15º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2024

DETERMINANTES DO DESEMPENHO ACADÊMICO E ABANDONO DE DISCIPLINAS: APLICANDO CIÊNCIA DE DADOS AO DATASET DE DUAS ESCOLAS PORTUGUESAS

BIANCA CARVALIO ANTONIETTI[1], LUCAS BUENO RUAS DE OLIVEIRA[2]

[1] Graduanda em Análise e Desenvolvimento de Sistemas, Bolsista PET, IFSP/São Carlos, bianca.carvalio@aluno.ifsp.edu.br
[2] Docente do IFSP/São Carlos, lucas.oliveira@.ifsp.edu.br

Área de conhecimento (Tabela CNPq): 1.03.03.04-9 Sistemas de Informação

RESUMO: A educação é fundamental para o progresso social e econômico de qualquer nação. Este estudo, focado em duas escolas públicas de Portugal, aplica ciência de dados à área de educação para explorar o desempenho acadêmico e o abandono de disciplinas, questões que desafiam sistemas educacionais em diversas partes do mundo. Por meio de uma análise descritiva e preditiva, é investigado como variáveis educacionais, comportamentais e socioeconômicas influenciam as notas dos alunos nas disciplinas de Matemática e Português. Na análise do conjunto de dados, são utilizadas diferentes variáveis independentes, como o nível de escolaridade dos pais, a participação em atividades extracurriculares e o acesso à internet. O estudo também examina as correlações associadas ao abandono de disciplinas, oferecendo informações iniciais sobre as causas desse fenômeno. Embora os dados utilizados sejam específicos de Portugal, a metodologia e as abordagens apresentadas têm o potencial de serem adaptadas a outras realidades educacionais, como no Instituto Federal de São Paulo (IFSP), onde desafios semelhantes podem ser observados, mesmo na ausência de dados abertos disponíveis.

PALAVRAS-CHAVE: ciência de dados; desempenho acadêmico, abandono de disciplinas.

DETERMINANTS OF ACADEMIC PERFORMANCE AND DISCIPLINE DROPOUT: APPLYING DATA SCIENCE TO THE DATASET OF TWO PORTUGUESE SCHOOLS

ABSTRACT: Education is crucial to the social and economic progress of any nation. This study, focused on two public schools in Portugal, applies data science to the field of education to explore academic performance and dropout rates, issues that challenge educational systems in different parts of the world. Through descriptive and predictive analysis, the authors investigate how educational, behavioral, and socioeconomic variables influence students' grades in Mathematics and Portuguese. In the analysis of the dataset, different independent variables are used, such as parents' level of education, participation in extracurricular activities, and internet access. The study also examines the correlations associated with dropout rates, offering initial insights into the causes of this phenomenon. Although the data used are specific to Portugal, the methodology and approaches presented have the potential to be adapted to other educational realities, such as at the Instituto Federal de São Paulo (IFSP), where similar challenges can be observed, even in the absence of available open data.

KEYWORDS: data science; academic performance; discipline dropout.

INTRODUÇÃO

O desempenho acadêmico dos alunos é um tema central para educadores, administradores escolares e formuladores de políticas educacionais. Decerto, é amplamente reconhecido que o desempenho acadêmico é um indicador crucial dos resultados de aprendizagem dos estudantes, além de ser determinante para suas futuras oportunidades de carreira e status socioeconômico (OECD, 2013). Portanto, a identificação dos fatores que influenciam o sucesso ou fracasso acadêmico pode fornecer informações valiosas para a criação de estratégias que melhorem a educação. Já o abandono de disciplinas representa um desafio significativo, pois impede a obtenção de uma educação completa e equilibrada. Essa questão não apenas afeta o progresso acadêmico individual, mas também pode impactar a retenção geral dos alunos, tornando-se um fator crucial a ser abordado para melhorar os resultados educacionais..

