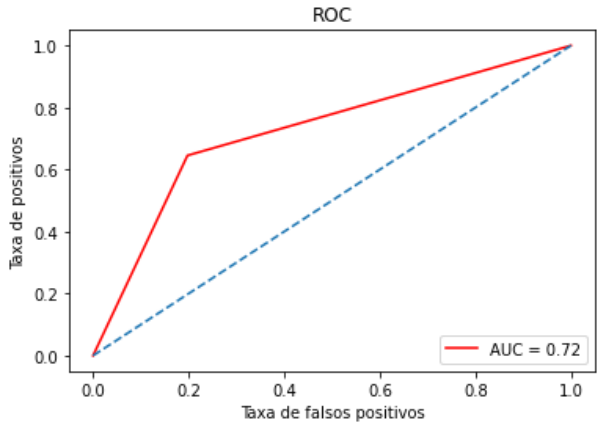

Fig. 10 -– Gráfico 'ROC', com Taxa de positivos *versus* Taxa de falsos positivos.

Com o gráfico obtido pôde-se concluir que o modelo teve um bom desempenho em termos de taxas de positivos e falsos positivos, apesar de ter tido uma menor performance em comparação ao obtido com o LDA.

Conforme foi visto até aqui, foram utilizados métodos de classificação linear e não-lineares; agora faz-se a comparação da performance desses métodos com a intenção de concluir se uma estrutura não linear melhora os resultados obtidos. Na figura a seguir tem-se a acurácia, porcentagem de classificações corretas, para os três métodos utilizados:

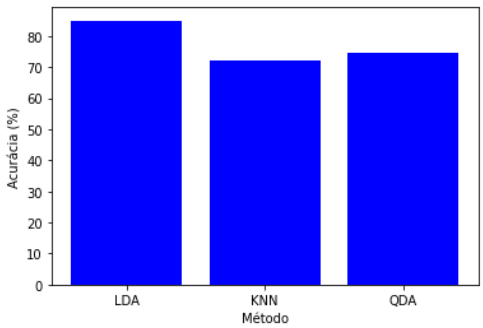

Fig. 11 – Gráfico mostrando a porcentagem de acurácia para cada um dos métodos utilizados no estudo.

Verifica-se que o LDA teve uma melhor performance em relação aos demais, entretanto, deve-se também verificar como está a distribuição dos falsos-positivos, alunos que não deveriam receber a bolsa e recebem, e falsos negativos, alunos que deveriam receber a bolsa e não recebem, conforme tem-se na tabela a seguir:

Método	Falsos-negativos	Falsos-positivos
LDA	45	33
KNN	55	89
QDA	65	67

Fig. 12 – Tabela de comparação dos métodos utilizados.

- Em relação ao Falsos-negativos, o método LDA apresentou a menor quantidade desse tipo de erro, ao passo que o KNN teve a 2ª melhor performance e em seguida o QDA.
- Em relação aos Falsos-positivos o LDA teve a menor quantidade de erros desse tipo, embora, dentre os métodos não lineares, o QDA se saiu melhor em comparação ao KNN.

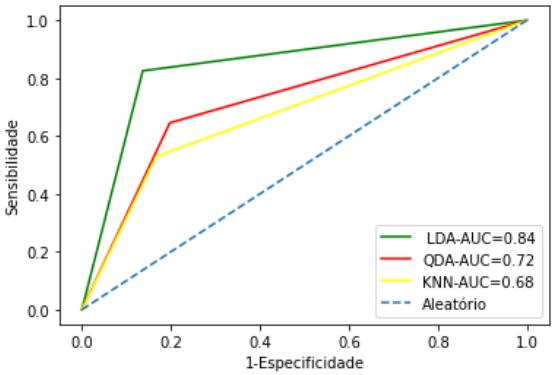

Fig. 13 – Curva ROC dos métodos utilizados.

Teoricamente, quanto maior a AUC (Área sob a curva), melhor o sistema. Pela Figura 11, conclui-se que o LDA foi o melhor método de classificação para esse conjunto de dados, seguido pelo QDA e o KNN, uma vez que a fronteira dos dados é aproximadamente linear; mesmo métodos não-lineares apresentando uma melhor performance na sequência de treino por ser mais flexível para o de teste, o LDA é melhor devido ao bias-variance trade off.

IV REFERÊNCIAS

[1] G. James, D. Witten, T. Hastie e R. Tibshirani, An Introduction to Statistical Learning with applications in R. Springer, 2013.

[2] M. Kuhn, K. Johnson, Applied Predictive Modeling. Springer, 2013.

[3] “Multidimensional binary search trees used for associative searching”, Bentley, J.L., Communications of the ACM (1975).

[4] “Five balltree construction algorithms”, Omohundro, S.M., International Computer Science Institute Technical Report (1989)1(1,2).

[5] “The Elements of Statistical Learning”, Hastie T., Tibshirani R., Friedman J., Section 4.3, p.106-119, 2008.

[6] Ledoit O, Wolf M. Honey, I Shrunk the Sample Covariance Matrix. The Journal of Portfolio Management 30(4), 110-119, 2004.

[7] R. O. Duda, P. E. Hart, D. G. Stork. Pattern Classification (Second Edition), section 2.6.2.

[8] Linear Discriminant Analysis. Disponível em: https://sebastianraschka.com/Articles/2014_python_lda.html. Acesso em 12 mar. 2021.

[9] Matriz de Confusão. Disponível em: https://pt.wikipedia.org/wiki/Matriz_de_confus%C3%A3o. Acesso em 14 mar. 2021.

[10] KNeighborsRegressor. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html. Acesso em 14 mar. 2021.

[11] Classificação e regressão com K-nearest neighbors. Disponível em: https://profes.com.br/julio.c.p.rocha/blog/classificacao-e-regressao-com-k-nearest-neighbors. Acesso em 16 mar. 2021.