Using Data Mining techniques to follow students trajectories in secondary schools of Uruguay

Luiz Antonio Macarini[1], Cristian Cechinel[1], Henrique Lemos dos Santos[2],
Xavier Ochoa[3], Virginia Rodés[4], Guillermo Ettlin Alonso[5], Alén Pérez Casas[4]
[1]Universidade Federal de Santa Catarina (UFSC), Araranguá, Brasil
[2]Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brasil
[3]Escuela Superior Politecnica del Litoral (ESPOL), Guayaquil, Ecuador
[4]Universidad de la Republica (UDELAR), Montevideo, Uruguay
[5]Administración Nacional de Educación Pública (ANEP), Montevideo, Uruguay
luiz.buschetto@posgrad.ufsc.br, cristian.cechinel@ufsc.br, hlsantos@inf.ufrgs.br
xavier@cti.espol.edu.ec, virginia.rodes@cse.edu.uy, gettlinal@anep.edu.uy, alen.perez@cse.edu.uy

***Resumo*—It is possible to observe an enormous increase on the number of researches focused on automatically find patterns and factors that affect students behavior and performance during their learning process. The fields of Learning Analytics and Educational Data Mining are in constant growing, developing new and innovative tools. Furthermore, new methodologies are being created to follow and help students and professors inside the many different types of educational settings. At the same time, it is also possible to see that the majority of the existing works are still restricted to small and controlled experiments, conducted on samples of students data. The present work describes the first step of an international collaboration focused on implementing Learning Analytics on a national scale. Precisely, this work describes the methodology applied to find rules that can be used to follow students' trajectories in secondary schools in Uruguay. The results points out for the possibility of delivering rules by analyzing patterns of students clusters based on their success (or failure) in the school year. Among other findings, this work shows a strong relationship between students grades and their number of absences in the classes.**

***Index Terms*—*Educational Data Mining*, *Learning Analytics*, *Rules*, *Clustering*, *At-risk students*.**

I. Introdução

Com a evolução da tecnologia, houve um aumento na quantidade de informações disponíveis nas bases de dados. Estes grandes volumes são uma fonte de conhecimento que pode ser aplicado em diversos contextos [1]. Dados coletados de diversas fontes precisam de um método apropriado para que seja possível extrair conhecimento e auxiliar nas tomadas de decisões [2]. Isso porque os seres humanos possuem uma capacidade limitada para extrair conhecimento de dados não tratados [3]. É com o intuito de encontrar informações úteis dentro destas grandes coleções de dados que técnicas de *Data Mining* vem sendo utilizadas [4].

A aplicação de *data mining* no contexto educacional, conhecida como *Educational Data Mining* (EDM), conciliada com técnicas de *Learning Analytics* se apresentam como áreas de pesquisa muito importantes. Estas visam encontrar conhecimento nas bases de dados educacionais [5], tais como regras de associação, classificação e clusterização [2].

De um ponto de vista prático, técnicas de EDM permitem a descoberta de conhecimento baseado em dados provenientes dos próprios estudantes, visando validar sistemas educacionais. Além disso, alguns aspectos da qualidade da educação podem ser melhorados, criando uma base para um processo de aprendizado mais efetivo [6]. Por este motivo, nos últimos anos houve um aumento na quantidade de pesquisas visando encontrar os motivos que influenciam o desempenho dos alunos [7].

A habilidade de prever a performance de estudantes é beneficial aos sistemas educacionais modernos [5]. Porém, esta não é uma tarefa fácil [8]. Recentemente, técnicas de *data mining* foram utilizadas para fornecer novos *insights* para este problema, já que muitos fatores podem influenciar o desempenho do estudante [5]. Técnicas de predição auxiliam no momento de realizar intervenções, visando evitar uma possível reprovação e/ou evasão escolar. Isso porque geralmente os alunos apresentam sinais antes de se desligar formalmente de um curso [9].

Neste contexto, o presente trabalho propõe uma abordagem, através da utilização de técnicas de *data mining*, para buscar informações em uma base de dados educacional do ensino secundário do Uruguai. O objetivo é encontrar uma relação entre a quantidade de matérias que possuem notas abaixo de um ponto de corte na primeira e na segunda reunião e o *fallo* (resultado final) do aluno ao término do ano letivo. A partir destes resultados, espera-se encontrar regras (padrões) que auxiliem os profissionais da educação no acompanhamento da trajetória dos alunos, com *feedbacks* mais precisos. O presente trabalho é parte dos resultados de um amplo projeto. Este, por sua vez, visa a implementação de um sistema nacional de monitoramento da trajetória acadêmica dos estudantes nos níveis primário e secundário do Uruguai [10].

O restante do trabalho está organizado da seguinte maneira: a Seção II apresenta os trabalhos relacionados a este tema. Na Seção III é apresentado um breve referencial teórico para auxiliar no entendimento do trabalho. Já na Seção IV é dada uma descrição do sistema educacional uruguaio, assim como uma breve explicação sobre a estrutura da base de dados. Na

Seção V é apresentada a metodologia utilizada, detalhando cada etapa da realização do trabalho. Seguindo para a Seção VI, são apresentados os resultados obtidos e as discussões sobre os mesmos. Fechando o trabalho, na Seção VII, são apresentadas as considerações finais, assim como os trabalhos futuros.

II. Trabalhos relacionados

A adoção de tecnologias educacionais proporcionou uma nova oportunidade de compreensão do aprendizado dos estudantes. Seus dados podem ser analisados para identificar padrões de comportamento em relação ao aprendizado [11]. Pesquisas envolvendo *Educational Data Mining* foram feitas visando revelar informações úteis provenientes de bases de dados educacionais. Por exemplo, tenta-se prever o sucesso dos estudantes, fato que é muito benéfico aos sistemas educacionais modernos [5]. Vários trabalhos foram realizados visando desenvolver modelos que possam indicar estudantes em risco.