Nesse contexto, a Ciência de Dados surge como uma ferramenta poderosa para transformar a educação, oferecendo métodos apropriados para a análise e interpretação de dados. Ao unir conceitos de Computação, Matemática e Estatística, a Ciência de Dados tem se mostrado eficaz na identificação de padrões em conjuntos de dados educacionais, contribuindo para a previsão de resultados acadêmicos e para o desenvolvimento de intervenções mais direcionadas (Siemens & Baker, 2012). Neste artigo, aplicamos técnicas de ciência de dados a um conjunto de dados de duas escolas em Portugal, para explorar os determinantes do desempenho acadêmico e do abandono de disciplinas. As técnicas utilizadas incluem análises descritivas, modelagem preditiva e identificação de correlações entre variáveis. A intenção é estabelecer uma base sólida que permita, futuramente, aplicar essas técnicas a dados de instituições de ensino do Brasil, como as do Instituto Federal de São Paulo.

MATERIAL E MÉTODOS

O conjunto de dados utilizado neste estudo foi obtido do repositório de aprendizado de máquina da Universidade da Califórnia, Irvine (UCI Machine Learning Repository)[1]. Ele contém informações abrangentes sobre alunos do ensino secundário de duas escolas portuguesas, Gabriel Pereira (GP) e Mousinho da Silveira (MS), totalizando 649 instâncias e 33 variáveis, das quais 17 são categóricas e 16 são numéricas. O ensino secundário em Portugal corresponde aos últimos anos de educação formal antes da entrada no ensino superior, abrangendo alunos de 15 a 18 anos em sua maioria, embora haja casos de alunos mais velhos no conjunto, com idades que variam até 22 anos.

As informações no conjunto de dados incluem características demográficas, familiares, acadêmicas e comportamentais. Por exemplo, o tipo de endereço residencial é categorizado como "U" para urbano e "R" para rural; o tamanho da família é classificado como "LE3" (menor ou igual a 3 membros) ou "GT3" (maior que 3 membros); e o estado de coabitação dos pais é indicado por "T" (vivendo juntos) ou "A" (separados). A escolaridade da mãe é registrada na variável “Medu”, variando de 0 (nenhuma escolaridade) a 4 (ensino superior), e a escolaridade do pai é indicada pela variável “Fedu”, também variando de 0 a 4. As ocupações dos pais são representadas por “Mjob”, que classifica a mãe em cinco grupos: "teacher" (professor), "health" (área da saúde), "services" (serviços civis), "at_home" (ocupação doméstica) e "other" (outras ocupações). A variável “Fjob” refere-se à ocupação do pai, com as mesmas categorias de “Mjob”.

Outros fatores incluem o motivo para escolha da escola (“reason”), tempo de deslocamento até a escola (“traveltime”), tempo de estudo semanal (“studytime”), e número de reprovações anteriores (“failures”). As variáveis relacionadas ao comportamento e suporte educacional dos alunos incluem participação em atividades extracurriculares (“activities”), suporte educacional extra (“schoolsup”), aulas pagas (“paid”), e a qualidade das relações familiares (“famrel”). O conjunto de dados também registra a frequência de consumo de álcool durante a semana e nos finais de semana (“Dalc” e “Walc”), além de outros fatores como estado de saúde (“health”) e número de faltas (“absences”).

As notas dos alunos nos três períodos são medidas em uma escala de 0 a 20, sendo G1 a nota do primeiro período, G2 a nota do segundo período, e G3 a nota final, que é a variável-alvo deste estudo. Como descrito no repositório, o conjunto de dados não possui valores nulos. O conjunto de dados é dividido em escalas fixas e bem definidas, portanto, todos os valores foram considerados válidos. Já os

[1] O *conjunto de dados* está disponível em https://archive.ics.uci.edu/dataset/320/student+performance

valores extremos observados em variáveis como 'absences' e 'age' refletem a diversidade natural de uma população estudantil. Tais variações são normais e esperadas, e não indicam erros nos dados.

Em suma, o conjunto de dados é composto por variáveis categóricas, predominantemente binárias, e variáveis nominais, que capturam atributos como ocupação e motivos escolares. As variáveis numéricas são inteiras e, em sua maioria, ordinais, refletindo escalas graduadas como nível de educação e tempo dedicado a atividades específicas. Variáveis como idade, número de faltas e as notas (G1, G2 e G3) utilizam escalas mais amplas e contínuas, embora sejam tratadas como intervalares. Para uma análise mais detalhada, as notas finais foram avaliadas separadamente para as disciplinas de Matemática e Português.

