Protocolo de Revisão Sistemática de Literatura

Equipe: Gian Giovanni Rodrigues da Silva,
Jessyca Jordanna Barroso de Moraes,
Tammy Hikari Yanai Gusmão,
Thalita Naiara Andre Alves.

1. Tema

Análise preditiva para identificação de fatores da evasão no ensino superior.

2. Sub-tema

Utilização de técnicas de *machine learning* no mapeamento de fatores que influenciam a evasão no ensino superior.

3. Perguntas de Pesquisa (RQ)

I. Quais fatores motivam a evasão dos alunos de graduação?

II. Qual ou quais cursos possuem maiores índices de evasão?

III. Existem pesquisas na área da educação utilizando *machine learning*?

IV. Quais as técnicas de ML utilizadas na análise preditiva?

4. Tipos de busca

Buscas Automáticas: IEEE, Scopus.

5. Strings de Busca

IEEE Xplore (período de publicação 2015 a 2021):

(("drop out" OR "withdraw*") AND ("education*" OR "college" OR "university") AND ("machine learning" OR "data science" OR "predictive analy*"))*

Total de Resultados retornados: 58.

Scopus (período de publicação 2015 a 2021):

TITLE-ABS-KEY (("dropout" OR "dropping out" OR "withdraw") AND ("education" OR "college" OR "university") AND ("machine learning" OR "data science" OR "predictive analysis"))

Total de Resultados retornados: 194.

6. Critérios de Inclusão e Exclusão

Inclusão:

- Trabalhos que contenham os seguintes termos (e seus derivados) no título, abstract ou palavra-chave: (1) *machine learning*, *data science, data mining* ou *predictive analysis*, (2) *college*, *education*, ou *university*, (3) *dropout*, *withdraw*.

Exclusão:

Trabalhos que:

1. não tenham relação com Ciência de Dados e a área de Educação.
2. não tratam ou incluem uma análise preditiva.
3. não tratam de evasão no ensino superior.
4. não foram escritos em inglês ou português.
5. estão disponíveis somente na forma de resumos/pôsteres e apresentações.
6. são literatura cinzenta, incluindo "white papers", teses e artigos que não foram revisados por pares.

No total, foram aceitos 50 artigos. Dos quais 13, 5 da IEEE e 8 da Scopus, foram selecionados como amostra.

7. Questões de Qualidade

- QQ1: O trabalho utiliza ML como uma ferramenta de predição da evasão no ensino?
- QQ2: O trabalho é algum tipo de experimento ou estudo de caso?

- QQ3: O trabalho foi feito sobre alunos de graduação?
- QQ4: O trabalho identifica fatores/razões responsáveis pela evasão?

Trabalhos obtidos através do IEEE:

IEEE1: "Predictive analytics using data mining technique" **- H. Gulati**

IEEE2: "Machine learning model for detecting high school students as candidates for drop-out from a study program" **- Đ. Pašić; D. Kučak**

IEEE3: "Predicting Dropout Using High School and First-semester Academic Achievement Measures" **- B. Kiss; M. Nagy; R. Molontay; B. Csabay**

IEEE4: "Detection of Potentially Students Drop Out of College in Case of Missing Value Using C4.5" **- S. Mutrofin; A. M. Khalimi; E. Kurniawan; R. V. H. Ginardi; C. Fatichah; Y. A. Sari**

IEEE5: "Educational Data Mining: Analysis of Drop out of Engineering Majors at the UnB - Brazil" **- R. da Fonseca Silveira; M. Holanda; M. de Carvalho Victorino; M. Ladeira**

Trabalhos	QQ1	QQ2	QQ3	QQ4	Total Trabalho
IEEE1	1	1	1	1	4
IEEE2	1	1	0	0	2
IEEE3	1	1	1	0	3
IEEE4	1	1	1	0	3
IEEE5	1	1	1	1	4
Total QQ	5	5	4	2	16

Tabela 1. Artigos Retornados pela IEEE.