Em [7] foram aplicadas técnicas de *data mining* para prever reprovações em escolas da cidade de Zacatecas, no México. Vários experimentos foram realizados visando melhorar a acurácia na predição do desempenho do estudante, e de forma mais específica, quais deles poderiam reprovar. Foram obtidos bons resultados, incluindo regras de associação que puderam ser utilizadas para encontrar os estudantes que estavam em risco de reprovação. Uma abordagem similar foi adotada em [3], onde foram utilizados métodos de *business intelligence* e *data mining* para analisar o desempenho de estudantes do ensino médio em duas escolas públicas de Portugal. Além disso, tentou-se identificar quais seriam as variáveis chaves que afetam o desempenho educacional. Dados demográficos, sociais e educacionais foram coletados através de questionários e relatórios das escolas.

Em [12] também foram utilizados métodos de *data mining* para criar um modelo visando encontrar estudantes em risco no primeiro ano da faculdade, na *New York Instutute of Technology*. A solução apresenta medidas de risco para cada novo estudante, identificando os fatores que podem fazer com que o mesmo se retire da universidade. Segundo os autores, o resultado foi um modelo que pode ser utilizado para identificar e intervir de maneira precoce, tendo uma boa taxa de acerto, mas com espaço para possíveis melhorias.

Em [13], os autores propuseram uma ferramenta colaborativa *open source* chamada de *Open Academic Analytics Initiative* (OAAI), visando criar alertas para estudantes acadêmicos em risco. Os autores conduziram uma pesquisa em busca da portabilidade entre instituições. Os modelos de predição foram criados utilizando dados demográficos, de aptidão e do ambiente de gestão de aprendizado. Foram obtidos resultados promissores, alcançando uma portabilidade maior do que a inicialmente esperada. Os resultados tiveram um impacto positivo na efetividade da intervenção em relação aos estudantes em risco.

Ainda utilizando dados provenientes do sistema de gestão de aprendizado, em [14] foi conduzido um estudo na *Open University* visando prever quais estudantes estavam em risco de reprovação em um determinado módulo. Mostrou-se que é possível prever sua reprovação observando as mudanças nas atividades desenvolvidas no ambiente de aprendizado. Isto pode ser feito comparando com comportamentos anteriores ou com estudantes que possuem um comportamento de aprendizado similar. Em [9] foi proposto um modelo preditivo criado para a Universidade de *Phoenix*, onde o objetivo era identificar os estudantes que estavam em perigo de reprovação. Os dados foram coletados no sistema de gestão de aprendizado, sistema financeiro e sistema dos estudantes. Os resultados foram utilizados para fazer intervenções e propor recursos adicionais para estes alunos.

Em [15] os autores propuseram uma solução para intervenção acadêmica chamada de *Course Signals*. Esta foi desenvolvida com o intuito de permitir que os instrutores tenham a oportunidade de utilizar técnicas de *Learning Analytics* para dar um *feedback* em tempo real aos estudantes. Além das notas, são levadas em conta características demográficas, histórico acadêmico e esforço do estudante, baseado na interação com o sistema de gestão de aprendizagem. O modelo classifica os estudantes em três categorias: risco de reprovação alto, médio e baixo, simbolicamente representados pelas luzes de um semáforo. A abordagem obteve sucesso com alunos do primeiro e segundo ano, bem como aumentou a retenção global na universidade. Mais de 23 mil estudantes foram afetados em 100 cursos da universidade, além dos 140 instrutores utilizaram o sistema.

III. Referencial Teórico

Nesta seção serão apresentados os conceitos necessários para o entendimento do trabalho.

A. Clusterização

Clusterização (do inglês, *clustering*) consiste no processo de dividir os dados em grupos de acordo com alguma similaridade. Cada grupo, ou *cluster*, é formado por objetos que são similares entre si e diferentes dos objetos pertencentes a outros grupos. Este é um processo de aprendizado não-supervisionado. Quando aplicado em projetos de *data mining*, existem algumas complicações, como por exemplo: bases de dados extremamente grandes, objetos com muitos atributos e objetos dos mais diferentes tipos [16].

1) k-Means Clustering: O *k-Means Clustering* [17] é um método comumente utilizado para particionar automaticamente um conjunto de dados em k grupos. É um processo iterativo que pode ser dividido em basicamente três etapas. A primeira visa escolher duas centróides (ou mais, dependendo da quantidade de *clusters* escolhidos) randomicamente ou a partir da utilização de um algoritmo de inicialização. A centróide representa o centro (ou média) de cada grupo. Então, a distância entre cada ponto e as centróides é calculada e os dados são atribuídos para cada grupo de acordo com a centróide mais próxima. O último passo é calcular a média dos dados, visando encontrar as novas posições das centróides. O segundo e o terceiro passos são repetidos até que o critério de parada seja satisfeito. Este pode ser o número máximo de

iterações ou quando uma acurácia específica é encontrada, por exemplo [18].

IV. Sistema Educacional Uruguaio

No Uruguai, o ensino secundário tem duração de seis anos e se divide em dois ciclos: o *Ciclo Básico* (do primeiro ao terceiro ano) e o *Bachillerato Diversificado* (do quarto ao sexto ano). A partir do quinto, o aluno pode escolher entre quatro diversificações (em espanhol, *diversificaciones*) para cursar. Se este for aprovado, no ano posterior deve escolher entre aquelas disponíveis ao sexto ano, de acordo com a diversificação cursada no ano anterior. A Figura 1 apresenta as opções disponíveis aos estudantes.