Os métodos utilizados para a análise do conjunto de dados foram variados. Começou-se, primeiramente, por uma análise exploratória dos dados para entender melhor a distribuição e as características das variáveis no conjunto de dados. Essa análise descritiva incluiu a avaliação das estatísticas resumidas para as variáveis numéricas, como a média, mediana, desvio padrão, valores mínimos e máximos. Para as variáveis categóricas, foram geradas tabelas de frequência para identificar as categorias predominantes e observar qualquer desequilíbrio na distribuição. Gráficos de barras foram utilizados para visualizar a distribuição das variáveis categóricas, enquanto histogramas e boxplots foram aplicados para as variáveis numéricas, permitindo a identificação de assimetrias e tendências centrais.

Depois da análise univariada, focada na descrição de cada variável, fez-se uma análise bivariada para explorar a relação entre as variáveis independentes e a variável-alvo. Para isso, foi utilizado técnicas como tabelas de contingência, gráficos de dispersão e correlações de Pearson e Spearman, de modo a identificar possíveis padrões ou correlações que pudessem indicar fatores determinantes no desempenho acadêmico dos alunos. Por outro lado, para a análise dos fatores que influenciam o abandono de disciplinas, foi realizada uma investigação detalhada das variáveis que poderiam indicar um comportamento de abandono das disciplinas. Como parte dessa investigação, foi desenvolvido um novo atributo, que combina a ausência de faltas (valor de “absences” igual à 0) com o desempenho final nulo (“G3” igual à 0), a fim de identificar padrões que possam sugerir abandono. Essa análise também considerou a relação entre a frequência de faltas, o histórico de reprovações, e outros fatores comportamentais e familiares.

Como observado anteriormente, o conjunto de dados possui uma variável-alvo que se enquadra em um problema de regressão, pois as notas finais são representadas em uma escala de 0 a 20. No entanto, para fins de análise, as notas também foram categorizadas em três grupos: "excelente", "regular" e "insuficiente". Essa categorização permite explorar diferentes abordagens de modelagem, considerando tanto a previsão contínua quanto a classificação. A categorização foi realizada da seguinte forma: "excelente" para notas acima de 15, "regular" para notas entre 10 e 15, e "insuficiente" para notas abaixo de 10. Essa dualidade de tratamento possibilita a aplicação de técnicas de regressão e classificação, permitindo uma análise mais ampla dos fatores que influenciam o desempenho acadêmico dos alunos. Não somente isto, mas esta categorização desempenha um papel importante na análise descritiva, pois permite uma interpretação mais clara e detalhada dos dados. Ao transformar as notas em categorias, podemos utilizar gráficos mais inteligíveis para representar as distribuições e comparações entre diferentes grupos de alunos.

Antes de aplicar qualquer técnica de modelagem, os dados foram submetidos a um processo de pré-processamento. Inicialmente, as variáveis categóricas foram convertidas em variáveis numéricas utilizando técnicas de codificação adequadas. Para as variáveis binárias, foi aplicada a codificação simples (0 e 1). Já para variáveis com múltiplas categorias utilizou-se a codificação one-hot. Quando necessária, também foi realizada a normalização das variáveis numéricas que apresentavam escalas amplas, como 'age' e 'absences', garantindo que todos os atributos tivessem um peso comparável.

A etapa seguinte foi dividir o conjunto de dados em dois conjuntos principais: um conjunto de treinamento, composto por 80% dos dados, e um conjunto de teste, com os 20% restantes. Para refinar as variáveis utilizadas na modelagem, foi conduzida uma análise de correlação, visando identificar as variáveis que tinham maior relação com a variável-alvo 'G3'. Para o mesmo objetivo, foi utilizada técnicas de análise de importância de variáveis em modelos baseados em árvores para selecionar as variáveis mais relevantes. Este processo permitiu reduzir a dimensionalidade do conjunto de dados, focando em variáveis que realmente influenciam o desempenho acadêmico e o abandono de disciplinas. Dentre os algoritmos de aprendizado de máquina utilizados para prever a nota final dos

alunos, destacaram-se a Regressão Linear, Regressão Ridge e Árvore de Decisão. Cada modelo foi treinado e ajustado com base no conjunto de dados de treinamento, levando em consideração a natureza do problema e as características intrínsecas dos dados. A seleção desses modelos foi orientada pela busca em capturar tanto as relações lineares quanto as não lineares entre as variáveis independentes e a variável-alvo.

