Análise de modelos de classificação na previsão de pedidos de subsídios para pesquisas bem-sucedidos

Matheus Ferreira Lessa
Universidade Federal do Ceará
Fortaleza, Brasil
mathewslessa@gmail.com

Talles Felix Gomes
Universidade Federal do Ceará
Fortaleza, Brasil
tallesfelixg95@gmail.com

Paulo Victor Felício Lage
Universidade Federal do Ceará
Fortaleza, Brasil
victorlage7@gmail.com

***Resumo*—Seja um médico diagnosticando um paciente, um investidor escolhendo onde investir ou simplesmente uma caixa de e-mail identificando spans, em todos esses casos e em muitos outros está presente a classificação de observações em diferentes categorias, um problema bem conhecido na área de aprendizado de maquinas e que conta com diversas técnicas de abordagem. O objetivo desse trabalho consiste em fazer uma rápida introdução ao mundo dos classificadores e comparar a performance de três famosos modelos de classificação - Regressão logística, Máquina de vetores de suporte e K vizinhos mais próximos - na previsão do sucesso de pedidos de subsídios para pesquisas da universidade de Melbourne. Ao final do trabalho os resultados obtidos são apresentados e comparados.**

***Index Terms*—Classificação linear, classificação não linear, regressão logística, support vector machines, K vizinhos mais próximos (K nearest neighbor), matriz de confusão**

I. Introdução

Tomar decisões a partir de conhecimentos prévios pode uma tarefa difícil e muitas vezes tomar a decisão certa é de grande importância. É por isso que a classificação é uma das técnicas mais utilizadas no aprendizado de maquinas e conta com uma ampla gama de aplicações, incluindo detecção de spam, direcionamento de anúncios, avaliação de risco, diagnóstico médico e classificações de images.n

As técnicas de classificação são responsáveis por prever uma resposta qualitativa para uma observação, ou seja, lhe atribuir uma categoria/classe discreta. Diferindo assim das técnicas de regressão onde o resultado das previsões são valores contínuos. Embora alguns modelos de classificação também produzam predições contínuas, muitas vezes o foco está na predição discreta.

Os principais modelos de classificação podem ser divididos em dois grupos: os modelos lineares e os não lineares. No primeiro grupo encontram-se técnicas como a regressão logística, que apresenta certa semelhança com a regressão linear, e a análise discriminante linear (LDA[1]) que tem como base[2] o famoso classificador de Bayes, o classificador com a menor taxa de erro possível. Embora os modelos de regressão logística podem ser aplicados para a classificação de múltiplas classes, o método de análise discriminante linear é mais popular quando se tem mais de duas classes.

Dentre os modelos não lineares, temos a análise discriminante quadrática (QDA[3]) que assim como a LDA tem como base o classificador de Bayes, mas se torna um pouco mais flexível (quadrática ao invés de linear) que a LDA ao assumir que cada classe tem sua própria matriz de covariância. Além do QDA, temos também o modelo dos vizinhos mais próximos (KNN[4]). Como o seu nome sugere, ele classifica novas amostras com base na proximidade dela no espaço dos preditores com as amostras de treino cuja as classes são conhecidas. Um outro modelo não linear que nos últimos anos se tornou uma das ferramentas de aprendizado de máquina mais flexíveis e eficazes disponíveis são as Máquinas de vetores de suporte (SVM[5]). Nele cada amostra é vista como um ponto no espaço n-dimensional, com o valor de cada preditor sendo o valor de uma determinada coordenada. A classificação é então feita encontrando o hiperplano que melhor diferencia as classes. Por fim, temos as populares redes neurais, que se inspiram nos neurônios biológicos para a classificação de amostras.

Inúmeras são as aplicações práticas dos modelos citados anteriormente. Temos, por exemplo, o uso da regressão logística multivariada para o diagnósticos de diabetes de forma simples e prática [2] e também para a previsão de aprovação de empréstimos a partir de um conjunto de informações de um candidato [3]. Além disso, no campo da agricultura e da avicultura estão presentes diversas aplicações de máquinas vetoriais de suporte multi-classes [4]. Já o modelo dos vizinhos mais próximos mostra seu potencial, por exemplo, na identificação do câncer de cólon classificando o tipo de câncer de cólon com alta precisão e baixo custo computacional [5]. Por fim, as aplicações das redes neurais se estendem por várias disciplinas que incluem computação, ciência, engenharia, medicina, meio ambiente, agricultura, mineração, tecnologia, clima, negócios, artes e nanotecnologia [6].

