Desenvolvimento de um *Score* para análise de risco de evasão de estudantes do Ensino Superior baseado em Aprendizado de Máquina

Robinson Crusoé da Cruz
Centro Universitário do Planalto de Araxá
Araxá, MG, Brasil
robinsoncruz@uniaraxa.edu.br

Renato Correa Juliano
Centro Universitário do Planalto de Araxá
Araxá, MG, Brasil
renatocorrea@uniaraxa.edu.br

Alinne Cristinne Correa Souza
Universidade Tecnológica Federal do Paraná
Dois Vizinhos, PR, Brasil
alinnesouza@utfpr.edu.br

Francisco Carlos Monteiro Souza
Universidade Tecnológica Federal do Paraná
Dois Vizinhos, PR, Brasil
franciscosouza@utfpr.edu.br

ABSTRACT

Dropping out of Higher Education contributes to great social, economic and academic loss. Among the main reasons for dropping out are the student's difficulty in following the content, the structure proposed by the course and the lack of financial resources. In recent years, several studies have emerged to try to identify groups of students at risk of dropping out, either by identifying the factors that can contribute to dropout, or by creating classifiers based on Machine Learning. However, researches focus essentially on categorical indicators, that is, with binary results, which denote that the student is or is not in the risk group. This type of analysis is important, however, it does not show the variation in the student's performance during their academic life, in addition to not offering a score within a performance score. Differently, this project use Machine Learning techniques in the creation of a Score, in order to provide a thermometer to analyze how close the student is or not to the dropout group. Preliminary results are promising, because when using *KNN* to create the Score, it was possible to develop a Score with the best result of hyperparameters found in the experiments.

KEYWORDS

Dropout from Higher Education, Score, Machine Learning, KNN

1 INTRODUÇÃO

Alunos que iniciam os estudos no Ensino Superior e não finalizam, geram uma grande perda social, econômica e acadêmica. Entre os principais motivos do abandono estão a dificuldade do aluno em acompanhar o conteúdo, a estrutura proposta pelo curso e a falta de recursos financeiros [1].

A evasão do Ensino Superior é um assunto amplamente discutido pelas Instituições de Ensino Superior (IES) e pelos órgãos governamentais com objetivo de definir estratégias para minimizar este problema [1]. Em 2010, o Ministério da Educação (MEC) realizou uma pesquisa com estudantes do Ensino Superior tanto da rede pública quanto da rede privada no período de 2010 até 2014 a fim de verificar a permanência desses estudantes no mesmo curso de ingresso. De acordo com os resultados, 49% dos estudantes que começaram no Ensino Superior em 2010 abandonaram seus cursos e o desempenho acadêmico está entre os principais motivos que os levaram à desistência [2].

Algumas pesquisas [3, 4], foram conduzidas com a finalidade de auxiliar a identificação dos motivos da evasão no Ensino Superior. Em grande parte dessas pesquisas, foi aplicada a técnica de Aprendizado de Máquina (*Machine Learning*) para identificar fatores que possam contribuir com a evasão e na criação de classificadores.

O trabalho [3] visou identificar grupos de risco de evasão. Para isso, foram aplicadas técnicas de mineração de dados com algoritmo de classificação J48 nos dados dos alunos de doze cursos da Universidade Federal do Rio Grande (FURG), obtendo uma acurácia de 90,70%. Por outro lado, na pesquisa [4] foi utilizada a técnica de Regressão Logística para identificar grupos com risco de evasão. Para isso, foram analisados dados acadêmicos e demográficos de 32.538 alunos da Universidade de Washington nos EUA, obtendo uma acurácia de 66,59%. Esse tipo de abordagem tem como objetivo principal, identificar grupos de risco, com classificações binárias, ou seja, apresentam resultados se o aluno está ou não no grupo de risco de evasão.

A busca por classificadores é promissora. No entanto, com esta técnica não é possível identificar a situação de um aluno em relação ao grupo de risco. Por exemplo, ao aplicar um classificador, um aluno pode ser classificado no grupo sem risco de evasão. Por outro lado, com *Score*, além de classificar o aluno como pertencente ao grupo sem risco de evasão, é possível definir se ele está próximo ou não do grupo de risco.

O presente artigo visa realizar uma análise do risco de evasão de estudantes do Ensino Superior por meio da definição de um *Score*. Este *Score* foi desenvolvido com a utilização dos dados dos alunos de uma Universidade sem fins lucrativos (Fundação). Por fim, foi conduzido um experimento a fim de avaliar a eficácia do *Score* proposto.

Dessa forma, as principais contribuições do presente trabalho podem ser sumarizadas como: *i)* utilização do algoritmo de aprendizado de máquina KNN para classificação dos alunos como desistentes e não desistentes; *ii)* definição de um score para analisar a zona de risco de evasão com base nos resultados do algoritmo; e *iii)* realização de um experimento para verificar a eficácia da proposta.