Por fim, para avaliar o desempenho dos modelos, foram utilizadas métricas como o coeficiente de determinação (R^2) e o erro quadrático médio (MSE). A validação cruzada foi aplicada para garantir a robustez dos resultados, reduzindo o risco de sobreajuste (“overfitting”). Outra técnica empregada foi a de matriz de confusão, a qual serve para avaliar o desempenho em classificações binárias específicas, como a identificação de alunos em risco de reprovação

RESULTADOS E DISCUSSÃO

Nos resultados iniciais obtidos, destaca-se, primeiramente, a identificação de padrões consistentes relacionados ao desempenho acadêmico dos alunos e aos fatores associados ao abandono de disciplinas. A análise descritiva inicial revelou que variáveis como a escolaridade dos pais ("Medu” e “Fedu"), o tempo de estudo semanal ("studytime") e o número de reprovações anteriores (“failures”), apresentaram correlações significativas com as notas finais. Essa análise provém, primordialmente, de análises de regressão linear e correlações Pearson e Spearman.

Utilizando as mesmas técnicas de correlação, foi identificado que o abandono de disciplinas na disciplina de matemática é influenciado por quatro variáveis: “failures”, “paid”, “Medu” e “higher”. O histórico de reprovações (“failures”) aponta para dificuldades acadêmicas que podem desmotivar o aluno, enquanto a falta de aulas pagas (“paid”) sugere a ausência de suporte educacional adicional. A escolaridade da mãe (“Medu”) influencia o nível de incentivo familiar, uma vez que pode manter o engajamento acadêmico. Já a aspiração de continuar os estudos em níveis superiores (“higher”) pode motivar os alunos a se dedicarem mais para alcançar seus objetivos, diminuindo o risco de abandono da disciplina. Esses fatores, em conjunto, ajudam a explicar por que certos alunos abandonam ou negligenciam disciplinas desafiadoras como matemática.

Na parte de seleção de atributos, foram empregadas cinco estratégias distintas para identificar as variáveis mais relevantes, tanto para a predição de desempenho acadêmico quanto para o abandono de disciplinas. Foram utilizadas a “SelectKBest” com “f_regresion” para a predição de desempenho acadêmico e “f_classif” e qui-quadrado (“chi2”) para o abandono de disciplinas, adaptando as técnicas às naturezas de regressão e classificação, respectivamente. Foi aplicado também o “Recursive Feature Elimination” (RFE) com modelos de regressão para o desempenho acadêmico com classificadores para o abandono de disciplinas, removendo iterativamente os atributos menos importantes. A importância das variáveis foi avaliada, ainda, com o “RandomForest”, utilizando “RandomForestRegressor” para o desempenho acadêmico e “RandomForestClassifier” para o abandono de disciplinas, ambos baseados na redução da impureza das árvores. O “VarianceThreshold” foi usado em ambos os casos para eliminar atributos com baixa variação, consideradas menos informativas, e a árvore de decisão, com “DecisionTreeRegressor” e “DecisionTreeClassifier”, ajudou a selecionar os atributos com base na importância Gini.

Após treinar diversos modelos com base em algumas variáveis selecionadas, foram obtidos os seguintes resultados. No contexto da análise do conjunto de dados de Matemática, foi utilizado o modelo de Regressão com Árvore de Decisão. Construído com as variáveis “absences”, “G2”, “studytime” e “failures”, esse modelo apresentou resultados expressivos na predição da nota final (G3). O Mean Squared Error (MSE) de 2.22 indica que o erro médio das previsões em relação aos valores reais foi relativamente baixo, o que demonstra a precisão do modelo. Já o coeficiente de determinação R2R^2R2 foi de 0.89, o que sugere que o modelo conseguiu explicar 89% da variação nas notas finais, destacando a forte capacidade preditiva das variáveis selecionadas. Já na análise do abandono de disciplinas, utilizamos o modelo K-Vizinhos Mais Próximos (“KNN”) com as variáveis “failures”,”"paid”,”"studytime” e “Medu”. Após dividir os dados em conjuntos de treinamento e validação, o modelo apresentou uma acurácia de 89.8%. Esse resultado destaca a eficácia do “KNN” em identificar padrões de abandono de disciplinas, auxiliando na compreensão dos fatores que contribuem para a desistência dos alunos.