O objetivo deste trabalho consiste em fazer uma análise comparativa de três dos principais modelos de classificação - Regressão logística, Máquina de vetores de suporte e K vizinhos mais próximos - na previsão dos pedidos de subvenção para pesquisas bem-sucedidos da Universidade de Melbourne na Austrália. Essa análise foi dividida em duas

[1]Linear discriminant analysis

[2]Ele assume que cada classe é normalmente distribuída, possuem a mesma variância e encontra os parâmetros por estimação

[3]Quadratic discriminant analysis

[4]K-Nearest neighbors

[5]Support vector machines

partes principais. Na primeira parte (seção II) os dados são analisado e pré-processados. Além disso, uma visão global das técnicas de análise e dos três modelos usados neste trabalho são apresentados. Já na segunda parte (seção III), os modelos treinados têm suas performances avaliadas com o conjunto de teste e, por fim, os resultados são apresentados juntamente com as conclusões que puderam ser tiradas da análise.

II. Métodos

O conjunto de dados usado neste trabalho são dados reais das candidaturas para subvenção de pesquisas de 2004 a 2008 fornecidos pela Universidade de Melbourne no contexto de uma competição[6] prever se um pedido de subsídio seria ou não aceito. Os dados originais podem ser encontrados no site Kaggle[7].

Ao todo são 8.707 amostras e um total de 249 características das equipes de pesquisa candidatas, dentre elas: o papel de cada indivíduo na equipe, além da data de nascimento, língua materna, escolaridade, nacionalidade, número de subsídios anteriores bem-sucedidos (e malsucedidos) dentre outras informações de cada membro. Sem esquecer do domínio da pesquisa, objetivos socio-econômicos, data da submissão e valor almejado. É importante destacar que os dados brutos fornecidos não são adequados para a modelagem e portanto são necessárias algumas adaptações. Um versão modificada contando com 1.882 preditores é proposta em [1]. No entanto, nesse novo conjunto como parte dos dados são zero por causa das variáveis *dummy*, existem preditores altamente correlacionados e muitos outros com faltam alguns dados um novo conjunto reduzido foi então proposto contando com apenas 252 preditores. Esse último é o conjunto usado nas análises deste trabalho.

Antes de continuarmos é interessante destacar que o criação de tal classificador é interesse para as universidades em geral pois ele possibilitaria que menos tempo seja desperdiçado em pedidos que dificilmente serão bem-sucedidos. Outrossim, este estudo pode ajudar as universidades a entender os fatores que são mais relevantes na previsão do sucesso de uma candidatura.

Uma vez apresentados os objetivos desse trabalho e o conjunto de dados utilizados, na seção II-A é feita uma análise mais detalhadas dos dados e um pré-processamento dos dados é realizado. Em seguida na seção II-B, são apresentadas as métricas e técnicas utilizadas para a avaliação dos modelos ao longo do trabalho, essas métricas são de grande importância na comparação dos modelos. Nas seções seguintes (II-C, II-D e II-E), os três modelos que estão sendo comparadas são apresentados e as principais ideias por trás cada uma deles são exploradas.

A. Análise e pré-processamento dos dados

Como os dados que contém 252 preditores já escolhido com o objetivo de não possuir preditores altamente correlacionados nem preditores com variância próxima de zero, o que poderia afetar fortemente a performance de alguns modelos em particular a Regressão Logística. No entanto, os dados não estão centrados nem escalonados o que não é recomendado para modelos como o KNN que depende muito da distancia entre os dados. Além disso, é possível notar também uma certa assimetria em alguns preditores. Esses problemas serão resolvidos conforme a necessidade de cada modelo e portanto, o tipo de pré processamento utilizado por cada modelo será especificado quando o ajustamento dos dados for realizado.

Uma análise mais detalhada dos dados pode ser encontrada no seção 12.1 da obra [1].

