Sistemática para previsão de desempenho discente balizada por técnicas de regressão

Maximilianno Meneghini (UFRGS) – *maxmeneghini125@gmail.com*

Michel José Anzanello (UFRGS) – *anzanello@producao.ufrgs.br*

Resumo

O ingresso a universidades no Brasil dá-se principalmente através de uma prova unificada, que demanda dos candidatos aptidão em múltiplas áreas de conhecimento. Visando a melhorar suas chances de entrada, alunos recorrem, com frequência, aos chamados cursos preparatórios, ou pré-vestibulares, como são popularmente conhecidos. Este artigo propõe uma sistemática para previsão de desempenho discente na prova nacional unificada, aplicando ferramentas de regressão e análise multivariada a um banco de dados coletados de uma empresa de educação online de Porto Alegre. Número de aulas assistidas, exercícios resolvidos e simulados submetidos pelos alunos estão entre as variáveis analisadas. Foram testadas três técnicas de regressão para realização das predições: Regressão Linear Múltipla (MLR), Regressão por Componentes Principais (PCR) e Regressão por Mínimos Quadrados Parciais (PLS). Através da análise do erro percentual de predição gerado pelas ferramentas utilizadas em uma validação cruzada, concluiu-se que a MLR possui a melhor capacidade preditiva, com erro percentual médio de 13,04%. Observando as equações construídas pela metodologia, foi possível, ainda, identificar tendências entre as variáveis e o impacto causado por elas no desempenho final dos estudantes.

Palavras-chave: educação online, cursos preparatórios, ferramentas de análise multivariada, Regressão Linear Múltipla (MLR), Regressão por Componentes Principais (PCR), Regressão por Mínimos Quadrados Parciais (PLS).

Abstract

The entrance to universities in Brazil happens mainly through a unified exam which demands aptitude in multiple fields of knowledge from the applicants. In an attempt to improve their chances, students frequently appeal to the so called preparatory courses, or prep courses, as they are popularly referred to. This article proposes a systematic for student performance prediction in the national unified exam, using regression and multivariate analysis tools over a set of data collected from an online education company based in Porto Alegre. Number of classes attended, exercises solved and simulated tests submitted by students, clients of the company, are among the analyzed variables. Three regression techniques were tested in order to perform the predictions: Multiple Linear Regression (MLR), Principal Component Regression (PCR) and Partial Least Squares Regression (PLS). Through the analysis of the percentage prediction error generated in the application of the tools in a cross validation, the MLR was found to have the best predictive capacity with a mean percentage error of 13,04%. Observing the equations built through the method, it was also possible to identify tendencies among the variables as well as the impact caused by them in the final performance of the students.

Keyword: online education, preparatory courses, multivariate analysis tools, Regression (MLR), Principal Component Regression (PCR), Partial Least Squares Regression (PLS).

1. Introdução

O atual modo de ingresso de estudantes em universidades brasileiras apoia-se, em grande maioria, em uma prova nacional única. De tal forma, é imprescindível o desenvolvimento de técnicas e métodos de ensino, buscando gerar melhores oportunidades de ingresso aos candidatos às vagas. Visto que, nesta realidade, uma grande quantidade de estudantes não se encontra apropriadamente preparada para a realização da prova nacional (tipicamente devido a diferenças de ensino entre as escolas brasileiras e diversos fatores socioeconômicos), muitos dos mesmos recorrem a cursos de ensino preparatório. Paralelamente, uma nova forma de preparação tem se tornado cada vez mais popular: os cursos preparatórios online. Isso se deve, principalmente, ao recente desenvolvimento tecnológico, à maior agilidade e praticidade no serviço e à impossibilidade financeira de contratar um curso presencial.

O uso dos denominados cursos preparatórios online proporciona inúmeros benefícios aos estudantes que os atendem. Entre esses benefícios, pode-se citar a facilidade de acesso, uniformidade e concisão nos métodos de ensino, baixo custo em comparação aos cursos presenciais, flexibilidade de tempo e ritmo de aprendizado, dentre outros (BRINSON, 2015). Da mesma forma, tais cursos trazem potenciais desvantagens quando comparados aos presenciais, especialmente em relação à quase total impessoalidade do serviço. Esse nível de impessoalidade pode acarretar problemas ao aluno, como a não sincronização entre estudante e professor, impossibilidade de reconhecer as possíveis deficiências na formação educacional do aluno e a dificuldade na geração de um *feedback* preciso para o mesmo (EL-MAGBOUB *et al*., 2016).

Devido a esse caráter impessoal do serviço provido pelas empresas nos cursos preparatórios online, ainda há significativo nível de rejeição aos mesmos por parte de alunos em preparação para o exame nacional, preferindo cursos presenciais em busca de algum grau de personalização no serviço. Acredita-se, no entanto, que se analisados os dados que descrevem os perfis destes alunos, seria possível transformar tais dados em informações de alto valor, com potencial para a criação de um serviço mais pessoal, mesmo que em uma plataforma online. Dentre tais ações, seria possível ainda prever o desempenho dos estudantes no exame nacional, além de identificar as variáveis que mais contribuem para os melhores desempenhos.

Utilizando dados coletados de uma empresa de ensino online, fundada e baseada em Porto Alegre, Rio Grande do Sul, este artigo propõe a aplicação de ferramentas de análise multivariada sobre os dados que descrevem os perfis de alunos com vistas à previsão do desempenho dos mesmos nos exames de ingresso. Além disso, busca-se entender a relação entre os dados coletados e os resultados alcançados, de forma que possam ser identificadas as variáveis que possuem maior significância para a análise. Com isso, objetiva-se oferecer meios para que a empresa consiga aproximar-se de seus estudantes, suprindo a necessidade de tornar seu serviço mais pessoal e auxiliá-los a identificar e a trabalhar em suas dificuldades específicas.

Segundo Johnson e Wichern (2007), ferramentas de análise multivariada podem ser utilizadas tanto para predição quanto para análise de relação entre variáveis. Portanto, serão aplicados três métodos de análise multivariada para realização de predições de desempenho dos estudantes: (i) regressão linear múltipla, que utiliza uma equação linear para modelar a relação entre as variáveis de entrada e de saída; (ii) regressão PLS (*Partial Least Squares Regression*), utilizada para reduzir as variáveis de previsão a um conjunto relativamente menor de elementos não correlacionados para posterior predição; (iii) regressão por componentes principais, que utiliza um modelo similar ao de uma regressão linear, mas com variáveis independentes representadas por variáveis oriundas da Análise de Componentes Principais (PCA). Ao final dos testes, será possível selecionar o modelo mais adequado à modelagem do cenário em análise.

Finalmente, esse trabalho será dividido em cinco seções, além desta introdução: Referencial Teórico, Procedimentos Metodológicos, Resultados, Conclusões e Referências, Apêndices e Anexos.

2. Referencial Teórico

2.1. Contexto educacional

Esta seção do referencial teórico visa introduzir, revisar e contextualizar conceitos e aspectos educacionais abordados neste artigo, abordando informações relativas à educação à distância, educação online e cursos preparatórios, relevantes para o estudo desenvolvido.

2.1.1. Educação online

Educação à distância é um campo da área de educação com enfoque em pedagogia, tecnologia e planejamento de sistemas instrutivos, efetivamente incorporados no ato de

instruir o estudante, em que professor e aluno podem se comunicar de forma síncrona ou assíncrona (AL-ARIMI, 2014). De acordo com Kaplan e Haenlein (2016), define-se por educação à distância qualquer forma de transmissão de conhecimento em que aluno e professor não se encontrem presente no mesmo espaço físico e em que o material pedagógico é planejado e preparado por uma instituição de ensino. Inicialmente utilizada para resolver o problema de alcance do sistema de ensino convencional, a educação à distância veio a constituir um importante modo de ensino, sendo capaz de complementar o sistema educacional presencial ao utilizar recursos tecnológicos que facilitam a comunicação entre estudantes e professores e, dessa forma, quebrando as barreiras de tempo e espaço (RODRIGUES *et al*., 2014).

A evolução histórica da educação à distância pode ser dividida em três principais períodos, cada um correspondente ao veículo de mídia utilizado: materiais impressos, televisão e a Internet (KAPLAN e HAENLEIN, 2016). O último período citado, ao ser utilizado com o propósito de ensino, desencadeou a vasta utilização da educação online. Rodrigues *et al*. (2014) conduziram um estudo em diferentes unidades da Universidade de Anhanguera, envolvendo cerca de 160 alunos e 30 professores relacionados a cursos totalmente ou com suporte online, em que foi constado que 80% dos alunos se utiliza de meios de comunicação online para responder questões, e que 95% utiliza livros a artigos virtuais como fonte de pesquisa (denotando alto índice de aceitação por parte dos estudantes aos métodos de ensino).