Já no contexto da análise do conjunto de dados de Português, resultados semelhantes foram obtidos, com destaque para duas novas variáveis impactantes: “famrel” e “schoolsup”. Decerto, obter melhores notas anteriores, em G1 e G2, impactam bastante na nota final. Do mesmo modo, não ter reprovações anteriores impactou significativamente o desempenho em “G3” em Português. Quanto às novas variáveis, “famrel” indica que um bom relacionamento familiar impacta no desempenho acadêmico mais que o esperado. Já a variável “schoolsup” (suporte escolar) também se revelou crucial. No entanto, é interessante notar que, no conjunto de dados de Matemática, “schoolsup” apresenta uma correlação negativa com “G3”, contrastando com o efeito positivo observado em Português. Isto pode indicar que estudantes que recebem suporte extra tendem a ter notas finais mais baixas devido às dificuldades que eles já estavam apresentando. Esses resultados destacam a importância de considerar o contexto específico das disciplinas ao analisar o impacto do suporte escolar e outros fatores no desempenho acadêmico.

Fazendo uma comparação estatística entre os estudantes que desistiram de Português e aqueles que não, observamos a importância das seguintes variáveis:

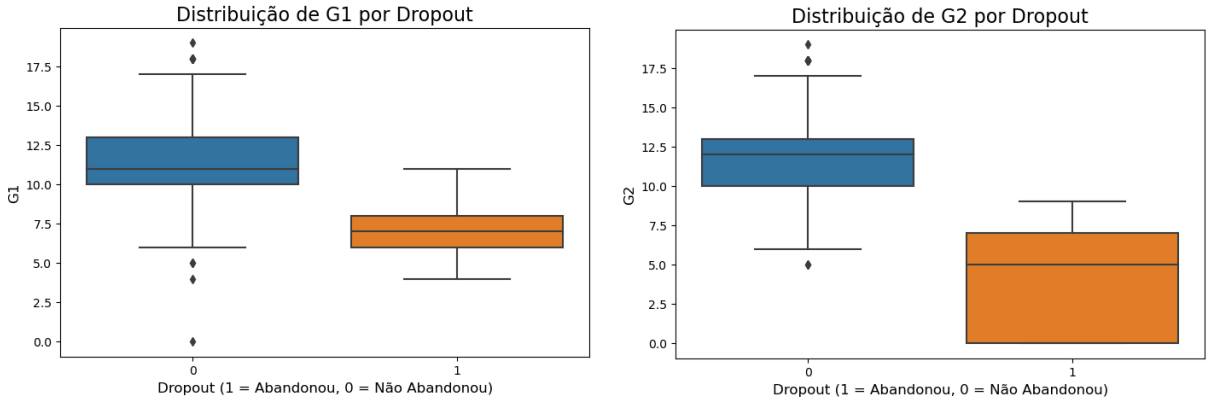

Figura 1. Captura da tela de gráficos feitos no Jupyter Lab. Fonte: Autores

Os gráficos mostram que as notas anteriores, “G1” e “G2”, são significativamente mais altas para os alunos que não abandonaram a disciplina (0 = Não Abandonou) em comparação com aqueles que abandonaram (1 = Abandonou). Esse padrão sugere que um desempenho insatisfatório em “G1” e “G2” pode ser desestimulante ao ponto de levar o aluno a desistir da disciplina.