B. Métricas para medir o desempenho dos modelos

A métrica principal adotada neste trabalho é a **matriz de confusão**. Ela consiste em uma ferramenta para medir o desempenho de um modelos de classificação, verificando com que frequência suas previsões são precisas em relação à realidade e permitindo uma melhor compreensão dos tipos de erros cometidos.

A partir dos resultados e previsões esperados, a matriz indica o número de previsões corretas e incorretas para cada classe organizadas de acordo com a classe prevista. Cada linha da tabela corresponde a uma classe prevista e cada coluna corresponde a uma classe real.

Nas linhas sob as classes reais, as previsões são inseridas. Elas podem ser a indicação correta de uma previsão positiva como ”verdadeiro positivo”(VP) e uma previsão negativa como ”verdadeiro negativo”(VN) ou uma previsão positiva incorreta como ”falso positivo”(FP) e uma previsão negativa incorreta como ”falso negativo”(FN).

Uma vez criada a a matriz de confusão, existem alguns conceitos que ajudam a entender melhor as informações contidas nela, são eles

$$\textbf{Precisão} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}$$

que diz quanto de forma geral quanto o modelo acertou das previsões.

$$\textbf{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

que indica qual proporção de positivos foram identificados corretamente. Em outras palavras, quão bom o modelo é na previsão de pedido de subsídios bem-sucedidos.

$$\textbf{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

que indica qual proporção de negativos foram identificados corretamente. Em outras palavras, quão bom o modelo é na previsão de pedido de subsídios recusados. As duas últimas são importantes pois os custos dos erros de positivos e negativos são geralmente diferentes.

Uma outra métrica importante é a curva ROC (Receiver Operating Characteristic Curve) é um gráfico que mostra o

[6]Competição realizada em 2011 com um prêmio de US$5.000 para o vencedor

[7]https://www.kaggle.com/c/unimelb

desempenho de um modelo de classificação em todos os limites de classificação. Esta curva mostra a taxa de verdadeiros positivos em função da taxa de falsos positivos. A diminuição do valor do limiar de classificação permite que mais elementos sejam classificados como positivos, o que aumenta o número de falsos positivos e de verdadeiros positivos.

Por fim, temos a AUC[8] que significa "área sob a curva ROC". Esse valor mede toda a área bidimensional abaixo da curva ROC inteira de (0,0) a (1,1). A AUC fornece uma medida agregada de desempenho para todos os limiares de classificação possíveis. Quanto mais próxima de 1 é a AUC melhor é o modelo.

C. Regressão Logística

A Regressão Logística é um método supervisionado muito parecido com a Regressão Linear dando um resultado continuo a partir de uma combinação linear do(s) preditor(es). No entanto, ela possui algumas particularidades pois seu resultado deve estar contido no intervalo entre 0 e 1 indicando a probabilidade de ocorrência de um determinado evento. O objetivo é então encontrar uma relação entre $p(X) = Pr(Y = 1|X)$ e X, ou seja, a relação entre a probabilidade de uma observação pertencer a classe $Y = 1$ dada a ocorrência do preditor X na observação (o caso multi preditores será abordado posteriormente).

A restrição de que $p(0)$ deve estar entre 0 e 1 implica que a relação entre essas duas partes não pode ser linear, pois isso faria com que os valores de $p(0)$ estivessem no intervalo de $-\infty$ a $+\infty$. Uma solução para esse problema seria o uso da função logística que dar uma saída entre 0 e 1 para todos os valores de X, essa função tem a seguinte forma

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

Assim como na regressão linear é necessário ajustar encontrar os valores para os coeficientes que melhor ajusta o modelo aos dados. Infelizmente aqui a técnica dos mínimos quadrados não é aplicável visto que os resultados são classes e não números reais. No entanto, é possível aplicar a técnica da máxima verossimilhança para a estimar os coeficientes β_0 e β_1.

Na posse de todos os coeficientes do modelo é possível fazer previsões das classes de novas amostras. Contudo, a função logística fornece como resultado a probabilidade de uma dada amostra pertencer a uma determinada classe e não diretamente a classe em questão. Para realizar essa passagem de uma probabilidade ao pertencimento a uma classe ou a outra pode-se estabelecer um limiar acima do qual a amostra pertence a uma classe e caso contrario perceber à outra classe. A escolha desse limiar está diretamente relacionada as métricas sensibilidade e especificidade e é o motivo da existência da curva ROC.