Considerando vantagens e desvantagens da educação online em relação ao método presencial convencional, El-Magboub *et al*. (2016) apresentaram os resultados de um experimento em que grupos de discussão realizariam suas reuniões online, através de videoconferências, mediadas por professores. O estudo ressaltou, além da facilidade do uso da tecnologia por parte dos estudantes, a conveniência do método online, em que os alunos não precisavam encontrar-se no mesmo espaço físico. Porém, o estudo mostrou uma queda na interatividade entre alunos e entre alunos e o professor, mesmo que as discussões fossem de forma síncrona, demonstrando o caráter mais impessoal do método de ensino.

Brinson (2015) e Al-Arimi (2014) acrescentam algumas das vantagens da modalidade, as quais incluem a flexibilidade de distância e tempo, ritmo de aprendizado determinado pelo próprio aluno, eficiência (tanto para a instituição quanto para o aluno) em tempo e custos, a melhor acessibilidade para deficientes físicos, conteúdo acessível de forma constante, fácil e

rápida e a possibilidade de um número massivo de participantes sem grande dificuldade de gerenciamento. Como desvantagens, os autores citam a sujeição aos recursos tecnológicos e infraestrutura, a impessoalidade do método, a diminuição da interação entre professores e estudantes, dificuldade de identificar deficiências no ensino e a submissão do conteúdo ao meio de comunicação (como aulas práticas, por exemplo).

2.1.2. Cursos preparatórios

Os cursos preparatórios, também chamados de cursos pré-vestibular, têm se tornado cada vez mais comuns para a tentativa de ingresso discente a universidades. Isso tem ocorrido principalmente pelo fato de muitos estudantes sentirem-se despreparados para a realização das provas de ingresso. Sparta e Gomes (2005) ressaltaram que os alunos provenientes de escolas públicas indicaram com maior frequência a opção de realização de cursos preparatórios ao finalizar o ensino médio, em contraste a alunos de escolas privadas, que indicaram com maior frequência a realização direta das provas de vestibular. Pode-se atribuir isso à preocupação com a qualidade da educação pública nos níveis fundamental e médio. Porém, esse estudo desconsidera a renda familiar dos estudantes, levando a crer que muitos dos estudantes provenientes do sistema de ensino público não são capazes de pagar pelos cursos preparatórios particulares, devido a seus elevados custos.

Os cursos pré-vestibular têm impacto positivo no desempenho dos alunos, funcionando como um diferencial principalmente para estudantes oriundos de famílias de maior renda, devido aos custos de um curso particular (SAMPAIO *et al*., 2011). Sabe-se ainda da existência de algumas variantes nos cursos preparatórios. Uma destas constitui-se dos cursos populares ou comunitários, que possuem o mesmo formato dos cursos preparatórios particulares, mas com enfoque em alunos com renda familiar mais baixa. Outro modelo que tem se tornado popular, principalmente nos últimos anos, é o de cursos preparatórios online. Este formato de curso incorpora, nos cursos preparatórios, os aspectos dos métodos de ensino online, citados na seção acima.

2.2. Ferramentas de análise multivariada

Ferramentas de análise multivariada são métodos que buscam analisar as relações entre conjuntos de variáveis ou sistemas. Existem diversos métodos, que podem ser classificados de várias maneiras, desde objetivos até dependência de variáveis. Entre as aplicações e objetivos mais comuns das ferramentas de análise multivariada, estão a redução de dados e simplificação estrutural, ordenação e agrupamento de variáveis, investigação de

dependência entre variáveis, predição, formulação e teste de hipóteses (JOHNSON e WICHERN, 2007).

Nesta seção, serão descritos os métodos de Regressão Linear Múltipla (MLR), Análise de Componentes Principais (PCA) e Regressão por Mínimos Quadrados Parciais (PLS), bem como aplicações dos mesmos.

2.2.1. Regressão Linear Múltipla (MLR)

Esta ferramenta considera a relação linear entre uma variável dependente *y*, também chamada variável de resposta, e múltiplas variáveis independentes *x*, também chamadas de variáveis de previsão (RENCHER, 2002). A regressão linear múltipla apresenta uma lógica similar à de uma regressão linear simples, em que existe apenas uma variável dependente e uma independente, porém com duas ou mais variáveis de entrada (FÁVERO, 2009).

Rencher (2002) apresentou o modelo da seguinte forma. Cada variável de resposta *y* é expressa como a função linear de múltiplas variáveis de previsão *x*, em adição a um erro aleatório ε, em uma amostra de *n* observações:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_q x_{1q} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_q x_{2q} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_q x_{nq} + \varepsilon_n \end{aligned} \tag{1}$$

Neste modelo, os chamados coeficientes de regressão são denotados nas equações pelos β's, enquanto que o número de variáveis independentes *x* presentes na mesma é dada por q. É possível sumarizar essas equações como segue:

$$y_n = \sum_{i=1}^{q} \beta_i \, x_{ni} + \varepsilon_n \tag{2}$$

Usando uma notação matricial, as equações acima, que consideram uma amostra de *n* observações, podem ser escritas mais concisamente como:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{3}$$

Ou na forma condensada equivalente:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (4)$$

Geladi e Kowalski (1986) apontam três casos que uma regressão linear múltipla pode seguir:

(i) $q > n$. O número de variáveis é maior do que o de amostras, o que implica que existem infinitas soluções para o vetor $\boldsymbol{\beta}$, sendo que todas elas satisfazem a equação.

(ii) $q = n$. O número de variáveis e de amostras é igual, que apresenta uma única solução para o vetor $\boldsymbol{\beta}$, considerando que a matriz X não é deficiente.

(iii) $q < n$. O número de variáveis é menor do que o de amostras, o que não permite uma solução exata para $\boldsymbol{\beta}$. Porém, é possível atingir uma solução ao minimizar-se o vetor residual.

Uma forma de se realizar o que foi descrito em (iii) é através do método dos mínimos quadrados, que Rencher (2002) descreveu como:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (5)$$

Nesta equação, $\widehat{\boldsymbol{\beta}}$ representa o valor esperado do vetor $\boldsymbol{\beta}$ ao minimizar-se o comprimento do vetor residual. Considera-se, também, que $n > q + 1$ e que nenhum valor pertencente a matriz **X** é uma combinação linear de outros *x*'s.

O método de regressão linear múltipla tem sido utilizado em diversos campos de estudo e com distintos propósitos. Uyanik e Güler (2013) utilizaram uma regressão linear múltipla com dados de estudantes da Universidade de Sakarya, Turquia, com o objetivo de analisar quais variáveis afetavam de forma mais significativa a nota dos alunos nos testes de entrada para cargos governamentais. Vivaracho-Pascual *et al*. (2016) utilizaram o modelo em conjunto com parâmetros robustos para estimar com precisão o ponto de decisão de leitores biométricos e ressaltaram, além do desempenho, a simplicidade e facilidade de implementação e compreensão do método. Dan e Shi (2013) coletaram dados de um cruzamento entre ruas em Shangai, China, visando estimar o atraso de atravessamento dos pedestres, bem como as variáveis que mais impactam neste atraso, descobrindo que a maior causa era a grande quantidade de carros virando à direita – que não necessitam da abertura do sinal – ao obter um alto índice de correlação. Herrig *et al*. (2015) utilizaram dados coletados

das águas do Rio Lahn, Alemanha, juntamente com um método de regressão linear múltipla para predizer a qualidade da água em determinadas áreas do rio e atingindo um modelo que explicava grande parte dos casos e eficiente em custos. Por fim, Turóczy e Marian (2012) aplicaram uma regressão linear em uma indústria cerâmica para descobrir como o lucro era influenciado por cinco variáveis independentes: os autores constataram que apenas três das mesmas influenciavam significativamente nas predições.