Este artigo está organizado da seguinte forma: na Seção 2 são abordados trabalhos relacionados que contribuíram para o cenário

científico desta pesquisa. Na Seção 3 é apresentada a forma como foi conduzido o experimento. Na Seção 4 são apresentados e discutidos os resultados alcançados. Por fim, na Seção 5 as considerações finais e trabalhos futuros são apresentados.

2 TRABALHOS RELACIONADOS

Alguns trabalhos têm discutido a criação de classificadores e métodos para analisar o grau de evasão do Ensino Superior. No geral, esses trabalhos propõem a criação de classificadores binários, para definir se um aluno está ou não no grupo de risco de evasão.

O trabalho de da Silva and Adeodato [5] analisou dados de oito cursos da Universidade Federal do Pernambuco (UFPE), entre os anos de 1998 e 2018. Uma amostra de 11.036 estudantes e 415.327 resultados da performance dos estudantes foi analisada. Foi aplicado um método de regressão e outras técnicas com objetivo de produzir um modelo para predição e auxílio na análise da evasão dos cursos. Nos resultados os autores sugerem a utilização de dados socioeconômicos, culturais e comportamentais. Análise dos dados supracitados pelos autores, fazem parte deste estudo, conforme apresentados na Tabela 3.

Em Manhães et al. [6] foi utilizado o modelo *Naive Bayes* para visualizar os fatores que distinguem os alunos que tiveram sucesso ou fracasso em seus cursos. Nesse contexto, foram analisados dados referentes ao período de 2003 a 2014 de 155 cursos de graduação de 28 unidades da Universidade Federal do Rio de Janeiro (UFRJ). Sete algoritmos de classificação foram comparados antes da escolha do *Naive Bayes*.

Rabelo et al. [7] realizaram uma revisão sistemática sobre os motivos que levam a evasão dos alunos e as técnicas mais utilizadas para auxiliar na análise da evasão. Apesar do estudo estar relacionado com EaD e seja diferente da proposta deste trabalho, seus resultados serviram para auxiliar na análise das variáveis a serem analisadas e os algoritmos de classificação utilizados.

Lanes and Alcântara [3] realizaram uma análise dos dados da Universidade Federal do Rio Grande (FURG) e aplicaram um método para analisar a evasão. Durante o estudo foi utilizado o algoritmo J48. A análise foi conduzida com alunos que tinham no mínimo um ano de curso e com dados demográficos e acadêmicos. De acordo com o resultado do estudo, foi possível obter uma acurácia de 90% na classificação do grupo de risco de evasão.

Em Aulck et al. [4] foram analisados dados acadêmicos e demográficos de 32.538 alunos da Universidade de Washington nos EUA, com a obtenção de uma acurácia de 66% ao utilizar a Regressão Logística para identificar grupos de risco de evasão.

Em de Almeida Teodoro and Kappel [9] foi realizado um estudo com dados de uma base pública disponibilizada pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), com a utilização de cinco técnicas de Aprendizado de Máquina: *Naive Bayes*, *K-Nearest Neighbors*, Árvores de Decisão, *Random Forest* e Redes Neurais. Como parte dos resultados, os autores realizaram uma comparação entre as técnicas. *Random Forest* foi a que obteve o melhor desempenho.

Em de Brito et al. [8] foram analisadas duas base de dados: (i) acadêmica contendo informações dos alunos; (ii) dados demográficos e socioeconômicos do Brasil, por meio dos indicadores do IBGE. Nesta pesquisa foram analisadas as acurácias de acordo com as áreas dos cursos, obtendo um acurácia geral de 70%. Por fim, na pesquisa de e Robinson Noronha e Celso Kaestner [10] foi realizado um estudo de seleção de variáveis para auxiliar na previsão da evasão do Ensino Superior. As variáveis desse estudo possuem similaridades com as utilizadas nesta abordagem.

Na Tabela 1 é apresentado um resumo dos trabalhos relacionados apresentados. A primeira coluna lista os estudos relacionados, os algoritmos e/ou técnicas utilizadas (segunda coluna), acurácia alcançada (terceira coluna), o período que a pesquisa foi realizada (quarta coluna), sujeitos (quinta coluna) e o número de cursos (sexta coluna). Para as informações não disponibilizadas pelos estudos foi utilizado "Não Informado (NI)" e quando "Não Aplicadas (NA)". Vale ressaltar, que os trabalhos que utilizaram classificadores, no geral, apresentaram a acurácia geral e nos resultados deste projeto, a precisão de cada classe (1 - Não Desistente e 2 - Desistente) foi um fator essencial para definir quais hiperparâmetros seriam utilizados na criação do *Score*.

As pesquisas apresentadas nesta Seção, focam essencialmente em indicadores categóricos. Ou seja, com resultados binários, que denotam que o aluno está ou não no grupo de risco. Os autores, em linhas gerais, indicam pontos que precisam ser pesquisados para aprimorar a identificação de estudantes com risco de evasão. Por exemplo, o tratamento do desbalanceamento entre as classes e utilização de novas variáveis e métodos. Nesta perspectiva, este trabalho não visa apenas a classificação, mas propor uma nova abordagem que classifique o estudante de acordo com seu *Score*.