Uma vez apresentada a teoria para um preditor a generalização para o caso com muitos preditores se torna relativamente simples e generalizando a equação 1 temos

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_P X_P}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_P X_P}} \tag{2}$$

onde $X = (X_1, ..., X_P)$ representam P preditores e a estimação dos coeficientes β_0, ..., β_P é feita também por meio da máxima verossimilhança.

É fácil notar nas equações (1) e (2) que não existe uma relação linear entre $p(X)$ e X, porém manipulando um pouco a equação (1) é possível obter

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \tag{3}$$

onde é possível perceber uma relação linear entre X e o termo do lado esquerdo da equação que é conhecido como 1881. O interessante aqui é que mesmo tendo se distanciado um pouco da regressão linear, aqui encontramos uma relação linear implícita muito semelhante a regressão linear.

O potencial do modelo de regressão logística será explorado na prática na seção III-A na solução do problema da universidade de Melbourne.

D. Máquinas de vetores de suporte

Máquinas de vetores de suporte do inglês *Support Vector Machines* (SVM) são uma classe de modelos estatísticos desenvolvidos pela primeira vez em meados da década de 1960 por Vladimir Vapnik e que vêm evoluindo para uma das ferramentas de aprendizado de máquina mais flexíveis e eficazes disponíveis [1].

O classificador SVM é amplamente utilizado devido à sua alta precisão bem como capacidade de lidar com dados de alta dimensão, como por exemplo expressão de genes, e flexibilidade na modelagem de diversas fontes de dados [8].

Ele consiste em um modelo de aprendizado supervisionado, com o objetivo de classificar determinado conjunto de dados que são mapeados para um espaço de características multidimensional. De forma generalizada o SVM realiza a separação de um conjunto de objetos com diferentes classes.

Os SVMs pertencem à categoria geral dos métodos do kernel, que é um algoritmo que depende dos dados apenas por meio de produto escalar. Quando esse é o caso, o produto escalar pode ser substituído por uma função do kernel que calcula um escalar no espaço de recursos possivelmente com alta dimensão, tendo assim a capacidade de gerar limites de decisão não lineares usando métodos projetados para classificadores lineares.

Vapnik definiu uma métrica alternativa chamada margem onde a margem é a distância entre o limite de classificação e o ponto de ajuste de treinamento mais próximo. Na terminologia SVM, a inclinação e interceptação da fronteira que maximiza a buffer entre o limite e os dados é conhecido como o classificador de margem máxima.

Suponha que tenhamos um problema de duas classes, o classificador de margem máxima cria um valor de decisão $D(x)$ que classifica amostras, se $D(x) < 0$, preveríamos que uma amostra fosse da classe 1, caso contrário, classe 2.

[8]Area under the ROC Curve

Para uma amostra desconhecida u, a equação de decisão pode ser escrita de forma semelhante à função discriminante linear que é parametrizada em termos de interceptação e declives.

$$D(\boldsymbol{u}) = \beta_0 + \boldsymbol{\beta'u} = \beta_0 + \sum_{j=1}^{P} \beta_j u_j \tag{4}$$

Esta equação funciona do ponto de vista dos preditores e pode ser transformada para que o classificador de margem máxima possa ser gravado em termos de cada ponto de dados na amostra.

$$D(\boldsymbol{u}) = \beta_0 + \sum_{i=1}^{n} y_i \alpha_i \boldsymbol{x_i'u} \tag{5}$$

com $\alpha_i > 0$.

A função de previsão é apenas uma função das amostras do conjunto de treinamento que estão mais próximas do limite e são previstas com a menor quantidade de certeza. Como a equação de previsão é suportada apenas por esses pontos de dados, o classificador de margem máxima é geralmente chamado de máquina de vetores de suporte.

E. K vizinhos mais próximos - KNN

O método dos K vizinhos mais próximos é um dos métodos de classificação mais simples. Ele consiste em um método não paramétrico que armazena as observações do conjunto de aprendizagem para a classificação dos dados do conjunto de testes.