2.2.2. Análise de Componentes Principais (PCA)

Uma das maiores desvantagens de uma MRL é lidar com a presença de multicolinearidade. Porém, pode-se eliminar essa colinearidade ao transformar-se a matriz original de observações, descrita na seção 2.2.1, em um conjunto de dados não-correlacionados (FACCHIN, 2005). O propósito da aplicação de uma PCA é a transformação e redução dimensional de dados. Tal técnica permite transformar variáveis originalmente correlacionadas em um novo conjunto de variáveis não-correlacionadas, denominadas componentes principais (*principal components* - PC's) (HE *et al*. 2011). O primeiro PC é a combinação linear do conjunto original de variáveis com maior variância, ou seja, procura-se por uma dimensão em que as observações estejam maximamente separadas entre si. Por sua vez, o segundo PC é a combinação linear com maior variância em uma direção ortogonal a do primeiro PC, e assim por diante (RENCHER, 2002).

Considere um conjunto de vetores de q observações $\mathbf{x_1}$, $\mathbf{x_2}$, ..., $\mathbf{x_q}$ que formam um aglomerado de pontos em um espaço de n dimensões. Se as variáveis x_1, x_2, ..., x_n em $\mathbf{x}$ são correlacionadas, o aglomerado elipsoidal de pontos não está orientado paralelamente aos eixos representados por x_1, x_2, ..., x_n e, como objetiva-se encontrar os eixos naturais das observações com origem no vetor média $\bar{\mathbf{x}}$, é necessário transmitir a origem para $\bar{\mathbf{x}}$ e então rotacionar os eixos (RENCHER, 2002). Ao final deste processo, as novas variáveis serão não-correlacionadas. O vetor média $\bar{\mathbf{x}}$ pode ser obtido através da seguinte equação (HE *et al.*, 2011):

$$\bar{\mathbf{x}} = \frac{1}{q} \sum_{i=1}^{q} \mathbf{x}_q \tag{6}$$

De acordo com Rencher (2002), é possível rotacionar os eixos multiplicando cada vetor $\mathbf{x_i}$ por uma matriz ortogonal $\mathbf{A}$, que contém os pesos de cada variável no conjunto de observações. Através dessa multiplicação, cria-se um novo ponto $\mathbf{z_i}$ a partir de $\mathbf{x_i}$, tal que:

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_p' \end{pmatrix} \tag{7}$$

$$\mathbf{z}_i = \mathbf{A}\,\mathbf{x}_i \tag{8}$$

Como **A** é ortogonal, a distância em relação a origem continua inalterada ao fazer a transformação, como mostra a Equação (9).

$$\mathbf{z}_i'\,\mathbf{z}_i = (\mathbf{A}\,\mathbf{x}_i)'(\mathbf{A}\,\mathbf{x}_i) = \mathbf{x}_i'\mathbf{A}'\mathbf{A}\,\mathbf{x}_i = \mathbf{x}_i'\,\mathbf{x}_i \tag{9}$$

Ao final da transformação, os componentes principais alcançados são dados pelas variáveis $z_1 = \mathbf{a}_1'\mathbf{x}$, $z_2 = \mathbf{a}_2'\mathbf{x}$, ..., $z_n = \mathbf{a}_n'\mathbf{x}$ em $\mathbf{z} = \mathbf{A}\,\mathbf{x}$ ou, de forma genérica, $z_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n$.

É possível, através deste método, incorporar uma regressão linear em um modelo chamado Regressão por Componentes Principais (PCR). O objetivo é utilizar os componentes principais mais significativos encontrados a partir do método de PCA para obter uma relação linear estável com uma variável de resposta (FACCHIN, 2005). Assim, é possível utilizar a Equação (5), descrita na seção 2.2.1 deste artigo, para encontrar os coeficientes lineares a serem utilizados na Equação (4). As Equações (10) e (11) descrevem a forma alterada das Equações (5) e (4), respectivamente, em que **y** representa o vetor de variáveis dependentes.

$$\widehat{\boldsymbol{\beta}} = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y} \tag{10}$$

$$\mathbf{y} = \mathbf{z}\,\widehat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon} \tag{11}$$

O PCA e a PCR possuem aplicações em diversas áreas de estudo. He *et al*. (2011) utilizaram-se de uma PCA em um sistema de *multisite testing*, ou sistema de testes paralelos, e conseguiram evidências empíricas de que o modelo proposto era capaz de identificar instrumentos defeituosos no sistema, bem como propor uma ferramenta online de monitoramento de instrumentos. Faed *et al*. (2016) propuseram um modelo de manuseio inteligente de reclamações de clientes, utilizando PCA para análise dos dados, resultando na identificação eficaz de clientes chave, que possuíam pontos de vista de maior influência. Ranamukhaarachchi *et al*. (2017) aplicaram uma PCA, seguida de uma MLR, em uma pesquisa para estimar o potencial antioxidante de proteínas de soja hidrolisadas, identificando uma forte relação entre os componentes principais alcançados e as variáveis de resposta coletadas. Por fim, Marhaba *et al*. (2009) conduziram um estudo sobre caracterização de

matéria dissolvida em corpos de água e utilizaram um PCR para estimar um modelo individual de concentração de cada tipo de matéria, por necessitar de um número muito inferior de amostras em relação a uma MLR e pela não-correlação entre os componentes principais, obtendo resultados significativos.

2.2.3. Regressão PLS (*Partial Least Squares Regression*)

Partial Least Squares (PLS) é uma vasta classe de métodos utilizada para modelar relações entre conjuntos de variáveis observáveis por meio de suas variáveis latentes, compreendendo tarefas de classificação e regressão, bem como técnicas de redução dimensional e ferramentas de modelagem (ROSIPAL e KRÄMER, 2006). A regressão PLS tem por objetivo encontrar um conjunto de componentes, chamados vetores latentes, que executem uma decomposição simultânea das variáveis de predição e de resposta, buscando explicar a covariância entre esses conjuntos de variáveis (ABDI, 2010). O conceito permite a representação dos dados em gráficos de coeficientes, de resíduos e de pesos, em que a identificação de grupos, detecção de discrepâncias e análise de dados tornam-se tarefas visuais e intuitivas (ANDERSSON, 2009).

Ainda de acordo com Andersson (2009), geralmente há apenas uma variável de resposta em uma regressão PLS, representada pelo vetor **y**, ou objetiva-se modelar e otimizar o desempenho de cada uma das variáveis de resposta separadamente. Esse caso é geralmente denominado regressão PLS1, enquanto que é denominado PLS2 nos casos em que se utiliza uma matriz de resposta multivariada **Y**.

Helland (1988) descreve o modelo de forma simplificada, apresentado da seguinte forma. Considerando um conjunto de observações representado pela matriz $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_q)$ e um vetor de variáveis de resposta **y**, relacionados por uma equação similar à (4) e em que os vetores $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_q$ e **y** possuem *n* dimensões correspondentes às observações em *n* unidades. Considera-se, também, que esses vetores já foram escalonados, ou seja, deles já foram subtraídas suas médias $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \cdots, \bar{\mathbf{x}}_i$, $\bar{\mathbf{y}}$. Pode-se então, descrever a matriz de observações **X** como uma combinação linear como mostra a Equação (12).

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \cdots + \mathbf{t}_a\mathbf{p}_a' + \mathbf{E}_a \qquad (12)$$

Os vetores $\mathbf{t}_a$, de dimensão *n*, são as variáveis latentes; os vetores $\mathbf{p}_a$, de dimensão *k*, são os coeficientes e $\mathbf{E}_a$ é a matriz residual. Como a base de um método PLS é a relação entre

y e as variáveis latentes de **X**, pode-se descrever **y** de forma semelhante, como na Equação (13), abaixo, em que q_a são os coeficientes e $\mathbf{f}_a$ é o vetor residual.

$$\mathbf{y} = \mathbf{t}_1 q_1 + \mathbf{t}_2 q_2 + \cdots + \mathbf{t}_a q_a + \mathbf{f}_a \tag{13}$$

Em seguida, busca-se descrever $\mathbf{t}_a$ como a combinação linear de **X** e um vetor peso $\mathbf{w}_a$, de *k* dimensões, como mostra a Equação (14). Ainda, é desejável que $\mathbf{t}_a$ seja altamente correlacionado com **y**, e uma escolha razoável para tal é tornar cada componente $\mathbf{w}_{ia}$ proporcional à covariância entre $\mathbf{x}_i$ e **y**, de acordo com a Equação (15).

$$\mathbf{t}_a = \sum_{i=1}^{k} \mathbf{x}_i \mathbf{w}_{ia} = \mathbf{X}\mathbf{w}_a \tag{14}$$

$$\mathbf{w}_{ia} = \mathbf{x}_i' \mathbf{y}\,, \text{ ou } \mathbf{w}_a = \mathbf{X}'\mathbf{y} \tag{15}$$

Por fim, para que se atinjam os melhores valores para $\mathbf{y} = \mathbf{t}_a q_a + \mathbf{f}_a$, utiliza-se o coeficiente de regressão q_a, como na Equação (16), ao passo que para $\mathbf{x}_i = \mathbf{t}_a p_{ia} + \mathbf{e}_{ia}$ utiliza-se o coeficiente p_{ia}, como na Equação (17), para *i* = 1, ..., *k*.

$$q_a = \mathbf{y}'\mathbf{t}_a / \mathbf{t}_a'\mathbf{t}_a \tag{16}$$

$$p_{ia} = \mathbf{x}_i'\mathbf{t}_a / \mathbf{t}_a'\mathbf{t}_a\,, \text{ ou } \mathbf{p}_a = \mathbf{X}'\mathbf{t}_a / \mathbf{t}_a'\mathbf{t}_a \tag{17}$$

Após o procedimento, o número de fatores a serem retidos na equação final pode ser determinado por um processo de validação cruzada. Esse processo se dá pela separação dos dados em partes, seguida de calibragem do modelo em uma delas e validação do modelo na outra. O número de fatores é escolhido ao minimizar-se o erro estimado de predição.