3 METODOLOGIA EXPERIMENTAL

Nesta seção serão detalhados os experimentos conduzidos, visando avaliar a eficácia do *Score* proposto para identificar o risco de evasão ao longo da vida acadêmica do estudante no Ensino Superior.

3.1 Conjunto de dados

Para a condução do experimento foram utilizados dados de uma Universidade Privada Sem Fins Lucrativos. Este tipo de Universidade consiste em uma organização não governamental, mantida por uma Fundação, constituída por tradicionais universidades confessionais, filantrópicas e comunitárias, onde espera-se que os recursos obtidos por meio de mensalidades, sejam revertidos na Instituição [11].

A amostra utilizada nos experimentos contém dados de cinco semestres letivos, entre o primeiro semestre de 2019 até o primeiro semestre de 2021, com um total de 9.801 registros de estudantes distribuídos entre 24 cursos presenciais, com 20 variáveis acadêmicas e 14 variáveis relacionadas à pesquisa de perfil.

3.2 Configuração do experimento

Inicialmente foram realizadas análises exploratórias dos dados. Em seguida, os experimentos foram divididos em três fases: *(I)* aplicação do classificador *KNN* para analisar a classificação dos estudantes desistentes e não desistentes com a utilização das variáveis da Tabela 2; *(II)* aplicação do classificador *KNN* com as variáveis da fase I (Tabela 2) em conjunto com as variáveis da pesquisa sobre o perfil do aluno (Tabela 3), com mesmo objetivo da fase I, com a particularidade de analisar se a pesquisa realizada com os estudantes pode melhorar os resultados; *(III)* desenvolvimento do *Score* de acordo com o melhor resultado ao aplicar o classificador KNN (fases I e II).

Table 1: Visão geral dos trabalhos relacionados.

Estudos	Algoritmos	Acurácia	Período	Amostra	Número de cursos
Manhães et al. [6]	Naive Bayes	NA	2003 a 2014	NI	15
Lanes and Alcântara [3]	J48	90,70%	2012 a 2017	916	12
Aulck et al. [4]	Regressão	66,59%	1998 a 2006	32538	NI
da Silva and Adeodato [5]	Regressão	NA	1998 a 2018	11036	NI
de Brito et al. [8]	*Random Florest*	70,00%	a partir de 2010	51175 e 31482	59
Proposta	*KNN*	83,00%	2019 a 2021	9801	24

Table 2: Variáveis analisadas no período 2019-1 a 2021-1

VARIÁVEL	DESCRIÇÃO	VALORES
Serie	Série que o aluno estava cursando (ocupando vaga)	1 a 10
IdadeDias	Idade (em dias) que o aluno tinha quando finalizou o semestre	6.000 a 36.500
PrimeiroBimestre	Média da Nota obtida no primeiro bimestre	0 a 10,00
SegundoBimestre	Média da Nota obtida no segundo bimestre	0 a 10,00
ACQG	Nota da Avaliação de Controle de Qualidade da Graduação	0 a 2,00
MediaProvaFinal	Médida da Nota obtida na prova final	0 a 10,00
MediaNotaFinal	Médida da Nota final do aluno	0 a 10,00
PorcFalta	Porcentagem de falta do aluno	0 a 1
QtdeDisciplinaCursando	Número de disciplina que o aluno estava cursando no semestre	0 a 10
QtdeAprovacoes	Número de disciplinas que o aluno foi aprovado no semestre	0 a 10
QtdeReprovacoes	Número de disciplinas que o aluno foi reprovado no semestre	0 a 10
CargaHorariaTotal	Carga horária total das disciplinas	20 a 800
CargaHorariaTeorica	Carga horária teórica total das disciplinas	0 a 800
CargaHorariaPratica	Carga horária prática total das disciplinas	0 a 800
QtdeMatCalculo	Número de disciplinas de matemática e cálculo que o aluno estava cursando	0 a 10
TotalAcessoLMS	Total de acesso em disciplinas no ambiente virtual	0 a N
MoraCidade	Define se o aluno mora na cidade onde fica a Universidade	1=Sim e 2=Não
Sexo	Sexo do aluno	0=Feminino, 1=Masculino e 2=Não definido
TipoEnsinoMedio	Tipo do Ensino Médio que o aluno cursou	1=Público e 2=Privado
EstadoCivil	Estado civil do aluno	1=Solteiro, 2=Casado, 3=Divorciado e 0=Outros
Situacao/Classe	Situação do aluno durante ou no final do semestre	1=Não Desistente e 2=Desistente