Na verdade, esse algoritmo é qualificado como aprendizagem preguiçosa porque não aprende nada durante a fase de treinamento. Para prever a classe de um novo dado de entrada, ele procurará seus vizinhos K mais próximos (usando a distância euclidiana, por exemplo) e escolherá a classe da maioria dos vizinhos. Com um pouco mais de rigor matemático [7], dado um inteiro K positivo e uma observação de teste x_0, o classificador KNN primeiro identifica um conjunto $\mathcal{N}_0$ contendo os K vizinhos nos dados de treinamento mais próximos de x_0. Em seguida, ele estima a probabilidade condicional para a classe j como a fração de pontos em $\mathcal{N}_0$ cujos valores de resposta são iguais a j. Temos então

$$\Pr\left(\mathrm{Y} = j \mid \mathrm{X} = x_0\right) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j) \tag{6}$$

onde $I(y_i = j)$ = 1 se $y_i = j$ e 0 caso contrario. Por fim, a observação x_0 é atribuída a classe com maior probabilidade.

Vale destacar que, apesar de simples de entender e não precisar de nenhuma suposição sobre dados muitas vezes o KNN pode produzir classificadores que estão surpreendentemente próximos do classificador ideal de Bayes.

A escolha do valor de K é tarefa muito importante e está diretamente relacionada com a boa performance do modelo. Valores muito pequenos de K pode dar origem a modelos excessivamente flexível que possuem baixo viés e uma alta variância, podendo assim não apresentar bons resultados com o conjunto de teste. Por outro lado, valores muito grandes de K tornam o classificador menos flexível apresentando uma baixa variância e um alto viés. Diante desse trade-off entre viés e variância a validação cruzada pode ajudar na escolha de K.

Dentre os pontos negativos desse modelo temos a alta necessidade de memória, já que todos os dados de treinamento devem estar presentes na memória para calcular os vizinhos K mais próximos. Além disso, em alguns casos ele pode se mostrar sensível a características irrelevantes e também ele é sensível à escala dos dados uma vez que estamos a calcular a distância até aos pontos K mais próximos, ou seja, uma padronização dos dados é necessária. Por fim, ele não funciona bem com dados dimensionais elevados pois com um grande número de dimensões, torna-se difícil para o algoritmo calcular a distância em cada dimensão.

Na seção III-C são apresentados os resultados desse modelo na predição dos pedidos de subsídios bem-sucedidos.

III. Resultados e conclusões

Após a apresentação dos modelos que serão analisados e das métricas que serão usadas, nesta seção os três métodos são testados de forma prática e os resultados obtidos são discutidos nas seções III-A, III-B e III-C. A comparação dos resultados e as conclusões são realizadas na seção III-D.

A. Regressão Logística

Para a modelagem utilizando a Regressão Logística foi utilizado inicialmente o conjunto completo com 1881 preditores e 8190 amostras para o realizar o ajuste o modelo. O conjunto completo foi escolhido inicialmente para ser possível a realização de uma comparação com o conjunto de preditores reduzidos, visto que a regressão linear é beneficiada pela a remoção de preditores com variância próximo a zero. Os resultados obtidos para o conjunto completo são mostrados na matriz de confusão na Tabela I.

		Classe esperada	
		Aprovado	**Recusado**
Classe predita	**Aprovado**	147	49
	Recusado	42	280

Tabela I: Matriz de confusão da Regressão Logística usando 1881 preditores

De acordo com os resultados, utilizando todos os 1881 preditores, o modelo obteve uma precisão de 82,4% na classificação, assim como, uma sensibilidade de 77,7% e uma especificidade de 85,1%, mostrando que o modelo consegue um resultado melhor na classificação de pedidos recusados. Para o conjunto completo, também foi encontrado uma AUC de 81,4%. Para comparação com o conjunto completo, foi realizado um segundo treino utilizando apenas 252 preditores, dos quais foram os que restaram da remoção de preditores com variância próxima a zero no conjunto completo. Para o treino com o conjunto reduzido, podemos observar a matriz de confusão da Tabela II.