Embora PLS tenha sido vastamente utilizada na área de quimiometria, a regressão possui aplicações em outras áreas de conhecimento. Borràs *et al.* (2016) conduziram um estudo para predizer descritores sensoriais de amostras de azeite de oliva a partir de provas sensoriais com participantes voluntários e métodos de análise de dados, incluindo uma regressão PLS, e conseguiram criar modelos explicativos para predição de alguns dos descritores, além de obterem resultados promissores para os demais. Aloglu *et al*. (2016) utilizaram uma regressão PLS para prever a atividade antioxidante total de uma família de plantas usando quatro diferentes representações de dados reunidos através de técnicas de cromatografia, permitindo que fosse determinada a melhor representação a ser utilizada. Por sua vez, Bian *et al*. (2016) criaram um modelo para determinação de conteúdo hidrocarbônico em amostras de óleo leve e diesel, coletadas através de espectroscopia ultravioleta e utilizando

um método de regressão PLS em combinação a outros métodos de decomposição de dados. Por fim, Teoh *et al*. (2016) utilizaram-se de uma regressão PLS, juntamente com uma ferramenta de prospecção de dados para identificar tendências de correlação entre metabólitos e fenótipos quantitativos de um indivíduo, permitindo validação de sua estratégia em otimização de fenótipos.

3. Procedimentos Metodológicos

A presente seção descreve a caracterização do método de pesquisa, em que se define a abordagem, natureza, objetivos e procedimentos do estudo. A seguir, detalha-se o método utilizado, descrevendo os passos desenvolvidos para elaborar a sistemática de previsão de desempenho discente.

Considerando a forma de abordagem, considera-se a pesquisa como quantitativa, pois as informações serão traduzidas em números para que sejam classificadas e analisadas. A natureza da pesquisa é aplicada, pois objetiva criar conhecimentos para uma aplicação prática, de modo a solucionar um problema específico proposto (DA SILVA e MENEZES, 2005). Os objetivos podem ser considerados tanto exploratórios como explicativos, pois a pesquisa visa criar familiaridade com o problema, bem como identificar fatores que influenciam o fenômeno, de modo a levantar hipóteses e encontrar as razões de ocorrência do mesmo. Por fim, do ponto de vista dos procedimentos técnicos, a pesquisa é caracterizada como experimental, pois busca-se o levantamento de variáveis e observação de resultados para formular um modelo (GIL, 2002).

O método de previsão de desempenho discente será dividido em quatro etapas, como mostra a Figura 1. Tais etapas são detalhadas na sequência.

Figura 1 – Etapas de desenvolvimento do método

Fonte: Autores.

3.1. Coleta de dados

A primeira etapa do método consiste na coleta dos dados disponíveis na base de dados da empresa. Estas informações compõem as variáveis, divididas em independentes e dependentes (ou variáveis de previsão e de resposta, respectivamente). Essas variáveis

descrevem fatores estudantis e sociais, como por exemplo, número de aulas assistidas e tipo de escola de proveniência, constituindo as variáveis de previsão. As variáveis de resposta referem-se ao desempenho final dos alunos na prova de ingresso ao ensino superior.

Dessa forma, gera-se uma matriz como na Tabela 1; são coletadas informações de *n* estudantes, descritos por *q* variáveis independentes e cinco variáveis dependentes, representando as cinco áreas de conhecimento da prova analisada.

Tabela 1 – Estrutura dos dados.

Fonte: Autores.

Observações	Desempenho Final					Variável 1	Variável 2	...	Variável q
Aluno 1	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	x_{11}	x_{12}	...	x_{1q}
Aluno 2	y_{21}	y_{22}	y_{23}	y_{24}	y_{25}	x_{21}	x_{22}	...	x_{2q}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
Aluno *n*	y_{n1}	y_{n2}	y_{n3}	y_{n4}	y_{n5}	x_{n1}	x_{n2}	...	x_{nq}

3.2. Validação cruzada

Após coletados e organizados, os dados devem ser divididos em dois grupos, para que as ferramentas de regressão multivariada possam ser aplicadas e validadas. Esses grupos são denominados treino e teste. O grupo de treino é utilizado para criação do modelo, sendo este aplicado no grupo de teste, para que seja possível verificar a aderência do modelo construído aos dados. A este processo dá-se o nome de validação cruzada.

A validação cruzada é realizada na aplicação de todas as ferramentas selecionadas, de forma a testar a capacidade preditiva dos modelos gerados por cada uma delas. O método *Holdout*, proposto por Kohavi (1995), utiliza dois terços dos dados como grupo de treino e um terço como grupo de teste. Devido à grande quantidade de dados presentes, são utilizados 75% dos dados como grupo de treino e os demais 25% como teste.

Os dados são separados através do algoritmo Kennard-Stone, método utilizado para realizar uma avaliação total da amostra de dados e selecionar os dados que devem ser utilizados na construção do modelo e na posterior validação do mesmo, criando uma separação uniforme de variáveis que melhor descreveriam os grupos (KENNARD e STONE, 1969). Assim, evita-se a geração de tendências nas porções de treino e teste, não favorecendo o método com uma partição específica.

Ao final deste processo, as variáveis independentes são normalizadas ao diminuir-se, do valor original das amostras, a média da variável em questão e em seguida dividindo-se o resultado pelo desvio padrão da mesma. Dessa forma, as variáveis têm magnitudes similares, o que possibilita a identificação de tendências nas equações finais a serem geradas pelos modelos. Em seguida, o processo acima descrito é novamente efetuado, visando a gerar um banco de dados secundário, com as variáveis normalizadas.

3.3. Aplicação de ferramentas multivariadas

Na sequência, o grupo de treino é modelado por cada uma das ferramentas de regressão multivariada descritas no referencial teórico: Regressão Linear Múltipla (MLR), Regressão por Componentes Principais (PCR) e Regressão por Mínimos Quadrados Parciais (PLS). Utilizando as variáveis independentes e dependentes do grupo de treino, cada uma dessas ferramentas deve gerar uma equação que permitirá prever o desempenho de alunos com base em variáveis independentes relativas à sua formação.

A Regressão Linear Múltipla deve satisfazer a Equação 4, representada por uma equação matricial em que o vetor **y** representa os resultados finais dos alunos, a matriz **X** representa os dados coletados referentes aos alunos, $\boldsymbol{\varepsilon}$ é o vetor de erro aleatório da equação e $\boldsymbol{\beta}$ representa os coeficientes de regressão. Analogamente, a Regressão por Componentes Principais deve satisfazer a Equação 11, encontrando os coeficientes de regressão alterados para os componentes principais descobertos. Por fim, a Regressão PLS deve encontrar os coeficientes relativos às variáveis latentes, representados na Equação 17, para que seja gerada uma matriz de coeficientes que possa ser multiplicada pela matriz de variáveis independentes originais **X**, resultando em uma equação geral similar à do MLR. A Regressão PLS pode ser operacionalizada de duas formas: aplicar o método apenas uma vez utilizando uma matriz de variáveis dependentes **Y**; ou utilizar múltiplos vetores de variáveis dependentes **y**, aplicando a ferramenta um número de vezes igual ao número de vetores. Ambas as formas serão testadas e avaliadas. Os mesmos procedimentos são executados, também, no banco de dados normalizados.