3.2.1 Fase I - Aplicação do classificador KNN

. Nesta fase, um total de 20 variáveis foram utilizadas na aplicação do algoritmo *KNN* [12] para analisar a classificação dos estudantes desistentes e não desistentes, conforme apresentadas na Tabela 2. Algumas variáveis como *QtdeDisciplinaCursando* e *QtdeMatCalculo* representam uma somatória ou média dos resultados de todas as disciplinas que o aluno cursou no semestre. A variável *QtdeDisciplinaCursando* representa o número de disciplinas que o aluno cursou no semestre, *QtdeMatCalculo*, representa o número de disciplinas com conteúdo de matemática. Além disso, algumas variáveis foram adicionadas para analisar o perfil do aluno, por exemplo, sexo, se o aluno mora na cidade onde a universidade está situada, idade que o aluno tinha no final do semestre letivo. Nas fases I e II, a amostra foi dividida em 20% para teste e 80% para treino. Na aplicação dos testes foram utilizadas a linguagem Python e a biblioteca scikit-learn.

3.2.2 Fase II - Análise com dados da pesquisa de Perfil

. Nos experimentos da primeira e segunda fase, o valor da Análise do Componente Principal (PCA), foi um fator preponderante, pois pode contribuir com a redução da dimensionalidade [13]. Nesta fase as 20 variáveis, apresentadas na Tabela 2, foram unificadas com a pesquisa relacionada ao perfil do estudante, conforme apresentado na Tabela 3. Esta pesquisa é realizada todo semestre letivo, desde de 2019/1 e contém 14 perguntas que foram propostas com os objetivos de coletar e identificar informações externas que possam interferir no rendimento acadêmico do aluno, e como consequência, fornecer

Table 3: Variáveis analisadas de acordo com o perfil dos estudantes

ID	PERGUNTA (VARIÁVEL)	VALORES E RESPOSTAS
797	Qual o seu estado civil?	1=Solteiro(a); 2=Casado(a); 3=Separado(a) judicialmente/divorciado(a); 4=Viúvo(a); 5=Outro(a)
798	Você tem filhos?	1=Não; 2=Sim, mas NÃO moram comigo; 3=Sim, moram comigo
799	Você possui acesso à internet em casa?	1=Sim; 2=Não
800	De onde vêm os recursos financeiros para pagamento da sua mensalidade?	1=Bolsa Parcial; 2=Bolsa Integral; 3= Financiamento Estudantil; 4=Familiares; 5=Trabalho; 6=Outros
801	Qual é o seu tempo gasto até chegar até a Universidade?	1=Menos de 10 minutos; 2=De 10 a 30 minutos; 3=Acima de 30 minutos; 4=Não moro em Araxá; 5=Sou aluno EAD
802	Assinale a alternativa que define sua condição de trabalho	1=Não estou trabalhando no momento; 2=Trabalho em casa; 3=Trabalho até 44 horas semanais; 4=Trabalho mais de 44 horas semanais
803	Você exerce atividade(s) relacionada(s) com seu curso, no trabalho ou estágio?	1=SIM, parcialmente; 2=SIM, integralmente; 3=NÃO
804	Assinale o(s) item(s) que descrevem seu tempo e sua(s) preferência(s) de lazer:	1=Saio entre uma a três vezes por semana; 2=Saio entre 3 a 5 vezes por semana; 3=Uso meu tempo de lazer para assistir séries/filmes; 4=Uso meu tempo de lazer em festas, shows ou barzinhos; 5=Não tenho tempo para o lazer
805	Você trabalha no regime de turnos?	1=Sim, por isso às vezes não consigo frequentar as aulas; 2=Sim, todavia não atrapalha a minha frequência às aulas; 3=Não
807	Quantas horas por semana você consegue dedicar aos estudos, excetuando as horas de aula?	1=Nenhuma, apenas participo das aulas, 2=De uma a três horas; 3=De quatro a sete horas; 4=De oito a doze horas; 5=Mais de doze horas
808	Quantos livros você leu nos últimos 6 meses (exceto os acadêmicos):	1=Nenhum; 2=Um livro; 3=Dois livros; 4=Três livros ou mais
809	Quais são suas formas de aprendizagem preferidas?	1=Vídeos; 2=Imagens; 3=Nivelamento; 4=Dinâmicas de grupo; 5=Explicações do professor; 6=Pesquisa de Campo/Visitas Técnicas
810	Quais são suas fontes de complementação (fora da sala de aula) de aprendizagem preferidas?	1=Materiais disponibilizados no AVA; 2=Pesquisas na Biblioteca (Física e/ou Virtual); 3=Noticiários; 4=Pesquisas de textos na internet; 5=Vídeo aulas; 6=Não busco alternativas de aprendizagem fora da sala de aula

insumos para análise da evasão. Por exemplo, nesta pesquisa é questionado o tempo de trabalho semanal, o tempo dedicado aos estudos extraclasse, o hábito de leitura do aluno, quais as preferências nas formas de aprendizagem, entre outros itens conforme apresentados na Tabela 3. Contudo, como a pesquisa não é obrigatória, após a junção entre os dados das Tabelas 2 e 3, a amostra foi reduzida de 9.801 para 2.907 registros.