		Classe esperada	
		Aprovado	**Recusado**
Classe predita	**Aprovado**	150	39
	Recusado	46	283

Tabela II: Matriz de confusão da Regressão Logística usando 252 preditores

		Classe esperada	
		Aprovado	**Recusado**
Classe predita	**Aprovado**	146	43
	Recusado	50	279

Tabela III: Matriz de confusão do SVM utilizando o kernel polynomial

De acordo com a matriz de confusão utilizando a redução de preditores, observamos uma precisão de 83.6%, assim como, uma sensitividade de 79,4% e uma especificidade de 86%. O que indica que a redução de preditores trouxe uma certa melhora na precisão geral do modelo e com isso a sua capacidade de classificar pedidos aceitos e recusados. Também mostrou uma melhora na AUC com 82,7%.

B. Support Vector Machines

Para a implementação do SVM é necessário a configuração de alguns parâmetros, um desses valores valores é o custo, que são usados para penalizar o número de erros, assim como o custo o modelo do Kernel deve ser escolhido, e essa combinação controla a complexidade do algorítimo, e devem ser ajustados adequadamente para que o modelo não se ajuste ao ponto de causar um overfit aos dados de treinamento.

Para saber quais parametros temos uma melhor taxa de acertos, fizemos a comparação usando kernel diferentes.

A partir da observação dos métodos utilizados através da matriz de confusão para ambos os métodos, percebe-se que o método polynomial, tem uma melhor classificação. O custo utilizado teve o valor de C = 1. Temos um total de 518 amostras, sendo 189 para o caso de sucesso e 329 para o caso de não sucesso, com isso teve-se uma precisão de 74% para os casos de sucesso e total de 87% para os casos de não sucesso. Com isso podemos perceber que o método polynomial consegue classificar de formar mais acretiva em relação ao radial basis function.

		Classe esperada	
		Aprovado	**Recusado**
Classe predita	**Aprovado**	139	50
	Recusado	62	267

Tabela IV: Matriz de confusão do SVM utilizando o kernel radial basis function

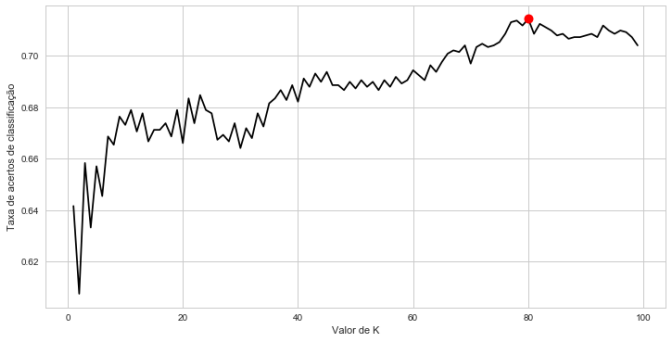

Figura 1: Comparação da precisão do modelo KNN para K variando de 1 a 100 usando 252 preditores

C. K vizinhos mais próximos - KNN

Como citado na seção II-E, existem dois pontos principais que devem ser considerados na modelagem com o KNN: a padronização dos dados e a escolha do número de vizinhos K.

Portanto, primeiramente, os dados de treino e teste com 252 preditores foram centrados e escalonados. Em seguida, para a escolha do K, as 8190 observações de treino foram divididas em 6633 observações (candidaturas pré 2008) para servirem como dados de referencia para o modelo e 1557 observações (75% das candidaturas de 2018) que foram usadas como suporte na escolha do K. Em resumo, a escolha do melhor valor de K se deu da seguinte forma: um conjunto de valores de K foram testados e para cada um deles foi feita uma predição sobre o conjunto de suporte com base no conjunto de referencia, o resultado obtido foi então apresentado como a percentagem de acertos de classificação. Na Figura 1 vemos os resultados para K variando de 1 a 100, onde K = 80 foi o valor com o qual o modelo apresentou a maior precisão (71%) na classificação dos dados de suporte.

Uma vez encontrado o melhor valor para K, usamos ele para testarmos a performance do modelo sobre um conjunto de dados ainda não utilizados composto de 518 observações (25% das candidaturas de 2018). Os resultados obtidos são resumidos na matriz de confusão V.