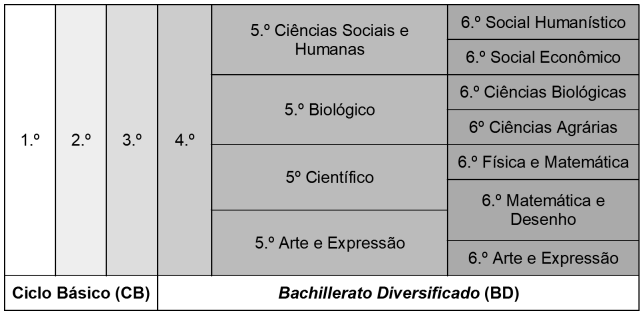

Figura 1: Visão geral da educação secundária no Uruguai e as diversificações disponíveis aos estudantes

Utilizando como exemplo um estudante que optou por cursar "Arte e Expressão" no quinto ano. Quando este é promovido ao terceiro ano do *bachillerato* (sexto ano no geral), pode optar entre "Matemática e Desenho" ou "Arte e Expressão".

Em 2006 houve uma reformulação no sistema educacional uruguaio, mais precisamente no Ciclo Básico e *Bachillerato*. O propósito era focar mais na aprendizagem (*aprendizaje*) do que no ensino (*enseñanza*). Assim, o aluno deve ser avaliado por seu comportamento e rendimento. Em relação a aprendizagem, são levados em conta itens como interesse, atitude em relação ao trabalho e integração social. Já sobre o desempenho, a avaliação está vinculada ao grau em que o aluno alcançou os objetivos propostos [19].

A reformulação permite que o aluno passe ao ano seguinte mesmo que não tenha atingido a qualificação necessária em até três matérias. Assim, ele terá uma assistência especial durante o andamento do curso (no ano seguinte) por parte dos professores destas matérias, para que através de estudos dirigidos possa superar as dificuldades encontradas. Os alunos que tiverem qualificação insuficiente em até 50% das matérias (levando em conta a quantidade que está cursando no ano letivo vigente e as com qualificação insuficiente dos anos anteriores) devem ir a exame. Serão reprovados (sem direito a exame) aqueles alunos que possuem qualificação insuficiente nas metade das matérias, mais uma [19].

A avaliação se dá através de reuniões. Acontecem no mínimo três durante o período letivo, onde é discutido o desempenho de cada estudante até o momento. Após cada reunião, o aluno recebe uma qualificação geral de acordo com cada matéria, até o momento em que a reunião acontece. Ao fim da terceira, o estudante recebe o seu *fallo*, que seria o resultado do seu desempenho durante o período letivo. Este, pode ser categorizado em *promovido*, onde o aluno obteve a sua aprovação; *repite por rendimiento*, quando o aluno reprova por baixo desempenho e *repite por inasistencia*, onde o aluno repete o ano por frequência insuficiente. Além disso, há o *fallo em suspenso*, que significa que o estudante ficou em exame. Para estes alunos, existe uma quarta reunião onde será decidido o seu *fallo final*, após a realização dos exames [19]. Esta última classificação não será explicitamente utilizada neste trabalho, já que será verificado apenas se o aluno foi aprovado ou reprovado.

Deste ponto em diante, o termo "resultado final" será utilizado para descrever o *fallo*. Além disso, *promovido* será tratado como "aprovado", *repite por rendimiento* como "reprovação por baixo desempenho" ou "reprovação por desempenho insuficiente". Já *repite por inasistencia* será tratado como "reprovação por faltas" ou "reprovação por frequência insuficiente".

A. Base de Dados

A base de dados utilizada neste trabalho foi cedida pela *Administración Nacional de Educación Pública* (ANEP) e contém dados da educação secundária no país. Possui cerca de 185 mil matrículas, distribuídas entre dois anos (2015 e 2016). São cerca de 135 mil estudantes e aproximadamente 8 mil *grupos*, sendo estes equivalentes as "turmas" no Brasil.

Dentre os dados demográficos dos estudantes, estão disponíveis o endereço, o departamento (similar ao que seria uma província), o gênero e a idade. Ainda, estão disponíveis dados de 254 centros (dentre escolas e academias). A base contém itens como a localidade, o departamento, se este se encontra em zona rural ou urbana e em qual região do país está.

Já em relação aos dados qualitativos, estão presentes os resultados finais dos alunos e o total de faltas, justificadas ou injustificadas, até o momento das reuniões. Além disso, tem-se as notas de qualificação geral de cada matéria até uma dada reunião. Por último, existem campos de texto livre, como o *juicio*, para cada reunião. Optou-se por não utilizar este último, já que seria necessário aplicar técnicas de análise textual.

V. Metodologia

A Figura 2 mostra a visão geral da metodologia usada. Esta foi dividida em etapas para facilitar sua explicação e replicação.

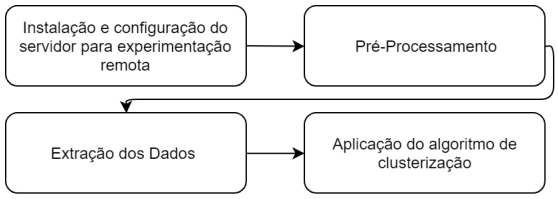

Figura 2: Visão geral da metodologia utilizada

A primeira etapa consiste em instalar e configurar o servidor para acesso ao banco de dados. Na etapa seguinte, foi feito o pré-processamento dos dados, extraindo-os da base de dados e gerando uma tabela contendo apenas os mais relevantes, visando facilitar a manipulação dos mesmos. A terceira etapa consiste na extração dos dados utilizando consultas SQL (*Structured Query Language*) a partir da tabela gerada no passo anterior. Na última etapa, um algoritmo de clusterização foi aplicado nos dados extraídos visando encontrar informações relevantes sobre o desempenho dos alunos. As etapas supracitadas serão explicadas em mais detalhes nas subseções seguintes.

Para este estudo, foram utilizados dados do ciclo básico (equivalente ao ensino médio no Brasil), onde estes são divididos em três *grados* (sendo similares as "séries")[1]. Além disso, foram levados em conta os dados dos alunos que frequentam as escolas que aderiram ao plano educacional proposto em 2006. Ainda, os dados foram limitados aos anos de 2015 e 2016.