3.2.3 Fase III - Desenvolvimento do Score

. Nesta fase o modelo de *Score* foi desenvolvido com base nos insumos gerados nas fases anteriores, com a utilização do melhor resultado de hiperparâmetros para o classificador *KNN*.

4 RESULTADOS E DISCUSSÕES

Nesta Seção, os resultados são apresentados de forma estrutural de acordo com as anuências e processos desenvolvidos durantes as três fases dos experimentos. As classes utilizadas nos experimentos foram definidas como: **1-Não Desistente; 2-Desistente**. Antes da condução dos experimentos, foram realizadas análises exploratórias dos dados para identificar fatores e variáveis que podiam ocasionar uma maior ou menor influência nos resultados e se condizem com o esperado dentro do contexto.

Na Figura 1 é apresentada a distribuição dos dados da média das variáveis Nota (*MediaNotaFinal*) e Frequência (*PorcFalta*). Nota-se que as notas pertencentes à classe 1 estão com uma maior concentração acima da nota 6, enquanto que as notas da classe 2 estão dispersas, apesar de ter uma proporção próxima a zero.

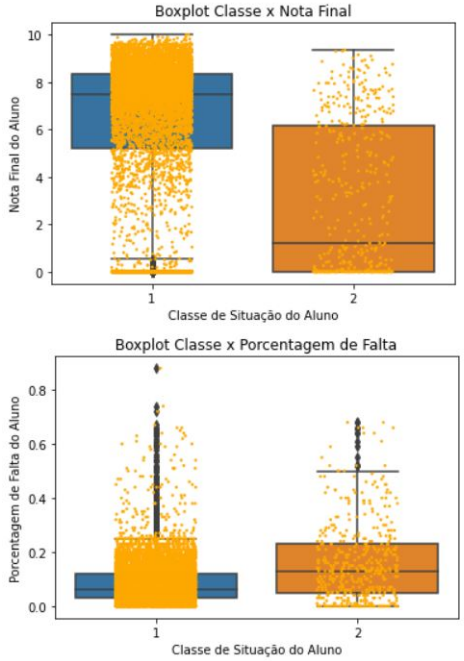

Figure 1: Distribuição da média de notas e frequências

Além disso, também foram analisadas correlações (Figura 2). De acordo com os resultados é notável que existe uma correlação alta entre notas do primeiro e segundo bimestre, indicando que quanto maior a nota do primeiro bimestre, maior a nota do segundo.

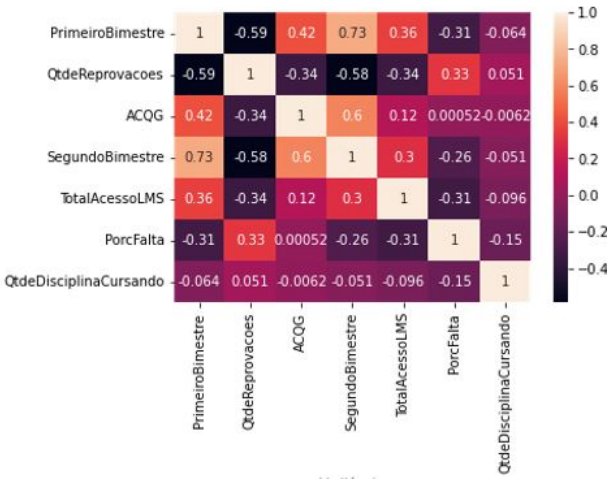

Figure 2: Algumas correlações analisadas

4.1 Fase I - Aplicação do classificador *KNN*

Nesta fase o classificador *KNN* foi aplicado para analisar a classificação dos estudantes desistentes e não desistentes. A amostra de 9.801 registros, sendo 9.337 da classe 1 (Não Desistentes) e 464 da classe 2 (Desistentes). Neste contexto, foram realizados três testes, conforme apresentado na Tabela 4. A acurácia alcançada foi de 0.95, conforme pode ser visualizado nos testes 1,2 e 3. Nos testes apresentados nesta Seção e nos subsequentes, as variáveis foram normalizadas antes da aplicação do classificador.

Table 4: Testes sem balanceamento.

	Qtde Amostra				Precisão			
Teste	**1**	**2**	**PCA**	**k**	**1**	**2**	**Acurácia**	**fi-score**
1	9337	464	Sem	5	0,96	0,69	0,959	0,68
2	9337	464	Sem	21	0,96	0,66	0,957	0,66
3	9337	464	14	21	0,96	0,59	0,958	0,67

O total da amostra possui 95,27% (9337) registros que pertencem a classe 1 (Não Desistentes). Como a literatura define que um desbalanceamento alto entre as classes pode interferir nos resultados dos classificadores [14], foi aplicado um balanceamento com o método *NearMiss* [15] que reduziu de forma aleatória e baseado em distância a amostra da classe 1 (Não Desistente). O resultado do Teste 6 após o balanceamento, conforme apresentado na Tabela 5, indica uma melhora na precisão da classe 2 (Desistente), na quantidade de vizinhos ($k = 5$) e na quantidade de Componentes Principais ($PCA = 9$). Nota-se que depois de balanceada, a variação do k entre os Testes 4, 5 e 6, foi inferior aos Testes 1, 2 e 3.