As equações resultantes, modeladas pelas ferramentas, são aplicadas nos dados presentes no grupo de teste, visando à validação de cada modelo. Assim, são geradas as previsões de notas finais para cada amostra e estas, por sua vez, são comparadas ao desempenho final original dos estudantes. Essa comparação gera um Erro Percentual Médio de Predição para cada matéria considerada e para cada ferramenta utilizada.

3.4. Análise dos resultados

Concluídas todas as etapas anteriores, o resultado final é uma equação geral para os dados coletados, representando as três ferramentas utilizadas, assim como o Erro Percentual Médio de Predição. Utilizando essas informações, é possível verificar se a aplicação dos modelos é viável, bem como identificar a ferramenta que apresenta os melhores resultados.

A seleção da ferramenta mais eficaz se dá pela análise da capacidade preditiva da mesma. Para tal, são comparados os Erros Percentuais Médios de Predição de cada ferramenta, representando a média dos erros percentuais individuais das cinco notas. A ferramenta mais adequada é aquela com menor erro de predição gerado. Por fim, é possível avaliar, com base nos coeficientes das equações geradas sobre os dados normalizados, as variáveis com maior influência sobre os resultados, identificando tendências e relações entre variáveis dependentes e independentes.

4. Resultados

Nesta seção é descrita a aplicação do método proposto em um conjunto de dados provenientes de uma empresa de ensino online de pequeno porte, situada na cidade de Porto Alegre, Rio Grande do Sul. A empresa utiliza-se de uma plataforma online de desenvolvimento e manutenção própria para a transmitir conhecimento aos alunos assinantes. Isso ocorre por meio da assinatura de cursos, presentes nas modalidades ensino médio, pré-vestibular e ensino superior. Ela opera a nível nacional, o que é facilitado pela forma de compra e entrega do produto, inteiramente virtual.

A empresa atualmente conta com uma base de cerca de dez mil assinantes, sendo cerca de seis mil na modalidade pré-vestibular. Por representar cerca de dois terços do total de assinaturas, grande parte das ações de marketing e *customer service* são destinadas a esta modalidade. Além de manter a base de dados dos alunos, são aplicadas pesquisas de forma recorrente, para que seja possível ampliar a base de informações e possibilitar ações eficazes e melhor direcionadas.

4.1. Coleta e avaliação dos dados

Para que fosse possível ter acesso ao desempenho final dos alunos, foi aplicada uma pesquisa em que os estudantes deveriam inserir a nota final alcançada em cada uma das cinco áreas de conhecimento da prova. As áreas de conhecimento presentes na prova são: Linguagens, Códigos e suas Tecnologias; Matemática e suas Tecnologias; Ciências Humanas

e suas Tecnologias; Ciências da Natureza e suas Tecnologias; e Redação. A pesquisa também contemplava questões de natureza social e psicológica, como a escola de curso durante o ensino médio e o objetivo ao submeter-se à prova. Porém, a quantidade de respostas obtidas para estas questões não foi suficiente para que fossem incorporadas ao modelo.

A base de dados já mantida pelo servidor da empresa continha diversas informações relacionadas à assiduidade dos alunos aos cursos, quantidade de exercícios resolvidos, desempenho em diversas categorias, etc. Avaliando todos os dados disponíveis, dez tipos de informação foram selecionados para auxiliar na construção do modelo, por estarem apresentadas em quantidade suficiente em relação aos alunos que responderam à pesquisa de desempenho final na prova. São eles: Número de Cursos Atendidos; Quantidade de Aulas Assistidas; Número de Exercícios Resolvidos; Quantidade de Acertos nos Exercícios; Quantidade de Erros nos Exercícios; Número de Simulados Submetidos; Número de Acertos nos Simulados; Número de Erros nos Simulados; Redações Submetidas; e Quantidade de Dias Estudados.

Com as informações selecionadas, a parcela de alunos assinantes apta a participar na construção do modelo continha duzentas e noventa amostras. Assim, foram geradas as matrizes de variáveis dependentes e independentes. A matriz **Y**, de variáveis dependentes, contempla as cinco notas alcançadas em relação a cada aluno da população amostral, como ilustra a Tabela 2. Analogamente, a matriz **X**, ilustrada na Tabela 3, representa as variáveis independentes, formadas pelos dez tipos de informações presentes na base de dados da empresa. Uma última matriz $\mathbf{X_N}$, de mesmo formato que a matriz **X**, contempla as variáveis independentes normalizadas. As variáveis y_{nj} representam as notas obtidas pelo aluno *n* na área de conhecimento *j*, enquanto as variáveis x_{nq} representam os valores do aluno *n* para o tipo de informação *q* obtidos.

Tabela 2 – Estrutura da matriz de variáveis dependentes **Y**.

Fonte: Autores.

Observações	Linguagens, Códigos e suas Tecnologias	Matemática e suas Tecnologias	Ciências Humanas e suas Tecnologias	Ciências da Natureza e suas Tecnologias	Redação
Aluno 1	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
Aluno 2	y_{21}	y_{22}	y_{23}	y_{24}	y_{25}
⋮	⋮	⋮	⋮	⋮	⋮

Aluno *n*	y_{n1}	y_{n2}	y_{n3}	y_{n4}	y_{n5}

Tabela 3 – Estrutura da matriz de variáveis independentes **X**.

Fonte: Autores.

Observações	Cursos	Aulas	Exercícios	Acertos	Erros	...
Aluno 1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	...
Aluno 2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	...
⋮	⋮	⋮	⋮	⋮	⋮	...
Aluno *n*	x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	...

Observações	...	Simulados	Acertos	Erros	Redações	Dias Estudados
Aluno 1	...	x_{16}	x_{17}	x_{18}	x_{19}	x_{110}
Aluno 2	...	x_{26}	x_{27}	x_{28}	x_{29}	x_{210}
⋮	...	⋮	⋮	⋮	⋮	⋮
Aluno *n*	...	x_{n6}	x_{n7}	x_{n8}	x_{n9}	x_{n10}

4.2. Aplicação das ferramentas de análise multivariada utilizando validação cruzada

Após coletados todos os dados necessários e elaboradas as matrizes de variáveis, partiu-se para a aplicação das ferramentas de análise multivariada. Seguindo a metodologia proposta, foi aplicado o algoritmo Kennard-Stone visando a separar os dados nas porções de treino e teste, de maneira uniforme. O mesmo foi feito para o banco de dados normalizados. Ao final da aplicação do algoritmo, as porções de treino contavam com duzentas e dezoito amostras, enquanto as porções de teste contavam com as setenta e duas amostras remanescentes, representando uma proporção aproximada de 75% para 25%.

Em seguida, as variáveis foram modeladas seguindo os métodos detalhados na Seção 2.2 deste artigo, que descrevem as ferramentas de análise multivariada. A Regressão Linear Múltipla (MLR) foi aplicada cinco vezes, uma para cada variável dependente, gerando cinco equações de MLR, uma para cada vetor de variáveis dependentes. O mesmo foi feito com as variáveis independentes normalizadas, gerando outras cinco equações, de mesmo formato,

sendo denominadas equações de MLRN. Os coeficientes encontrados em ambos os casos estão detalhados nas Tabelas 4 e 5, respectivamente.

Tabela 4 – Coeficientes de regressão de MLR.

Fonte: Autores.

Coeficientes	$\mathbf{y_1}$	$\mathbf{y_2}$	$\mathbf{y_3}$	$\mathbf{y_4}$	$\mathbf{y_5}$
β_0	611,39	653,76	645,93	590,80	703,25
β_1	-0,52	-0,65	-0,53	-0,12	0,78
β_2	0,00	0,01	0,01	0,01	0,01
β_3	-0,06	-0,13	-0,12	-0,07	-0,17
β_4	0,07	0,17	0,14	0,09	0,18
β_5	0,03	0,05	0,09	0,02	0,18
β_6	2,18	8,05	2,33	5,32	6,47
β_7	0,14	0,35	0,15	0,17	0,22
β_8	-0,33	-1,23	-0,43	-0,71	-0,72
β_9	2,42	-0,11	2,31	0,95	8,29
β_{10}	0,02	-0,14	0,04	-0,04	-0,17

Tabela 5 – Coeficientes de regressão de MLRN.

Fonte: Autores.