Trabalhos obtidos através da Scopus:

SC1: "Predicting nursing baccalaureate program graduates using machine learning models: A quantitative research study**" - Hannaford L., Cheng X., Kunes-Connell M.**

SC2: "Attributes selection using machine learning for analysing students' dropping out of university: A case study**" - Pehlivanova T.I., Nedeva V.I.**

SC3: "Predicting student academic performance by means of associative classification**" - Cagliero L., Canale L., Farinetti L., Baralis E., Venuto E.**

SC4: "Drop-Out Prediction in Higher Education Among B40 Students**" - Sani N.S., Nafuri A.F.M., Othman Z.A., Nazri M.Z.A., Nadiyah Mohamad K.**

SC5: "Application of decision trees for detection of student dropout profiles" - **Oliveira M.M.D., Barwaldt R., Pias M.R., Espindola D.B.**

SC6: "Adapting the Score Prediction to Characteristics of Undergraduate Student Data" - **Mai T.L., Do P.T., Chung M.T., Le V.T., Thoai N.**

SC7: "Dropout prediction system to reduce discontinue study rate of information technology students" - **Limsathitwong K., Tiwatthanont K., Yatsungnoen T.**

SC8: "Predictive modelling of student dropout using ensemble classifier method in higher education" - **Hutagaol N., Suharjito.**

Trabalhos	QQ1	QQ2	QQ3	QQ4	Total Trabalho
SC1	1	1	1	1	4
SC2	1	1	1	1	4
SC3	1	1	1	1	4
SC4	0	1	1	0	2
SC5	1	1	1	1	4
SC6	1	1	1	1	4
SC7	1	1	1	1	4
SC8	1	1	1	1	4
Total QQ	7	8	8	7	30

Tabela 2. Artigos Retornados pela Scopus.

Quais perguntas de qualidade foram respondidas?

- Dos cinco artigos apresentados na Tabela 1 (retornados pela IEEE):
    - IEEE2, 3 e 4 não respondem a pergunta QQ4, onde procura-se saber se o trabalho apresenta motivos ou fatores de evasão dos alunos.
    - Somente IEEE2 a pergunta QQ3, que deseja saber se houve participação de alunos de graduação do experimento realizado.
- Dos oito artigos apresentados na Tabela 2 (retornados pela Scopus):
    - SC1, 2, 3, 5, 6, 7 e 8 responderam todas as perguntas de pesquisa.

- SC4 respondeu somente às perguntas QQ2 e 3.

Quais os trabalhos mais aderentes à Revisão Sistemática?

Todos aqueles que somaram pontuação 4: IEEE1 e IEEE5, SC1, SC2, SC3, SC5, SC6, SC7 e SC8.

Quais foram suas conclusões a respeito dessas perguntas?

Apesar de terem sido selecionados artigos que se mostraram promissores à primeira vista para aplicação de QQ, alguns não continham informações que respondiam a todas as questões, como pode ser visto na Tabela 1 e na Tabela 2. O que levou os pesquisadores a incluir estes artigos na amostra foi o recebimento do critério de inclusão CI1. É possível que haja a necessidade de adicionar critérios de inclusão para estas situações.

Estatísticas:

De 58 artigos adquiridos na IEEE, 21 passaram no 1º filtro - ou seja, receberam CI1 como critério. Destes 21, 5 foram escolhidos para serem aplicados às Questões de Qualidade. Somente 2 artigos receberam pontuação 4 e passaram para o 2º filtro.

De 194 artigos adquiridos na Scopus, 30 passaram no 1º filtro. Destes 30, 8 foram escolhidos para serem aplicados às Questões de Qualidade. Desses artigos, 7 receberam pontuação 4 e passaram para o 2º filtro.

8. Extração de Dados / Síntese: Resultados da Revisão

Durante a extração dos dados, buscamos identificar informações que respondam às perguntas de pesquisa, como especificados na seção 3: fatores que motivam a evasão de alunos de graduação (independente do estudo de caso ter incluído alunos da pós-graduação), cursos com maiores índices de evasão ou o curso especificado no estudo de caso, se o trabalho é de fato uma pesquisa (análise) na área da educação que faz uso de *machine learning* e, se é, especificação das técnicas de ML utilizadas.