Table 5: Teste com balanceamento.

	Qtde Amostra				Precisão			
Teste	**1**	**2**	**PCA**	**k**	**1**	**2**	**Acurácia**	**f1-score**
4	464	464	Sem	5	0,73	0,92	0,801	0,80
5	464	464	Sem	3	0,74	0,91	0,806	0,80
6	464	464	9	5	0,74	0,92	0,833	0,83

4.2 Fase II - Análise com dados da pesquisa de Perfil

Conforme os trabalhos relacionados identificados, os fatores extraclasse podem interferir no rendimento acadêmico do estudo, e como consequência, influenciar na evasão. Neste contexto, foi realizada a junção entre as variáveis analisadas na fase anterior (Tabela 2) com os resultados da pesquisa (Tabela 3). O principal objetivo foi analisar se o acréscimo das variáveis da pesquisa, resultariam em melhores resultados. Contudo, como a pesquisa não era obrigatória, ao realizar a junção a amostra foi reduzida para 2.907 registros.

Nos primeiros experimentos desta Fase foram realizadas análises sem a aplicação de balanceamento, conforme apresentadas na Tabela 6. Nota-se que ocorreu uma melhora na acurácia (0.96), contudo, ao analisar a precisão da classe 2 (Desistente), os resultados foram inferiores aos experimentos anteriores.

Com a junção dos dados, 96% (2816) dos registros pertencem a classe 1 (Não desistente). Então foi realizado um balanceamento entre as classes que reduziu a amostra total para 182 registros, fato este,

Table 6: Testes sem balanceamento e com pesquisa de perfil.

	Qtde Amostra				Precisão			
Teste	**1**	**2**	**PCA**	**k**	**1**	**2**	**Acurácia**	**f1-score**
7	2816	91	Sem	5	0,97	0,38	0,965	0,57
8	2816	91	8	3	0,98	0,45	0,967	0,62

que pode ter interferido nos resultados do classificador, conforme apresentados na Tabela 7. A redução da amostra foi preponderante para que estes resultados fossem desconsiderados. Os resultados alcançados nos experimentos apresentados indicam que a utilização das variáveis coletadas na Pesquisa (Tabela 3) não contribuíram com os resultados.

Table 7: Testes com balanceamento e com pesquisa de perfil.

	Qtde Amostra				Precisão			
Teste	**1**	**2**	**PCA**	**k**	**1**	**2**	**Acurácia**	**f1-score**
9	91	91	Sem	5	0,62	0,88	0,675	0,64
10	91	91	2	9	0,70	0,79	0,729	0,72

4.3 Fase III - Desenvolvimento do *Score*

No desenvolvimento do modelo de *Score*, foi utilizado o melhor resultado dos experimentos da Fase I e II. A escolha do algoritmo *KNN* foi de suma importância na construção do modelo, pois como este algoritmo é baseado em distância entre vizinhos (k) e, espera-se que, se um aluno possui uma maior quantidade de vizinhos com risco de evasão (classe 2), então maior será a probabilidade dele pertencer a esta classe. Por outro lado, se ele possui uma menor quantidade de vizinhos da classe 2, ele poderá ter um menor risco de evasão, apesar de pertencer a classe 2.

Table 8: Matriz de confusão e variáveis.

Classe			Média das Variáveis		
Classificada no KNN	Correta	Total	Total AcessoLMS	QtdeDisciplina Cursando	Qtde Aprovacoes
2	2	**69**	31,33	5,55	1,38
2	1	**7**	41,86	5,00	4,71
1	1	**86**	44,71	5,45	4,70
1	2	**24**	48,63	5,29	4,00

A Tabela 8, apresenta detalhes dos resultados do classificador *KNN*, conforme os hiperparâmetros presentes na Tabela 5 referentes ao Teste 6 ($PCA = 9$ e $k = 5$). Este resultado, foi selecionado para o desenvolvimento do modelo de *Score*, devido ao balanceamento, valor do $f1 - score$, maior amostra de dados, melhor precisão das classes e os valores para k e PCA.

Na Figura 3 é apresentada a proposta do modelo que utiliza os valores de k como parâmetros para definir os limites e valores do *Score*. O valores do *Score* podem variar entre 1 (maior risco de evasão) e 1000 (menor risco de evasão). Como o melhor resultado para o classificador foi $k = 5$, o *Score* foi divido em 5 intervalos com 200 pontos ($1000/k$). Se um determinado aluno após a sua classificação, possuir $k = 3$ da classe 1 ou 2, ele será classificado no intervalo e 401 a 600. Este ponto de convergência indica que o aluno está na transição entre as classes.