Coeficientes	$\mathbf{y_1}$	$\mathbf{y_2}$	$\mathbf{y_3}$	$\mathbf{y_4}$	$\mathbf{y_5}$
β_0	602,00	635,36	640,39	596,36	736,44
β_1	-5,62	-7,87	-6,26	-1,45	13,29
β_2	-2,16	4,75	1,86	4,75	0,05
β_3	-79,25	-134,54	-149,48	-86,06	-214,20
β_4	60,87	123,53	114,71	78,48	146,65
β_5	12,51	15,63	35,78	8,63	73,63
β_6	14,61	55,99	15,47	36,52	43,24
β_7	19,27	47,50	19,63	22,56	30,53
β_8	-25,10	-93,59	-32,33	-54,41	-55,70
β_9	6,31	-0,61	6,23	1,79	22,05
β_{10}	0,93	-5,76	4,66	-0,63	-7,12

Para a Regressão por Componentes Principais (PCR), foi aplicado, primeiramente, o método PCA em ambos os bancos de dados. Essa aplicação gerou duas matrizes de variáveis latentes $\mathbf{Z}$ e $\mathbf{Z_N}$, em substituição às matrizes originais $\mathbf{X}$ e $\mathbf{X_N}$. Essas matrizes de variáveis independentes foram regredidas utilizando uma regressão linear múltipla, gerando cinco equações do formato da Equação 11 para a PCR e para a PCRN (com variáveis normalizadas). Os coeficientes podem ser encontrados nas Tabelas 6 e 7.

Tabela 6 – Coeficientes de regressão de PCR.

Fonte: Autores.

Coeficientes	$\mathbf{y_1}$	$\mathbf{y_2}$	$\mathbf{y_3}$	$\mathbf{y_4}$	$\mathbf{y_5}$
β_0	600,179	635,753	640,275	593,124	731,033
β_1	-0,002	0,004	0,002	0,005	0,012
β_2	-0,001	0,006	0,008	0,015	0,014
β_3	-0,049	-0,142	-0,057	-0,079	-0,024
β_4	0,061	0,039	0,020	0,027	0,137

Tabela 7 – Coeficientes de regressão de PCRN.

Fonte: Autores.

Coeficientes	$\mathbf{y_1}$	$\mathbf{y_2}$	$\mathbf{y_3}$	$\mathbf{y_4}$	$\mathbf{y_5}$
β_0	601,55	633,52	639,83	595,21	735,40
β_1	-0,37	2,98	1,94	3,37	10,30
β_2	5,93	4,11	0,09	-0,53	0,81
β_3	2,40	-1,63	1,06	2,89	13,11
β_4	6,18	-2,22	4,64	-0,84	21,42

A última ferramenta utilizada foi a Regressão por Mínimos Quadrados Parciais (PLS). O método foi operacionalizado das duas formas descritas na Seção 3.3 deste artigo, ou seja, aplicando a regressão com uma matriz de variáveis dependentes $\mathbf{Y}$ e com cinco vetores de variáveis dependentes $\mathbf{y}_j$. Além disso, ambos os casos foram aplicados nas duas porções de treino disponíveis. Esta ferramenta permite que, após a regressão via variáveis latentes, o modelo retorne um conjunto de coeficientes similares aos de uma MLR. Assim, o resultado destas aplicações foi um conjunto de cinco equações – de mesmo formato que a Equação 3 – para cada forma de operacionalização empregada. O conjunto para PLS utilizando a matriz $\mathbf{Y}$

foi denominado PLS2, enquanto a mesma aplicação para $\mathbf{X_N}$ foi denominado PLS2N. Da mesma maneira, o conjunto de equações utilizando os vetores $\mathbf{y}_j$ foi chamado PLS1 e sua contraparte normalizada PLS1N. Os coeficientes encontrados encontram-se detalhados nas Tabelas 8, 9, 10 e 11.

Tabela 8 – Coeficientes de regressão de PLS2.

Fonte: Autores.

Coeficientes	$\mathbf{y_1}$	$\mathbf{y_2}$	$\mathbf{y_3}$	$\mathbf{y_4}$	$\mathbf{y_5}$
β_0	603,1292	633,4327	637,4851	585,3794	710,6442
β_1	0,0014	-0,0005	0,0001	0,0002	0,0051
β_2	-0,0036	0,0029	0,0064	0,0132	0,0137
β_3	-0,0136	-0,0261	-0,0119	-0,0156	-0,0088
β_4	0,0271	0,0867	0,0342	0,0480	0,0125
β_5	-0,0307	-0,1017	-0,0411	-0,0570	-0,0013
β_6	0,0032	0,0030	0,0014	0,0019	0,0070
β_7	0,0781	0,0839	0,0385	0,0512	0,1627
β_8	0,0040	-0,0053	-0,0012	-0,0017	0,0183
β_9	0,0017	0,0012	0,0006	0,0008	0,0040
β_{10}	0,0002	-0,0033	-0,0007	-0,0007	0,0058

Tabela 9 – Coeficientes de regressão de PLS2N.

Fonte: Autores.

Coeficientes	$\mathbf{y_1}$	$\mathbf{y_2}$	$\mathbf{y_3}$	$\mathbf{y_4}$	$\mathbf{y_5}$
β_0	601,92	635,00	640,36	596,27	735,89
β_1	-3,96	-14,25	-3,32	-5,73	8,43
β_2	4,26	19,94	8,17	12,19	6,92
β_3	-1,44	4,32	1,28	2,93	-0,58
β_4	7,65	31,23	12,45	18,40	9,02
β_5	-18,07	-45,67	-19,22	-25,64	-17,32
β_6	7,10	15,77	6,96	9,16	11,61
β_7	18,12	45,50	19,58	26,27	24,32
β_8	-16,11	-53,72	-23,71	-32,71	-21,04
β_9	5,71	3,23	5,47	4,82	24,49

β_{10}	-2,81	-4,51	-1,56	-1,79	0,80

Tabela 10 – Coeficientes de regressão de PLS1.

Fonte: Autores.

Coeficientes	$\mathbf{y_1}$	$\mathbf{y_2}$	$\mathbf{y_3}$	$\mathbf{y_4}$	$\mathbf{y_5}$
β_0	603,122	634,047	637,235	584,637	709,936
β_1	-0,004	-0,022	-0,011	0,000	0,012
β_2	-0,003	0,006	0,007	0,013	0,014
β_3	-0,013	-0,043	-0,031	-0,020	-0,010
β_4	0,026	0,087	0,048	0,046	0,011
β_5	-0,030	-0,063	-0,015	-0,044	0,004
β_6	0,003	0,009	0,002	0,006	0,007
β_7	0,079	0,280	0,101	0,125	0,163
β_8	0,007	-0,212	-0,061	-0,075	0,021
β_9	0,002	0,001	0,004	0,002	0,006
β_{10}	-0,002	-0,057	-0,006	-0,002	0,011

Tabela 11 – Coeficientes de regressão de PLS1N.

Fonte: Autores.

Coeficientes	$\mathbf{y_1}$	$\mathbf{y_2}$	$\mathbf{y_3}$	$\mathbf{y_4}$	$\mathbf{y_5}$
β_0	601,95	635,14	640,27	596,29	736,60
β_1	-7,44	-3,72	-8,30	-1,97	12,13
β_2	2,48	9,51	6,69	8,96	0,27
β_3	-1,10	5,58	0,38	4,76	1,73
β_4	8,23	32,70	12,42	21,60	9,87
β_5	-17,47	-44,18	-19,35	-26,35	-8,58
β_6	5,81	29,18	5,50	17,32	21,92
β_7	18,39	50,51	19,54	23,86	31,65
β_8	-15,65	-69,87	-21,36	-41,39	-36,96
β_9	6,29	4,20	5,61	2,22	24,42
β_{10}	-0,26	-7,70	3,41	-1,80	-10,76

Em todas as tabelas acima exibidas, os vetores $\mathbf{y}_j$, com j variando entre um e cinco, representam as variáveis dependentes. Em outras palavras, eles representam cada uma das cinco notas preditas para cada aluno *n*, formando seu desempenho final. As constantes β_q, com q variando entre um e dez, e β_l, com l variando entre um e quatro, representam os coeficientes de regressão das equações. Estes são multiplicados pelos vetores de variáveis independentes que formam as matrizes $\mathbf{X}$, $\mathbf{X_N}$, $\mathbf{Z}$ e $\mathbf{Z_N}$, respectivamente, para alcançar as notas finais.