Após a extração, foram feitas as seguintes observações:

- As técnicas baseadas em ML mais populares foram: KNN, Random Forest, Naive Bayes, Redes Neurais, e Árvores de Decisão. Em alguns trabalhos foram especificadas as variantes de algumas das técnicas mencionadas.
- Os cursos de graduação mais estudados foram da área de Ciências Exatas, como engenharias, física, matemática, etc. Cursos das áreas de Saúde e de Humanas também foram incluídos em alguns estudos.
- Fatores de evasão variam de cunho socioeconômico, estado matrimonial, e rendimento acadêmico no ensino médio e/ou durante a graduação.

Os dados extraídos dos artigos são apresentados nas tabelas da seção 10.

9. Conclusões

Cada uma das etapas foi essencial para a seleção dos artigos, visto que, as mesmas foram mostrando trabalhos relevantes para a pesquisa. Foi aplicado o primeiro filtro (seleção) em 252 artigos, nos quais apenas 50 foram aceitos. Devido ao tempo e quantidade de artigos, escolheu-se 13 para aplicação do segundo filtro (extração). Destes, 9 receberam nota 4 nesse filtro e mesmo assim, alguns não conseguiram responder todas as questões de pesquisa. Portanto, é possível que haja a necessidade de adicionar critérios de inclusão ou de qualidade.

Grande parte dos artigos envolviam cursos de graduação da área de exatas e técnicas de *machine learning* como Árvore de Decisão e Random Forest. Além disso, fatores socioeconômicos, estado matrimonial, e rendimento acadêmico no ensino médio e/ou durante a graduação foram alguns dos fatores apontados pelos artigos que influenciam a evasão no ensino superior.

10.Anexos

Trabalhos do IEEE que receberam nota 4 nas Questões de Qualidade

Trabalho	RQ1	RQ2	RQ3	RQ4

IEEE1	Programa, area, sexo, data de nascimento, , créditos e medium.	-	Sim	Rule based classification algorithm (Jrip, NNge, conjuctive rule, DTNB, PART, etc) e Decision tree algorithm (J48, NBTree, REPTree, SimpleCart, etc)
IEEE5	Quantidade de vezes que a disciplina de Física 1 foi feita e quantidade de solicitação de licença do curso.	Cursos Não Especificados de Engenharia.	Sim	Generalized Linear Model (GLM), Gradient Boosting Machine (GBM) and Random Forest (RF)

Trabalhos do Scopus que receberam nota 4 nas Questões de Qualidade

Trabalho	RQ1	RQ2	RQ3	RQ4
SC1	GPA de faculdade, notas do curso, créditos obtidos.	Enfermagem	Sim	C5.0, Random Forest, xgboost, Redes Neurais, SVM, Naive Bayes, KNN, Regressão Logística.
SC2	Nível de estresse, idade, ano/período de graduação, comunicação professor x aluno, nível de educação dos pais, satisfação com o trabalho, status matrimonial do aluno, relacionamento com outros estudantes.	Cursos Não Especificados de Engenharia.	Sim	Naive Bayes
SC3	Notas baixas no ensino médio ou no teste de	Cursos Não Especificados de	Sim	L^3, C4.5, LIBSVM, Naive Bayes,

	vestibular, ser mais velho que os demais estudantes, uso limitado de material educativo, não ter começado a trabalhar desde o início do semestre, não focar em uma matéria por vez durante a semana de provas.	Engenharia.		KNN, Random Forest.
SC5	Restrições econômicas e financeiras, baixo desempenho acadêmico, desorientação vocacional e profissional e dificuldade de adaptação ao ambiente universitário	Cursos das faculdades de Ciências Naturais, Ciências da Saúde, Educação e Artes Além de áreas com fundamentos matemáticos como introdução às ciências naturais, formação básica, pedagogia, economia e contabilidade	sim	Decision tree
SC6	-	-	sim	Decision Tree, Random Forest, Rule Induction, Regression e Neural Network
SC7	As notas baixas dos cursos	-	sim	Decision Tree e Random Forest

	fundamentais de idioma japonês e de tecnologia da informação			
SC8	Gênero, idade, nível de educação e renda dos pais, notas dos trabalhos de casa e testes, total de crédito do curso e GPA (média das notas)	-	sim	KNN, Decision Tree e Naive Bayes