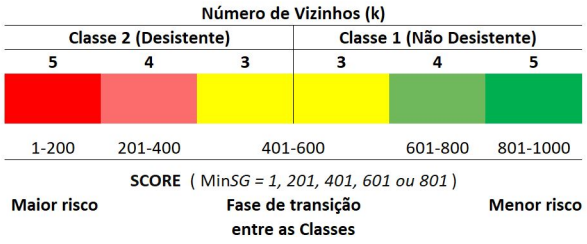

Figure 3: Modelo de *Score* com $k = 5$

Em relação a classe 2 (Desistente), se após a classificação um aluno for classificado na classe 2 e possuir 5 vizinhos (k) da classe 2, considera-se que além de pertencer a classe 2, possui um alto risco de evasão. Por outro lado, se um aluno for classificado na classe 2 e possuir 3 vizinhos (k) da classe 2, logo possui um menor risco de evasão, pois existe uma proximidade com 2 alunos da classe 1 (Não Desistente). Esta mesma análise, pode ser aplicada aos alunos que sejam classificados na classe 1 (Não Desistentes), pois se possuem um valor de $k = 3$, pertencem a classe 1, contudo, possuem um maior risco de evasão, pois estão na transição entre as classes 1 e 2.

Os resultados apresentados na Figura 3, indicam que é possível classificar o aluno de acordo com sua classe e *Score*. No entanto, nos experimentos foram identificadas variáveis que podem ser utilizadas para auxiliar no desenvolvimento do *Score*. Na Tabela 8, por exemplo, nota-se que alunos classificados na classe 2 (Desistente), possuem menor quantidade de aprovações e acessos no Ambiente Virtual (LMS).

Com estes indicativos, surgiu a necessidade de considerar a importância das variáveis da Tabela 2 para auxiliar a refinar o *Score* entre os intervalos. Por exemplo, um aluno pode pertencer a classe 1, com valor de $k = 5$, com classificação em um intervalo entre 801 e 1000. No entanto, com o modelo apresentado na Figura 3 não é possível definir o valor exato do *Score*, mas ao aplicar a Equação 1 de ajuste do *Score* , espera-se que seja possível definir este valor.

$$ScoreFinal = MinSG + \sum_{i=1}^{n} (v_i * p_i * 10) \quad (1)$$

onde n representa a quantidade de variáveis analisadas, v a variável normalizada entre 0 e 1, p o peso da variável entre 0 e 19,90, conforme a sua importância dentro do contexto e $MinSG$ o menor [1] *Score* do aluno de acordo com a classe e número de k. Independente da quantidade de variáveis (v), a soma dos pesos (p) será <= 19.90. Ou seja, o peso total será dividido em todas as variáveis utilizadas.

[1]Por exemplo, se o aluno for classificado com $k = 4$ e pertencer a classe 2-Desistente, o valor de $MinSG$ será 201.

Exemplo hipotético de um aluno: considere que apenas as variáveis *QtdeAprovacoes* e *AcessoLMS*, apresentadas na Tabela 8 sejam utilizadas no ajuste do *Score* de um aluno e tenham os valores normalizados 1 e 0,5, pesos 14,90 e 5 (Total=19,90), respectivamente e após aplicada a classificação *KNN*, o aluno seja classificado na classe 1 (Não Desistente), pois possui 4 vizinhos (k) da classe 1. Inicialmente, este aluno seria classificado na classe 1 (Sem Risco de Evasão) e no *Score* com intervalo entre 601 e 800. Com este resultado o valor de *MinSG* será 601 e o cálculo pode ser representado por:

$$ScoreFinal = 601 + ((1 * 14,90 * 10) + (0,5 * 5 * 10)) \quad (2)$$

O valor do *ScoreFinal* será 775. Ou seja, o aluno pertence a classe 1 (Não Desistente) e conforme o resultado da Equação 2, ele está próximo do grupo de alunos com baixo risco de evasão (801 a 1.000). Como o peso de cada variável será entre 0 e 1, na variável *QtdeAprovacoes* o aluno obteve 100% (1) e na variável *AcessoLMS* 50% (0,5). Ou seja, o aluno possui uma taxa alta de aprovações e taxa média de acessos no ambiente virtual (LMS).

O resultado parcial do modelo, apresentado na Figura 3, indica que é possível classificar o aluno e definir o seu grupo de *Score* dentro de um intervalo. Contudo, para refinar sua classificação seriam necessários novos estudos para identificar as variáveis e seus respectivos pesos, e assim, calcular o resultado exato, conforme proposto pela Equação 1.

Os resultados podem ser utilizados como indicativo de uma possível evasão e não como uma afirmativa, pois mesmo que o aluno esteja com dados que podem classificá-lo com risco de evasão, não é possível afirmar que o mesmo será um desistente. Nesta perspectiva, os resultados podem ser utilizados como insumos para auxiliar na análise da evasão.