Depois da construção dos modelos através das ferramentas, as equações gerais resultantes foram aplicadas nas porções de teste, separadas anteriormente. As variáveis independentes, x ou t, de acordo com a ferramenta, foram multiplicadas por seus respectivos coeficientes e somados às constantes. A capacidade preditiva dos modelos foi testada subtraindo-se a nota real do aluno *n* pela nota predita e, em sequência, dividida pela nota real. Esse cálculo resulta no Erro Percentual de cada nota, para cada aluno, como mostra a Equação 20.

$$\mathrm{EP}_{ni} = (\hat{y}_{nj} - y_{nj})/y_{nj} \qquad (18)$$

A nota real coletada é representada por y, para cada aluno *n* e área de conhecimento *j*. Analogamente, a nota predita pelo modelo é representada por $\hat{y}$, de mesmas dimensões. Em seguida, foram feitas as médias dos erros percentuais para cada modelo, apresentados na Seção 6 (apêndices A1, A2, A3 e A4). A Tabela 12 contém o Erro Percentual Médio obtido para cada ferramenta.

Tabela 12 – Erro Percentual Médio por ferramenta aplicada.

Fonte: Autores.

Ferramenta	Erro percentual médio	Ferramenta	Erro percentual médio
MLR	13,04%	MLRN	13,75%
PCR	13,36%	PCRN	13,91%
PLS2	13,36%	PLS2N	13,75%
PLS1	13,37%	PLS1N	13,77%

4.3. Análise dos resultados

Com base nos resultados adquiridos, é possível identificar uma grande similaridade entre todas as ferramentas, com Erro Percentual Médio variando entre 13% e 14%. Apesar da normalização de variáveis aproximar a magnitude das mesmas entre si, o erro percentual

apresentado pelos modelos normalizados é levemente superior ao de suas contrapartes regulares, inferindo uma capacidade preditiva igualmente inferior. A MLR, apesar se representar o modelo mais simples entre os utilizados, apresentou a melhor capacidade preditiva, como mostra a Tabela 12. Com base nesta informação, é possível dizer que há baixa correlação entre as variáveis de predição, ou a correlação entre elas pouco afeta a capacidade preditiva do modelo, visto que PCR e PLS operam com a transformação das variáveis originais de forma a anular essa correlação. Por outro lado, observando os coeficientes encontrados nos modelos normalizados de cada ferramenta, identifica-se tendências mais realistas para as equações de PLS.

Analisando individualmente os coeficientes das variáveis independentes, é possível identificar uma significante influência positiva da variável x_6 (Número de Simulados Submetidos), bem como x_7 (Número de Acertos nos Simulados) nas notas de todas as áreas de conhecimento. A variável x_9 (Redações Submetidas) claramente impacta positivamente nas notas das áreas de conhecimento "Redação" e "Linguagens, Códigos e suas Tecnologias", como esperado, por ser uma das relações que faria mais sentido.

De forma geral, pode-se constatar que as variáveis x_5 (Quantidade de Erros nos Exercícios) e x_8 (Número de Erros nos Simulados) causam grande parte do impacto negativo nas notas, enquanto x_4 (Quantidade de Acertos nos Exercícios) causa grande impacto positivo, assinalando uma tendência condizente com o esperado: os alunos tendem a repetir em provas reais os resultados alcançados em simulados e testes preparatórios. Também, nota-se uma leve influência positiva relacionada à variável x_2 (Quantidade de Aulas Assistidas), também condizente com o resultado esperado.

5. Conclusões

Neste trabalho foi proposta uma metodologia para a previsão de desempenho discente através da aplicação de ferramentas de análise multivariada em um conjunto de dados coletados de uma empresa de educação à distância de pequeno porte. Localizada na cidade de Porto Alegre, a empresa atua em nível nacional, criando a oportunidade para coleta de amostras de dados mais homogêneos em comparação a dados referentes a uma região específica.

Três ferramentas multivariadas foram utilizadas para avaliação do método, sendo elas a Regressão Linear Múltipla (MLR), Regressão por Componentes Principais (PCR) e Regressão por Mínimos Quadrados Parciais (PLS). As amostras foram separadas em porções

de treino e teste através do algoritmo Kennard-Stone, permitindo a construção dos modelos de cada ferramenta através da aplicação das mesmas sobre a primeira porção, seguido de comparação com os valores originais da segunda porção, sem que a partição das amostras favorecesse o método. Esta metodologia foi aplicada duas vezes, sendo uma delas utilizando o banco de dados inicial e outra para um conjunto das variáveis originais normalizadas.

A capacidade preditiva de cada uma das ferramentas em relação à amostra foi medida através do Erro Percentual Médio de cada uma delas. Todas as ferramentas apresentaram um erro percentual muito similar, com as ferramentas não normalizadas obtendo um resultado levemente superior. A ferramenta com melhor capacidade preditiva apresentada foi a MLR, apesar dos coeficientes de predição gerados pelos modelos de PLS apresentarem comportamento mais realista. Ainda, foi possível identificar algumas tendências individuais nas variáveis independentes. Um exemplo foi o grande impacto que o Número de Simulados Submetidos tem para aumento da nota. O mesmo ocorre para a o número de Redações Submetidas em relação às notas de “Redação” e “Linguagens, Códigos e suas Tecnologias”.

Pode-se afirmar que a metodologia é aplicável a este tipo de dados, visando previsão de desempenho de estudantes. O estudo mostra-se incompleto, porém, devido à baixa quantidade amostral disponível. Seria possível divisar comportamentos mais claros e concisos com uma quantidade maior de dados para análise. Além disso, este trabalho poderia ser aprimorado incorporando fatores qualitativos, como informações socioeconômicas e psicológicas dos alunos, indisponíveis em quantidade suficiente na coleta de dados.

6. Referências

ABDI, H. **Partial least squares regression and projection on latent structure regression (PLS Regression)**. Wiley Interdisciplinary Reviews: Computational Statistics. v. 2, p. 97-106, 2010.

AL-ARIMI, A. **Distance Learning**. Procedia - Social and Behavioral Sciences. v. 152, p. 82-88, 2014.

ALOGLU, A, K.; HARRINGTON, P. B.; SAHIN, S.; DEMIR, C. **Prediction of total antioxidant activity of Prunella L. species by automatic partial least square regression applied to 2-way liquid chromatographic UV spectral images**. Talanta. v. 161, p. 503-510, 2016.

ANDERSSON, M. **A comparison of nine PLS1 algorithms**. Journal of Chemometrics. v. 23, n. 10, p. 518-529, 2009.

BIAN, X.; LI, S.; LIN, L.; TAN, X.; FAN, Q.; LI, M. **High and low frequency unfolded partial least squares regression based on empirical mode decomposition for quantitative analysis of fuel oil samples**. Analytica Chimica Acta. v. 925, p. 16-22, 2016.

BORRÀS, E.; FERRÉ, J.; BOQUÉ, R.; MESTRES, M.; ACEÑA, L.; CALVO, A.; BUSTO, O. **Prediction of olive oil sensory descriptors using instrumental data fusion and partial least squares (PLS) regression**. Talanta. v. 155, p. 116-123, 2016.

BRINSON, J. R. **Learning outcome achievement in non-traditional (virtual and remote) versus traditional (hands-on) laboratories: A review of the empirical research**. Computers & Education. v. 87, p. 218-237, 2015.

DA SILVA, E. L.; MENEZES, E. M. **Metodologia da Pesquisa e Elaboração de Dissertação**. 4. ed. rev. atual. Florianópolis: UFSC, 2005.

DAN, L.; SHI, X. **Estimates of pedestrian crossing delay based on multiple linear regression and application**. Procedia - Social and Behavioral Sciences. v. 96, p. 1997-2003, 2013.

EL-MAGBOUB, A.; HAWORTH, I. S.; SUTCH, B. T.; ROMERO, R. M. **Evaluation of in-class and online discussion meetings in a biopharmaceutics problem-based learning class**. Currents in Pharmacy Teaching and Learning. In press, 2016.

FACCHIN, S. **Técnicas de Análise Multivariável aplicadas ao Desenvolvimento de Analisadores Virtuais**. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2005.

FAED, A.; CHANG, E.; SABERI, M.; HUSSAIN, O. K.; AZADEH, A. **Intelligent customer complaint handling utilising principal component and data envelopment analysis (PDA)**. Applied Soft Computing. v. 47, p. 614-630, 2016.

FÁVERO, L. P.; BELFIORE, P.; SILVA, F. L.; CHAN, B. L. **Análise de Dados - Modelagem Multivariada para Tomada de Decisões**. Rio de Janeiro: Elsevier, 2009.