A. Instalação do servidor para experimentação remota

Por questões legais, os dados disponibilizados não puderam ser retirados do país. Ou seja, estes não poderiam ser trazidos ao Brasil para se trabalhar localmente. Assim, foi necessário fazer a instalação de um servidor para acesso remoto ao banco de dados, também utilizado para a realização dos experimentos. Pelos mesmos motivos, toda a base de dados foi cedida com o seu conteúdo anonimizado. Assim, seria possível manter a privacidade dos estudantes que passaram pelo ensino uruguaio.

Inicialmente, o servidor disponibilizado pela *Universidad de la República* (UDELAR) foi configurado. Optou-se por colocá-lo em uma máquina virtual, com sistema operacional Ubuntu 16.04 LAMP (Linux, Apache, *MySQL*, PHP). Adicionalmente, foi instalado o *PostgreSQL 9.5.11* e foi configurado para acesso externo.

O *software Pentaho Data Integration CE 8.0* foi utilizado para extração das bases do servidor da ANEP. O *Pentaho* trabalha com ETLs (*Extract*, *Transform*, *Load*). Isto é, com processos que envolvem a extração, transformação e carga das bases de dados. Essa ferramenta utiliza uma base *MySQL* como repositório de metadados e permite que se programe (visualmente) o processo de ETL dos dados. Optou-se por utilizá-la devido a sua robustez, por ser multi-plataforma, ter uma interface gráfica intuitiva, inúmeras funções para pré-processamento, com uma documentação extensa, além de ser gratuita (em sua versão *community*). Além disso, permite conexão direta com o banco de dados, onde a extração dos dados pode ser feita através da utilização de consultas (*queries*) SQL.

Assim que a base de dados (anonimizada) do *Consejo de Educación Secundaria* (CES) foi disponibilizada, deu-se início ao processo de extração dos dados. Estes foram movidos a partir de uma base localizada no servidor da ANEP para o servidor da UDELAR, sendo armazenados em uma base *PostgreSQL*. A extração destes foi feita através do desenvolvimento de ETLs com o *Pentaho*.

[1]Deste ponto em diante, a palavra "série" ou "ano" será utilizada para denotar *grado*.

B. Pré-Processamento

De posse dos dados e utilizando o *Pentaho*, trabalhou-se em ETLs que pudessem gerar conjuntos de dados passíveis de mineração de regras (padrões). Assim, foram executadas ETLs que produziram tabelas contendo informações numéricas como a quantidade de faltas, idade e qualificação geral para uso posterior via arquivos CSV (*Comma Separated Values*) nos algoritmos de mineração de dados.

Para este trabalho, foi gerada uma tabela contendo 17 atributos, onde 16 deles foram utilizados como entrada do modelo e o resulto final do aluno foi utilizado para a avaliação dos *clusters*. A Tabela I apresenta (de maneira resumida) as variáveis utilizadas, juntamente com as suas respectivas descrições.

Tabela I: Variáveis utilizadas neste trabalho e suas descrições

Variável	Descrição
cant_materias	Quantidade de matérias que o aluno cursou
idade	Idade do aluno
pmat_mx_ry	Porcentagem de matérias (do total que este estava cursando) que o aluno obteve a qualificação geral (média) menor que x na reunião y, onde x varia de 3 à 7, e y varia de 1 à 2
inasinjust_rx	Quantidade de faltas não justificadas até a reunião x
inasjust_rx	Quantidade de faltas justificadas até a reunião x

É importante esclarecer que nas variáveis $pmat\_mx\_ry$, mx representa qual é a média (ponto de corte) que está sendo levado em conta. Seu valor varia de 3 à 7. Já ry indica a qual reunião a variável se refere e seu valor varia de 1 à 2. Por exemplo, $pmat\_m3\_r1$ é referente a nota média 3 até o momento da reunião 1. Ou seja, ela representa a quantidade de matérias (em porcentagem) em que o aluno estava com média menor que 3 até o momento da reunião 1. Logo, $pmat\_m3\_r2$ se refere a mesma média até a reunião 2; $pmat\_m4\_r1$ representa a média 4 até a reunião 1, e assim por diante. As variáveis $inasjust\_rx$ e $inasinjust\_rx$ seguem o mesmo padrão, onde rx indica qual reunião está sendo levada em conta. Isto resulta em duas variáveis de cada tipo.

C. Extração dos Dados

Novamente utilizou-se o *Pentaho* para extrair os dados da tabela gerada no passo anterior. Nesta etapa, apenas a função de extração via consultas SQL foi utilizada, pois os dados contidos nesta tabela foram pré-processados.

Porém, haviam algumas linhas com valores nulos, possivelmente causadas por erro humano no momento da inserção dos dados. Esta filtragem foi realizada na própria consulta SQL, excluindo-se as linhas que possuíam algum campo com valor nulo. Além disso, para diminuir a necessidade de limpeza dos dados antes da aplicação do algoritmo de clusterização, toda a seleção (delimitação por séries) também foi feita na consulta SQL. Ou seja, foram gerados três *datasets*, cada um referente a uma série (primeiro, segundo e terceiro ano).

Tendo estes dados extraídos da tabela, a última ação foi gerar o arquivo CSV que foi utilizado na etapa seguinte. É necessária apenas uma pequena configuração para que o arquivo gerado seja compatível com o *Weka* [20] e sua *Application Programming Interface* (API).

D. Aplicação do algoritmo de clusterização

Inicialmente optou-se pela mineração de regras de associação nos *datasets* gerados, a partir da utilização do algoritmo *Apriori* [21]. As regras de associação tem como objetivo mostrar condições que ocorrem ao mesmo tempo, de modo frequente, em um determinado *dataset* [22].