Em relação a reutilização deste modelo em outros cenários, nota-se que seria necessário ajustar os grupos de acordo com valor de k e a definição das variáveis e seus respectivos pesos. Isto mostra que o modelo é promissor e pode ser adaptado em outros contextos.

5 CONSIDERAÇÕES FINAIS

As pesquisas identificadas na literatura focam essencialmente em indicadores categóricos, ou seja, com resultados binários que denotam que o aluno está ou não no grupo de risco. Apesar desta análise ser importante, não apresenta a variação de performance do aluno durante sua vida acadêmica.

Os resultados preliminares alcançados indicam que a criação de um *Score* é promissora e pode ser uma alternativa aos estudos identificados na literatura, pois além de classificar um aluno de acordo com sua classe, é possível analisar o quão próximo ele está ou não do grupo de risco.

Apesar dos resultados obtidos com o modelo de Score serem parciais, indicam que ao aplicar os ajustes com a utilização da Equação 1, pode-se obter resultados importantes, pois leva em consideração o peso das variáveis. Por exemplo, se um determinado aluno possui uma frequência alta nas disciplinas, esta variável seria importante para aumentar o seu *Score*.

Como o Brasil passou por um período de Pandemia, isto trouxe mudanças no meio de aprendizado, avaliação, renda e fatores psicológicos. Então, estes fatores podem ter influenciado na desistência dos alunos, e como consequência, nos resultados alcançados. Contudo, parte da amostra pertence a um período anterior a Pandemia, e no futuro, novos experimentos serão aplicados, e caso seja necessário, os devidos ajustes serão realizados.

Por fim, como trabalhos futuros estão a definição das variáveis e seus respectivos pesos, para validar e ajustar a Equação 1 e a implantação deste modelo como um dos indicadores da plataforma de *Business Intelligence*, que atualmente é utilizada pela Instituição.

REFERENCES

[1] Delsi Fries Davok and Rosilane Pontes Bernard. Avaliação dos índices de evasão nos cursos de graduação da universidade do estado de santa catarina-udesc. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 21:503–522, 2016.

[2] MEC. Altos índices de desistência na graduação revelam fragilidade do ensino médio, avalia ministro, oct 2016. URL www.portal.mec.gov.br/ultimas-noticias/212-educacao-superior-1690610854/40111-altos-indices-de-evasao-na-graduacao-revelam-fragilidade-do-ensino-medio-avalia-ministro.

[3] Mariele Lanes and Cleber Alcântara. Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. 29(1):1921, 2018.

[4] Lovenoor Aulck, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*, 2016.

[5] Hadautho Roberto Barros da Silva and Paulo Jorge Leitão Adeodato. A data mining approach for preventing undergraduate students retention. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, June 2012. doi: 10.1109/IJCNN.2012.6252437.

[6] Laci Mary Barbosa Manhães, Sérgio Manuel Serra da Cruz, Raimundo J. Macário Costa, Jorge Zavaleta, and Geraldo Zimbrão. Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. *VIII Simpósio Brasileiro de Sistemas de Informação (SBSI2012)*, 2002.

[7] Humberto Rabelo, Aquiles Burlamaqui, Ricardo Valentim, Danieli Silva de Souza Rabelo, and Soraya Medeiros. Utilização de técnicas de mineração de dados educacionais para predição de desempenho de alunos de ead em ambientes virtuais de aprendizagem. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 28, page 1527, 2017.

[8] Bruno Claudino Pereira de Brito, Rafael Ferreira Leite de Mello, and Gabriel Alves. Identificação de atributos relevantes na evasão no ensino superior público brasileiro. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1032–1041. SBC, 2020.

[9] Leonardo de Almeida Teodoro and Marco André Abud Kappel. Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no brasil. *Revista Brasileira de Informática na Educação*, 28:838–863, 2020.

[10] José Júnior e Robinson Noronha e Celso Kaestner. Criação e seleção de atributos aplicados na previsão da evasão de curso em alunos de graduação. *Anais do Computer on the Beach*, 0(0):061–070, 2017. ISSN 2358-0852. doi: 10.14210/cotb.v0n0.p061-070. URL www.siaiap32.univali.br/seer/index.php/acotb/article/view/10725.

[11] Helena Sampaio. Ensino superior no brasil: o setor privado. *Cadernos de Pesquisa*, pages 213–213, 2000.

[12] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991. ISSN 1573-0565. doi: 10.1007/BF00153759. URL www.dx.doi.org/10.1007/BF00153759.

[13] Tom Howley, Michael G Madden, Marie-Louise O'Connell, and Alan G Ryder. The effect of principal component analysis on machine learning accuracy with high dimensional spectral data. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 209–222. Springer, 2005.

[14] D Ramyachitra and P Manikandan. Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, 5(4):1–29, 2014.

[15] Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126. ICML United States, 2003.