GELADI, P.; KOWALSKI, B. R. **Partial least-squares regression: a tutorial**. Analytica Chimica Acta. v. 185, p. 1-17, 1986.

GIL, A. C. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas, 2002.

HE, S.-G.; WANG, G.A.; COOK, D.F. **Multivariate measurement system analysis in multisite testing: An online technique using principal component analysis**. Expert Systems with Applications. v. 38, p. 14602–14608, 2011.

HELLAND, I. S. **On the Structure of Partial Least Squares Regression**. Communications in statistics - Simulation and Computation. v. 17, n. 2, p. 581–607, 1988.

HERRIG, I. M.; BÖER, S. I.; BRENNHOLT, N.; MANZ, W. **Development of multiple linear regression models as predictive tools for fecal indicator concentrations in a stretch of the lower Lahn River, Germany**. Water Research. v. 85, p. 148-157, 2015.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. New Jersey: Pearson Prentice Hall, 2007.

KAPLAN, A. M.; HAENLEIN, M. **Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster**. Business Horizons. v. 59, p. 441-450, 2016.

KENNARD, R. W.; STONE, L. A. **Computer Aided Design of Experiments**. Technometrics. v. 11, n. 1, p. 137-148, 1969.

KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. International Joint Conference on Artificial Intelligence. v. 14, n. 2, p. 1137-1145, 1995.

MARHABA, T. F.; BORGAONKAR, A. D.; PUNBURANANON, K. **Principal component regression model applied to dimensionally reduced spectral fluorescent signature for the determination of organic character and THM formation potential of source water**. Journal of Hazardous Materials. v. 169, p. 998-1004, 2009.

RANAMUKHAARACHCHI, S. A.; PEIRIS, R. H.; MORESOLI, C. **Fluorescence spectroscopy and principal component analysis of soy protein hydrolysate**

fractions and the potential to assess their antioxidant capacity characteristics. Food Chemistry. v. 217, p. 469-475, 2017.

RENCHER, A. C. **Methods of Multivariate Analysis**. New York: Wiley, 2002.

RODRIGUES, S. J.; AFFONSO, S. A.; QUINELATO, E.; MONTIEL, J. M. **Distance Learning in Undergraduate Education: The Challenges of Building a Collaborative Environment**. Procedia - Social and Behavioral Sciences. v. 116, p. 3499-3501, 2014.

ROSIPAL, R.; KRÄMER, N. **Overview and Recent Advances in Partial Least Squares**. Lecture Notes in Computer Science. v. 3940, p. 34-51, 2006.

SAMPAIO, B.; SAMPAIO, Y.; MELLO, E. P.; MELO, A. S. **Desempenho no vestibular, background familiar e evasão: evidências da UFPE**. Economia Aplicada. v. 15, n. 2, p. 287-309, 2011.

SPARTA, M.; GOMES, W. B. **Importância Atribuída ao Ingresso na Educação Superior por Alunos do Ensino Médio**. Revista Brasileira de Orientação Profissional. v. 6, n. 2, p. 45-53, 2005.

TEOH, S. T.; KITAMURA, M.; NAKAYAMA, Y.; PUTRI, S.; MUKAI, Y.; FUKUSAKI, E. **Random sample consensus combined with partial least squares regression (RANSAC-PLS) for microbial metabolomics data mining and phenotype improvement**. Journal of Bioscience and Bioengineering. v. 122, p. 168-175, 2016.

TURÓCZY, S.; MARIAN, L. **Multiple regression analysis of performance indicators in the ceramic industry**. Procedia Economics and Finance. v. 3, p. 509-514, 2012.

UYANIK, G. K.; GÜLER, N. **A study on multiple linear regression analysis**. Procedia - Social and Behavioral Sciences. v. 116, p. 234-240, 2013.

VIVARACHO-PASCUAL, C.; SIMON-HURTADO, A.; MANSO-MARTINEZ, E.; PASCUAL-GASPAR, J. M. **Client threshold prediction in biometric signature recognition by means of Multiple Linear Regression and its use for score normalization**. Pattern Recognition. v. 55, p. 1-13, 2016.

7. Apêndice

A1 – Erro Percentual Médio por Área de Conhecimento gerado pela MLR e MLRN.

	Nota média real	Nota média de previsão	Erro Percentual	Nota média real (N)	Nota média de previsão (N)	Erro Percentual (N)
y_1	617,94	603,29	7,30%	614,13	604,42	7,54%
y_2	639,60	640,36	15,52%	643,60	637,21	15,39%
y_3	639,82	641,59	8,21%	640,42	640,19	8,08%
y_4	589,75	592,19	17,28%	582,43	595,35	18,45%
y_5	720,90	723,47	16,89%	708,68	728,44	19,28%

A2 – Erro Percentual Médio por Área de Conhecimento gerado pela PCR e PCRN.

	Nota média real	Nota média de previsão	Erro Percentual	Nota média real (N)	Nota média de previsão (N)	Erro Percentual (N)
y_1	617,94	600,91	7,53%	614,13	602,61	7,78%
y_2	639,60	634,81	15,69%	643,60	629,83	16,00%
y_3	639,82	639,12	7,94%	640,42	637,92	7,83%
y_4	589,75	589,65	17,57%	582,43	590,74	18,88%
y_5	720,90	718,88	18,04%	708,68	724,26	19,06%

A3 – Erro Percentual Médio por Área de Conhecimento gerado pela Regressão PLS2 e PLS2N.

	Nota média real	Nota média de previsão	Erro Percentual	Nota média real (N)	Nota média de previsão (N)	Erro Percentual (N)
y_1	617,94	600,79	7,54%	614,13	604,09	7,64%
y_2	639,60	634,22	15,71%	643,60	635,79	15,43%
y_3	639,82	638,88	7,94%	640,42	640,07	8,00%
y_4	589,75	589,34	17,58%	582,43	595,02	18,62%
y_5	720,90	718,57	18,04%	708,68	726,22	19,07%

A4 – Erro Percentual Médio por Área de Conhecimento gerado pela Regressão PLS1 e PLS1N.

	Nota média real	Nota média de previsão	Erro Percentual	Nota média real (N)	Nota média de previsão (N)	Erro Percentual (N)
y_1	617,94	600,69	7,55%	614,13	604,21	7,58%
y_2	639,60	632,30	15,75%	643,60	636,33	15,53%
y_3	639,82	638,05	7,94%	640,42	639,68	8,06%
y_4	589,75	588,25	17,59%	582,43	595,07	18,42%
y_5	720,90	718,22	18,02%	708,68	729,09	19,25%

A5 – Amostra da base de dados utilizada.

y_1	y_2	y_3	y_4	y_5	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Linguagens, Códigos e suas Tecnologias	Matemática e suas Tecnologias	Ciências Humanas e suas Tecnologias	Ciências da Natureza e suas Tecnologias	Redação	Cursos	Aulas	Exercícios	Acertos	Erros	Simulados	Acertos	Erros	Redações	Dias estudados
601	376	669	443	900	48	1449	1261	718	531	20	382	234	5	141
601	376	669	443	900	48	1449	1261	718	531	20	382	234	5	141
599	694	623	640	840	32	1710	3086	1697	1319	8	256	104	3	121
597	604	632	624	760	32	860	2163	1551	600	8	172	86	3	137
439	721	610	610	740	27	1127	1921	1196	705	8	218	146	5	75
687	671	619	609	600	30	488	632	428	192	6	196	74	5	69
638	728	661	678	920	38	1951	2152	1703	402	2	74	16	10	118
680	795	756	690	880	21	1050	3152	2065	987	12	92	104	18	88
650	780	746	690	880	21	1050	3152	2065	987	12	92	104	18	88
627	734	674	597	880	47	1563	3543	2322	1207	2	24	6	1	204
546	443	590	490	680	17	1032	2340	1389	946	10	122	140	1	96
609	670	615	571	740	16	278	481	306	165	12	120	184	3	81
609	679	685	601	580	44	2538	1706	1211	476	6	114	30	1	191
675	645	681	575	680	12	373	1524	1081	437	8	174	96	8	69
644	710	642	595	940	26	149	159	82	64	2	14	16	10	42
627	807	632	614	600	21	278	822	602	217	8	30	32	1	48
641	681	666	624	740	23	741	1782	1134	636	20	442	280	4	122
613	582	689	625	940	8	129	838	467	363	4	92	88	1	13