Optou-se por utilizar a linguagem *R* [23] por possuir uma grande quantidade de bibliotecas, permitindo a sua aplicação em diversos contextos. Com esta, foram desenvolvidos *scripts* a partir da utilização da biblioteca *arules* [24], que disponibiliza o algoritmo *Apriori* para mineração de regras de associação. Desta forma, as primeiras foram geradas visando entender os fatores que influenciam o desempenho do aluno.

Porém, a utilização da abordagem citada não trouxe bons resultados. Assim, visando encontrar informações relevantes nos dados extraídos, optou-se por utilizar técnicas de clusterização. Para fazer a separação dos dados, escolheu-se o algoritmo *k-Means Clustering* por sua capacidade de apresentar resultados facilmente interpretáveis.

Neste trabalho, a implementação do *k-Means Clustering* foi feita em *Java*, utilizando a API do *Weka*. Optou-se por utilizar esta ferramenta por sua portabilidade, facilidade de uso e extensa documentação. Ainda, utilizou-se uma quantidade de *clusters* $k = 3$, sendo este valor referente a quantidade de categorias possíveis para o resultado final do aluno (aprovado, reprovado por baixo desempenho e reprovado por frequência insuficiente). A inicialização do algoritmo foi feita de modo randômico e a Distância Euclidiana foi utilizada como medida de distância.

VI. Resultados e discussões

A Figura 3 mostra a proporção das amostras distribuídas no *dataset*. Pode-se perceber que há um desbalanceamento em relação a quantidade de itens nas categorias, onde o número de alunos aprovados é maior do que os outros tipos de amostras.

As Tabelas II, III e IV mostram os centróides resultantes, obtidos através da utilização do algoritmo *k-Means Clustering*, levando em conta os resultados finais (*fallos*) dos alunos. As faltas não-justificadas são aquelas em que o aluno falta e não dá nenhum tipo de justificativa ao professor. Já nas justificadas, o estudante apresenta algum tipo de explicação, como um atestado médico, por exemplo.

Para as três séries, os alunos do primeiro e do segundo ano cursam aproximadamente 12 matérias, e os do terceiro cursam aproximadamente 13. Em relação a idade, pode-se observar que a média aumenta de acordo com o resultado final que está sendo analisado. Os alunos aprovados possuem uma média de idade menor do que aqueles que repetem por rendimento, que por sua vez é menor do que aqueles que repetem pela quantidade de faltas.

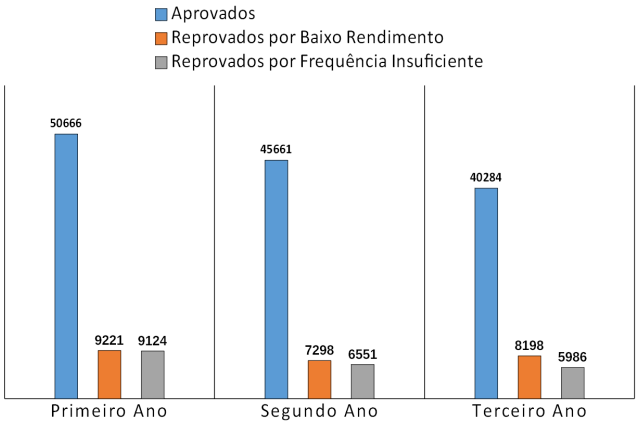

Figura 3: Quantidade de itens de cada categoria presente no *dataset*

Tabela II: Centróides resultantes para os alunos que foram aprovados

Variáveis	Primeiro Ano	Segundo Ano	Terceiro Ano
cant_materias	11.9583	11.9661	13.0005
idade	12.8453	13.8631	14.9177
pmat_m3_r1	8.0574	7.9251	5.7562
pmat_m3_r2	5.7074	5.6952	4.0129
pmat_m4_r1	8.2028	8.1311	5.984
pmat_m4_r2	5.8289	5.8404	4.1947
pmat_m5_r1	8.9928	9.0428	7.0879
pmat_m5_r2	6.448	6.5154	5.0949
pmat_m6_r1	14.1091	14.8976	13.5926
pmat_m6_r2	10.1624	10.6905	9.8643
pmat_m7_r1	35.3271	37.2662	37.4944
pmat_m7_r2	24.0937	25.7453	26.4785
inasinjust_r1	2.4528	3.4541	3.9361
inasinjust_r2	5.809	7.6511	8.9487
inasjust_r1	1.0515	1.1833	1.3963
inasjust_r2	2.6067	2.9512	3.5368

De acordo com a Tabela II, verifica-se que até o momento que acontece a segunda reunião, poucos alunos possuem uma média menor que 5. Porém, este número aumenta quando é levada em conta a nota 6, sendo esta a média limite adotada nas escolas secundárias do Uruguai para o plano do ano de 2006. A quantidade de alunos com a média abaixo de sete no momento da primeira reunião aumenta de modo mais acentuado, diminuindo na segunda reunião, já que estes são aprovados no final do ano letivo.

Além disso, é necessário dizer que o número de faltas é acumulativo. Ou seja, a quantidade de faltas registradas no momento da segunda reunião é igual a quantidade de faltas registrada no momento da primeira reunião, somadas às faltas que aconteceram entre as duas reuniões. Assim, é importante apontar que a quantidade de faltas (justificadas e não-justificadas) é a menor dentre os três grupos de estudantes.