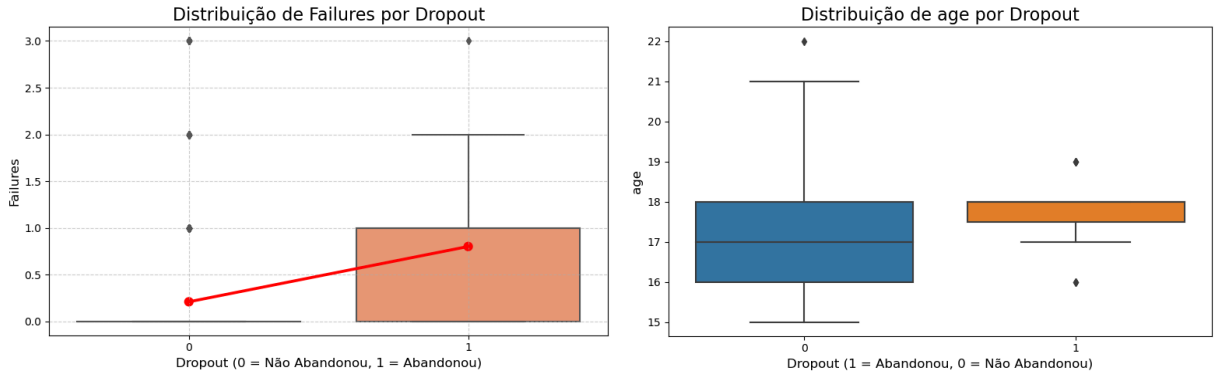

Figura 2. Captura de tela de gráficos feitos no Jupyter Lab. Fonte: Autores

Já a figura acima fornece informações sobre a relação entre “failures” e a idade com o abandono de disciplinas. O gráfico de boxplot à esquerda, que representa o “failures”, mostra que os alunos que abandonaram a disciplina tendem a ter mais reprovações, anteriores, enquanto o gráfico de

dispersão à direita revela que os alunos que abandonaram são ligeiramente mais velhos, sugerindo uma correlação entre idade e abandono.

Embora as variáveis citadas anteriormente sejam importantes nas análises descritivas e preditivas descritas, uma variável se destacou mais. A variável “failures” se revelou uma constante significativa em vários modelos de Aprendizado de Máquina e análises de correlação, demonstrando sua relevância no desempenho acadêmico e no abandono de disciplinas. Em todos os modelos treinados, “failures” mostrou-se como um dos preditores mais robustos e confiáveis, influenciando diretamente tanto as notas finais quanto a probabilidade de abandono de disciplinas. Sua consistência como um fator determinante destaca a necessidade de intervenções direcionadas para alunos que enfrentam reprovações anteriores, o que pode ser crucial para a implementação de estratégias de suporte e melhoria acadêmica.

CONCLUSÕES

Este estudo buscou identificar variáveis-chave que influenciam o desempenho acadêmico e a evasão de disciplinas, destacando a necessidade de atenção especial para alunos com histórico de reprovações. Os resultados evidenciam a importância da variável “failures” (reprovações) como um fator significativo no desempenho e no abandono de disciplinas. Além disso, iniciar o primeiro ano com boas notas (“G1”) e manter um desempenho consistente no segundo ano (“G2”) são cruciais para o sucesso dos alunos. A análise sugere que entender as causas subjacentes a essas dificuldades é fundamental para implementar estratégias eficazes de apoio e prevenção. Futuros trabalhos podem explorar intervenções específicas que abordem os fatores de risco identificados, como suporte educacional e melhoria nas relações familiares, contribuindo para um ambiente de aprendizado mais inclusivo e eficaz.

CONTRIBUIÇÕES DOS AUTORES

Bianca Carvalio Antonietti contribuiu na concepção do projeto, curadoria e análise dos dados, e na redação do artigo. Lucas Bueno Ruas de Oliveira atuou como orientador, fornecendo o material de estudo e apoio na redação do artigo. Ambos os autores participaram da revisão final do trabalho e aprovaram a versão submetida.

AGRADECIMENTOS

Gostaria de agradecer o professor Lucas Bueno Ruas de Oliveira pela supervisão e pelos ensinamentos ao longo deste projeto. Agradeço também ao professor Jorge Francisco Cutigi pela assistência com materiais de estudo e fornecimento do conjunto de dados, que foram cruciais para o desenvolvimento deste trabalho. Este projeto foi conduzido com o apoio do Programa de Educação Tutorial (PET) do Ministério da Educação (MEC).

REFERÊNCIAS

DHAR, V. Data Science and prediction: *Communications of the ACM.* v. 56, n. 12, 2013.

INEP. *Indicadores da Educação Básica 2023*. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2023.

GRUS, Joel. *Data Science from Scratch: First Principles with Python*. O'Reilly Media, 2015.

HARRIS, Joel. *Data Science do Zero: Noções Fundamentais com Python*. São Paulo: Novatec Editora, 2020.

OECD. *Education at a Glance 2013: OECD Indicators*. Paris: Organisation for Economic Co-operation and Development, 2013.

SIEMENS, George; BAKER, Ryan. Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK)*. Vancouver, 2012.