		Classe esperada	
		Aprovado	**Recusado**
Classe predita	**Aprovado**	108	63
	Recusado	81	266

Tabela V: Matriz de confusão do modelo KNN usando 252 preditores

Segundo a tabela V o modelo obteve uma precisão de 72.2%. Além disso, foi obtida uma sensibilidade de 57.1% e uma especificidade de 80.9% o que indica que o modelo é mais confiante na classificação de pedido recusado do que de pedidos aprovados. Isso é um bom sinal já que é desejado ter uma maior certeza ao recusar um pedido. Por fim, temos uma AUC de 0.69.

Apesar de uma precisão ligeiramente superior a taxa de não informação (67%), observa-se a precisão e principalmente a sensibilidade e a AUC estão com valores relativamente baixos. É importante notar a dificuldade do KNN em trabalhar em espaços de grandes dimensões, neste caso 252 dimensões. A fim de melhorar o desempenho do modelo, um conjunto de preditores ainda mais reduzido foi usado. Desta vez, foram escolhido apenas 19 preditores identificados como os mais relevantes na análise por outros algoritmos apresentados em [1]. Refazendo toda a análise, inclusive uma nova escolha do K, obtemos sobre o conjunto de teste os resultados apresentados na matriz de confusão VI.

		Classe esperada	
		Aprovado	**Recusado**
Classe predita	**Aprovado**	156	58
	Recusado	33	271

Tabela VI: Matriz de confusão do modelo KNN usando 19 preditores

Os novos resultados são significativamente melhores, com uma precisão de 82.4%, sensibilidade de 82.6%, especificidade de 82.4% e AUC de 0.825. Esse resultado se torna bastante interessante dada a simplicidade do KNN e a velocidade da computação ao se usar apenas 19 preditores. Na próxima seção é feita uma comparação com os demais modelos analisados neste trabalho.

D. Conclusões

Como podemos ver na tabela VII os resultados dos três modelos foram bastante próximo, sendo o KNN o que conseguiu um melhor equilíbrio entre as quatro métricas utilizadas. No entanto, um valor de especificidade mais elevado é desejado visto que é necessário ter mais confiança ao recusar um pedido do que ao dizer que possivelmente ele será aprovado.

Embora com resultados semelhantes alguns métodos acabam levando vantagem por sua simplicidade e velocidade de análise de novas amostras, como por exemplo no caso do KNN que utiliza apenas o 11 vizinhos mais próximos.

Por fim, concluímos que apesar de um pouco distante do AUC de 0.968, alcançado pelo campeão do competição, os resultados obtidos são significativos e podem contribuir positivamente na tomada de decisões das universidades. Uma análise mais detalhada sobre o tempo de cálculo de cada um deles pode ajudar como um critério de desempate.

	R. Logística	**SVM**	**KNN**
Precisão	0.836	0.82	0.824
Sensibilidade	0.794	0.773	0.826
Especificidade	0.86	0.848	0.824
AUC	0.827	0.81	0.825

Tabela VII: Comparação dos resultados

REFERÊNCIAS

[1] Kuhn, Max and Kjell. Johnson. 2013. Applied Predictive Modeling. New York: Springer.

[2] Herman, William H. (2002/11/01). A Multivariate Logistic Regression Equation to Screen for Diabetes. , 25, 1999-.

[3] A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval,"2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-6.

[4] Nurhanna, A.A. Othman, M.F.. (2017). Multi-class support vector machine application in the field of agriculture and poultry: A review. 11. 35-52.

[5] Pratiwi, Nor Magdalena, Rita Fuadah, Yunendah Saidah, Sofia. (2019). K-Nearest Neighbor for colon cancer identification. Journal of Physics: Conference Series. 1367. 012023. 10.1088/1742-6596/1367/1/012023.

[6] Arshad, Humaira (). State-of-the-art in artificial neural network applications: A survey. , 4, e00938-.

[7] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning with Applications in R, Springer (2017)

[8] B. Sch¨olkopf, K. Tsuda, and J.P. Vert, editors. Kernel Methods in Computational Biology. MIT Press series on Computational Molecular Biology. MIT Press, 2004