Observando a Tabela III, verifica-se que até o momento da segunda reunião aproximadamente 25% dos alunos que reprovam por baixo rendimento possuem notas médias menores que 5. Porém, esse valor tem um salto quando é levada em conta a média oficial (seis). No momento da primeira e da

Tabela III: Centróides resultantes para os alunos que reprovaram por baixo rendimento

Variáveis	Primeiro Ano	Segundo Ano	Terceiro Ano
cant_materias	11.9687	11.9698	12.994
idade	13.3956	14.339	15.3671
pmat_m3_r1	11.4198	11.2395	9.0989
pmat_m3_r2	9.3592	9.0174	7.7194
pmat_m4_r1	15.1444	14.7945	12.9739
pmat_m4_r2	14.4877	13.4724	12.5046
pmat_m5_r1	26.1789	25.4489	24.1916
pmat_m5_r2	26.3684	24.583	24.1661
pmat_m6_r1	52.3576	50.9826	50.3718
pmat_m6_r2	50.9396	48.7611	48.6528
pmat_m7_r1	83.0485	81.656	81.6131
pmat_m7_r2	79.1623	77.4063	77.5701
inasinjust_r1	5.8863	6.9282	7.1282
inasinjust_r2	13.5609	14.7964	15.6168
inasjust_r1	1.764	1.9489	2.3047
inasjust_r2	4.4637	4.8509	5.3797

segunda reunião, aproximadamente metade dos alunos estão abaixo da média. Ainda, cerca de 80% dos alunos possuem notas médias menores que sete, em ambas as reuniões. Já em relação as faltas, a quantidade é maior quando comparadas aos valores para os alunos aprovados. Percebe-se um aumento significativo de faltas não-justificadas registradas no momento da segunda reunião.

Tabela IV: Centróides resultantes para os alunos que reprovaram por frequência insuficiente

Variáveis	Primeiro Ano	Segundo Ano	Terceiro Ano
cant_materias	11.9635	11.9689	12.9838
idade	14.5923	15.231	16.1978
pmat_m3_r1	43.0403	37.6328	36.0106
pmat_m3_r2	67.6904	60.3698	58.3756
pmat_m4_r1	56.0875	50.1981	48.4257
pmat_m4_r2	79.1243	72.5361	70.5735
pmat_m5_r1	72.5946	67.207	65.6736
pmat_m5_r2	88.7003	84.0458	82.7778
pmat_m6_r1	88.5776	84.9484	84.1
pmat_m6_r2	95.6351	93.2301	92.4206
pmat_m7_r1	96.9901	95.6756	95.1195
pmat_m7_r2	98.7948	97.8617	97.6163
inasinjust_r1	20.2722	18.9524	17.9557
inasinjust_r2	52.2824	48.2996	47.0096
inasjust_r1	2.67	3.2957	3.7215
inasjust_r2	4.4413	5.8874	6.3571

A Tabela IV apresenta os dados dos alunos que reprovaram por frequência insuficiente. É possível perceber que já na primeira reunião, aproximadamente metade dos alunos possuem uma média menor que 4. A mesma quantidade de alunos dá um salto na segunda reunião. Se for levada em conta a média 6, nota necessária para ser aprovado, mais de 80% dos alunos estão abaixo desta já no momento da primeira reunião, chegando a 88% no primeiro ano. Em relação a nota média 7, verifica-se que mais de 95% dos alunos estão abaixo desta já na primeira reunião, em todos os anos.

Ainda, como estes alunos reprovaram por frequência insuficiente, pode-se perceber que principalmente o número de faltas não-justificadas na segunda reunião é bastante alto. Mas já no momento da reunião um, elas são relativamente altas, principalmente se comparadas aos valores dos outros dois anos. Fazendo uma análise superficial, é possível verificar que as faltas justificadas não apresentam grandes diferenças em relação aos outros grupos. A partir destes dados é possível apontar com certa confiança os alunos que serão reprovados por frequência insuficiente (ou que estão em risco).

Pode-se comparar também o perfil médio do estudante de cada categoria em um determinado ano (comparação inter-tabelas). Por exemplo, verifica-se que não há uma diferença significativa entre o perfil médio do estudante aprovado e do estudante que reprova por baixo rendimento. Para isso, deve-se levar em conta apenas as variáveis de ponto de corte baixos (até 4 - $pmat\_m3$ e $pmat\_m4$), em quaisquer um dos anos (comparação entre colunas análogas das Tabelas II e III).

Porém, essa diferença começa a se tornar notável a partir da variável $pmat\_m5$. Ou seja, para o ponto de corte na nota 5: enquanto que o aluno promovido no primeiro ano possui cerca de 9% de matérias abaixo de cinco na primeira reunião, o aluno que repete por rendimento contabiliza cerca de 26% das matérias abaixo de cinco. Uma diferença similar se repete para alunos do segundo e terceiro anos. É possível perceber também que, para quaisquer um dos anos, as variáveis de faltas injustificadas são as mais desiguais quando o *cluster* dos alunos que repetiram por frequência insuficiente é comparado com os outros dois *clusters*.

A Tabela V apresenta a Matriz de Confusão para os três anos analisados neste estudo. O "C_" antes de cada classe (nas colunas) indica que se tratam dos *clusters* gerados pelo algoritmo.

Tabela V: Matriz de Confusão para os três anos. AP - Aprovados; RR - Reprovados por Baixo Rendimento; FI - Reprovados por Frequência Insuficiente

Primeiro Ano			
C_AP	C_RR	C_FI	
34899	15629	138	AP
103	6714	2404	RR
68	3314	5742	FI
Segundo Ano			
C_AP	C_RR	C_FI	
27757	17655	249	AP
107	5432	1759	RR
42	2673	3836	FI
Terceiro Ano			
C_AP	C_RR	C_FI	
23933	16117	234	AP
104	6164	1930	RR
56	2468	3462	FI

Pode-se observar que a maioria das amostras dentro do *cluster* de alunos aprovados, para os três anos, foi classificada de modo correto. Porém, ainda há uma grande quantidade de alunos que foram aprovados e acabaram sendo classificados dentro do *cluster* de reprovados por baixo rendimento.

Em relação aos alunos que reprovam por frequência insuficiente, a maioria foi corretamente classificada. Porém, uma grande quantidade de amostras foi classificada como reprovações por baixo rendimento, sendo esta a classe que melhor se diferencia dos outros grupos. Pode-se observar que,

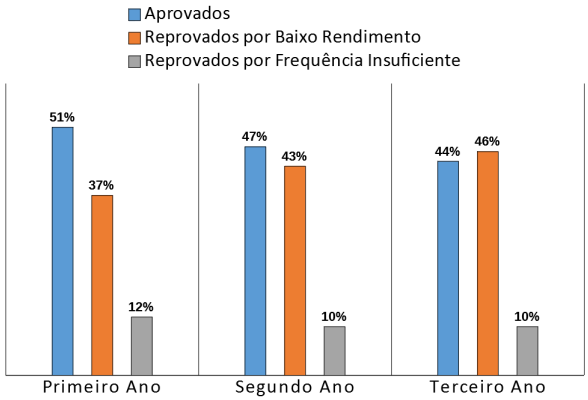

Figura 4: Divisão das amostras de acordo com cada *cluster*

em comparação com as outras duas categorias, esta foi a que obteve a menor proporção (levando em conta o número total de amostras desta categoria) de itens classificados de modo incorreto.

A Tabela VI mostra o resumo dos padrões encontrados observando os resultados obtidos através da aplicação do algoritmo de clusterização. Além disso, são apresentadas sugestões para o momento ideal do possível disparo de um alerta de risco de reprovação, considerando cada padrão.

Uma possível fraqueza da abordagem aqui proposta poderia ser verificada em um cenário onde um dado aluno possui características em comum tanto com o *cluster* de alunos aprovados, quanto com o *cluster* de alunos que reprovam. Para fins de exemplificação, suponhamos que na primeira reunião o aluno possua menos de 14% de suas matérias abaixo da nota 6 (característica média de alunos promovidos). Porém registrou mais de 20 ausências sem justificativa (característica média de alunos reprovados por frequência insuficiente). Uma situação semelhante também poderia ocorrer em reuniões diferentes. Isto é, na primeira reunião o aluno possui características de promoção. Porém, seu desempenho decai até o momento da segunda reunião, se aproximando de algum perfil médio de reprovação.

Para tratar este problema existem duas possibilidades: considerar o pior caso e disparar um alerta via sistema; ou assumir que o estudante pertence ao *cluster* cuja precisão teve maiores valores no experimento executado. Para o presente estudo, optou-se pela primeira solução. Em outras palavras, sempre que houver uma aproximação do estudante em direção a um perfil médio de reprovação, o alerta deveria ser emitido pelo sistema (considerando as dimensões formadas pelas variáveis de percentuais de matérias e quantidades de faltas).

A Figura 4 apresenta a proporção dos itens classificados dentre as categorias. A acurácia obtida para o primeiro ano foi de 68,62%, sendo esta a melhor dentre os três grupos. Já para o segundo, foi de 62,22%. Para o terceiro ano, obteve-se 61,61%.

Pode-se perceber que dentre os alunos classificados como aprovados, a maioria deles está no primeiro ano, e a quantidade destes vai diminuindo a medida que avançam os anos. A quantidade de alunos classificados como reprovados por falta de rendimento é menor no primeiro ano, aumentando a medida que os anos aumentam. Isto está de acordo com a conclusão anterior (diminuição dos alunos aprovados com o passar dos anos), já que a quantidade daqueles que reprovam por frequência insuficiente é a que possui menor variação.

A taxa de erro aumenta a medida que os anos avançam, tendo um salto maior do primeiro para o segundo. Um possível motivo seria a diminuição na quantidade de amostras, principalmente aquelas onde o resultado final do aluno é a aprovação. Para o caso anteriormente citado (alunos aprovados), o algoritmo mostrou uma melhor capacidade de classificação, devido a maior quantidade de amostras.

Apesar da baixa acurácia, com a utilização do algoritmo *k-Means Clustering* foi possível diferenciar os perfis médios dos estudantes, melhorando os resultados anteriormente obtidos através da utilização das regras de associação. Isto serviu para que fosse possível fazer uma análise inicial. Além disso, foi possível encontrar regras (padrões) que fornecem um *feedback* mais preciso ao profissional da educação. A utilização de técnicas como a anteriormente citada permite que seja possível fazer uma análise descritiva dos dados, diferentemente da adoção de algoritmos do tipo "caixa preta".

VII. Considerações finais

Nos últimos anos, houve um crescimento na investigação dos fatores que levam ao baixo desempenho escolar. Ainda, o avanço da tecnologia da informação fez com que houvesse um crescimento nas bases de dados, incluindo as educacionais. Estas, por sua vez, possuem informações valiosas como tendências e padrões, podendo ser utilizadas para melhorar a tomada de decisão. Uma solução promissora para encontrar informações nestas bases é a aplicação de *Data Mining*, que quando utilizada no contexto educacional recebe o nome de *Educational Data Mining*.

Este trabalho propôs uma abordagem baseada em *Data Mining* e *Learning Analytics* na tentativa de encontrar padrões a partir de informações contidas em uma base de dados do ensino secundário do Uruguai. Esta contém informações demográficas e dados qualitativos sobre cerca de 135 mil estudantes uruguaios. Foi utilizado o algoritmo *k-Means Clustering* para encontrar perfis dentro destes dados, divididos entre as séries do ensino secundário.

Foram apresentados dados que podem auxiliar os educadores a acompanhar a trajetória dos alunos de maneira mais precisa, dando um apoio importante no momento da tomada de decisão. O algoritmo de clusterização apresentou resultados facilmente interpretáveis, a ponto de ser possível extrair regras (padrões) que podem ajudar os profissionais durante a orientação educacional.

Analisando os dados dos alunos a partir da utilização desta abordagem, foi possível identificar alguns perfis médios para as variáveis observadas. A quantidade de faltas, por exemplo, se mostrou uma variável importante no momento da análise. Desconsiderando aqueles que repetem por frequência insuficiente, os que reprovam por falta de rendimento apresentam um

Tabela VI: Resumo dos padrões encontrados e sugestões de momentos para o alerta

Regra indicando risco de reprovação	**Sugestão de disparo do alerta**
Cinco faltas não-justificadas	Antes da quinta falta não-justificada
Quatro faltas justificadas	Antes da quarta falta justificada
50% das notas médias abaixo de cinco no momento da primeira reunião	Antes de obter 50% das notas médias abaixo de cinco, no momento da primeira reunião
Idade maior que a média para determinado ano (com reprovações anteriores) e com 50% das notas menor que seis no momento da primeira reunião	Antes de atingir 50% das notas menores que seis, no momento da primeira reunião

aumento acentuado de faltas não-justificadas entre a primeira e a segunda reunião. Há também um aumento nas faltas justificadas, apesar de este não ser tão acentuado quanto nas injustificadas.

A utilização do *k-Means Clustering* trouxe uma acurácia relativamente baixa, mostrando limitações. Isto acontece pelo fato de ser um trabalho preliminar, e pela opção de usar algoritmos que tragam resultados interpretáveis. Mostrou-se também que uma maior quantidade de amostras pode melhorar o desempenho do algoritmo. Ainda, através da matriz de confusão, pode-se notar que a quantidade superior de amostras de alunos aprovados fez com que grande parte destas fossem incorretamente classificadas como alunos que reprovam por falta de rendimento.

Dentre os trabalhos futuros, espera-se analisar a base de dados e verificar a relação entre as notas das matérias na primeira reunião e o resultado final do curso no ano letivo corrente e no próximo. Além disso, espera-se verificar a relação entre a quantidade de faltas em cada reunião e os resultados no fim do ano letivo. Ainda, pretende-se verificar quais são as matérias que levam os alunos a exames e os resultados finais dos mesmos. Um outro experimento seria analisar a diferença entre aqueles que vão a exame e passam facilmente de ano e aqueles que também são aprovados, mas com dificuldades.

AGRADECIMENTOS

Esse trabalho foi financiado pelo "*Fondo Sectorial: Inclusión Digital: Educación con Nuevos Horizontes 2016*" da ANII (*Agencia Nacional de Investigación e Innovación*) do Uruguai, por meio do projeto "*Modelos de predicción para la determinación de riesgo académico*" (código FSED_2_2016_1_130897).

REFERÊNCIAS

[1] M. F. Santos and C. S. Azevedo, *Preâmbulo [a] "Data mining: descoberta de conhecimento em bases de dados"*. FCA-Editora de Informática, Lda, 2005.

[2] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *arXiv preprint arXiv:1201.3417*, 2012.

[3] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[4] H. Mannila, "Data mining: machine learning, statistics, and databases," in *Scientific and Statistical Database Systems, 1996. Proceedings., Eighth International Conference on*. IEEE, 1996, pp. 2–9.

[5] A. Daud, N. R. Aljohani, R. A. Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi, "Predicting student performance using advanced learning analytics," in *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2017, pp. 415–421.

[6] C. Romero, S. Ventura, and P. De Bra, "Knowledge discovery with genetic programming for providing feedback to courseware authors," *User Modeling and User-Adapted Interaction*, vol. 14, no. 5, pp. 425–464, 2004.

[7] C. Marquez-Vera, C. Romero, and S. Ventura, "Predicting school failure using data mining." in *EDM*. ERIC, 2011, pp. 271–276.

[8] M. Ramaswami and R. Bhaskaran, "A chaid based performance prediction model in educational data mining," *arXiv preprint arXiv:1002.1144*, 2010.

[9] R. Barber and M. Sharkey, "Course correction: Using analytics to predict course success," in *Proceedings of the 2nd international conference on learning analytics and knowledge*. ACM, 2012, pp. 259–262.

[10] V. Rodés, C. Cechinel, H. L. D. Santos, X. Ochoa, and G. E. Alonso, "First steps towards the development of an academic system to follow the trajectories of primary and secondary uruguayan students," *I Conferencia Latinoamericana y Summer School de Analíticas de Aprendizaje*, jul 2018, guayaquil.

[11] D. Gašević, S. Dawson, and G. Siemens, "Let's not forget: Learning analytics are about learning," *TechTrends*, vol. 59, no. 1, pp. 64–71, 2015.

[12] L. Agnihotri and A. Ott, "Building a student at-risk model: An end-to-end perspective from user to data scientist," in *Educational Data Mining 2014*, 2014.

[13] S. M. Jayaprakash, E. W. Moody, E. J. Lauría, J. R. Regan, and J. D. Baron, "Early alert of academically at-risk students: An open source analytics initiative," *Journal of Learning Analytics*, vol. 1, no. 1, pp. 6–47, 2014.

[14] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek, "Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment," in *Proceedings of the third international conference on learning analytics and knowledge*. ACM, 2013, pp. 145–149.

[15] K. E. Arnold and M. D. Pistilli, "Course signals at purdue: Using learning analytics to increase student success," in *Proceedings of the 2nd international conference on learning analytics and knowledge*. ACM, 2012, pp. 267–270.

[16] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping multidimensional data*. Springer, 2006, pp. 25–71.

[17] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[18] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl *et al.*, "Constrained k-means clustering with background knowledge," in *ICML*, vol. 1, 2001, pp. 577–584.

[19] B. Nahum, *Historia de Educación Secundaria: 1935-2008*. Administración Nacional de Educación Pública (Uruguay), 2008, no. 373.9 ADMh.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[21] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 1994, pp. 487–499.

[22] V. Kumar and A. Chadha, "An empirical study of the applications of data mining techniques in higher education," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 3, 2011.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org

[24] M. Hahsler, B. Grün, and K. Hornik, "arules - a computational environment for mining association rules and frequent item sets," 2005.