Mineração de Dados Educacionais na Predição do Desempenho Acadêmico: um prognóstico a partir do percurso curricular realizado

Léo Manoel Lopes da Silva Garcia[1], Daiany Francisca Lara[1], Raquel Salcedo Gomes[2], Sílvio Cezar Cazella[2]

[1] Faculdade de Ciências Exatas – Curso de Ciências da Computação – Universidade do Estado de Mato Grosso(UNEMAT) – Barra do Bugres – MT – Brasil

[2] Programa de Pós-Graduação em Informática na Educação – Universidade Federal do Rio Grande do Sul (UFRGS) – Porto Alegre – RS – Brasil

leoneto@unemat.br, dflara@unemat.br, raquel.salcedo@ufrgs.br, silvio.cazella@gmail.com

Abstract. *This work presents the evaluation of predictive models for the identification of students at risk of failing in specific subjects. For this purpose, the curricular path previously carried out by the student before taking a certain course is used as a predictor attribute. The impact of using load balancing techniques on predictive model evaluation metrics is investigated. The results highlighted the best performances for the Random Forest, J48 and IBK algorithms, presenting an Accuracy from 71% to 81% and Recall from 75% to 93%, reflecting a significant improvement when using the SMOTE oversampling technique for balancing charge.*

Resumo. *Este trabalho apresenta a avaliação de modelos preditivos para a identificação de alunos com risco de reprovação em disciplinas específicas. Para tanto, é utilizado como atributo preditor o percurso curricular realizado previamente pelo aluno antes de cursar uma determinada disciplina. O impacto da utilização de técnicas de balanceamento de carga nas métricas de avaliação dos modelos preditivos é investigado. Os resultados destacaram os melhores desempenhos para os algoritmos Random Forest, J48 e IBK, apresentando uma Acurácia de 71% a 81% e Recall de 75% a 93%, refletindo uma significativa melhoria ao se utilizar a técnica de sobreamostragem SMOTE para o balanceamento de carga.*

1. Introdução

Após um longo período de expansão de vagas e matrículas na educação superior, as instituições de ensino superior (IES) têm sido pressionadas a apresentar resultados, principalmente no que concerne à capacidade de produzir concluintes no período previsto. Estes resultados provêm de trajetórias acadêmicas realizadas pelos alunos, nas quais o desempenho acadêmico emerge como um fator determinante para resultados positivos (diplomação) ou negativos (retenção e evasão) [GARCIA *et al*., 2020]. Dessa maneira, recaem sobre os gestores e professores a incumbência de formular estratégias educacionais que favoreçam o aprendizado e reduzam as adversidades que levam ao baixo desempenho. Neste contexto, tem se explorado o uso de tecnologias da informação na análise da grande quantidade de dados educacionais que têm se tornado cada vez mais disponíveis, sob o entendimento de que estes dados são valiosos para ações que envolvem o processo de

DOI: 10.5753/sbie.2022.225221

ensino-aprendizagem, amparando a tomada de decisão da gestão educacional [ROMERO e VENTURA, 2020].

Dentre os recursos tecnológicos disponíveis, destaca-se o uso de Mineração de Dados Educacionais (MDE), que compreende não só uma técnica, mas um campo de pesquisa voltado à aplicação de mineração de dados cujas origens são inerentes ao contexto educacional, com a intenção principal de produzir um entendimento confiável sobre os padrões de aprendizagem dos alunos e identificar seus processos e comportamentos de estudo para potencializar os resultados educacionais [ANOOPKUMAR, 2018]. Na mineração de dados educacionais, as abordagens podem ser divididas em duas categorias principais: os métodos descritivos e os preditivos. Os primeiros permitem identificar padrões nos dados, reconhecendo possíveis regras de causa e efeito. Os segundos, métodos preditivos, têm como objetivo fazer inferências sobre o futuro, possibilitando previsões de ocorrências por meio de indução baseada em ocorrências anteriores [AKMEŞE, 2021]. Tanto a identificação de padrões quanto as predições podem subsidiar ações para precaver-se de possíveis dificuldades no processo de educação. De posse dessas informações, instrutores podem monitorar o progresso dos alunos e intervir precocemente nos problemas acadêmicos, negociando rotas alternativas e discernindo estratégias para suprir carências ou fortalecer vulnerabilidades [AKMEŞE, 2021].

Explorando os potenciais benefícios que a MDE pode proporcionar às práticas educacionais, este estudo tem o objetivo de avaliar a utilização de modelos preditivos para a identificação de alunos com risco de reprovação em disciplinas específicas. A partir da matrícula do aluno em uma disciplina em uma determinada etapa de sua trajetória acadêmica, almeja-se obter um prognóstico acerca da probabilidade de sua aprovação ou reprovação. Para tanto, são utilizados como atributos previsores o percurso curricular realizado previamente pelo aluno antes de cursar uma determinada disciplina. Com esse empreendimento, busca-se trazer uma contribuição para a elaboração de intervenções para mitigação do problema do baixo desempenho acadêmico, considerando seu impacto direto nos resultados das trajetórias acadêmicas, podendo resultar em evasão ou retenção. Embora a retenção seja caracterizada pelo atraso na conclusão do curso, ela ocorre devido às reprovações ocorridas nas disciplinas (exceto em casos de trancamento). E essas reprovações também podem figurar como fator motivador da evasão por seu impacto negativo no senso de auto-eficácia do aluno, o que é destacado por Tinto (2017) como relevante em estudos acerca da trajetória acadêmica no ensino superior.

Da mesma forma, almeja-se contribuir para o arcabouço teórico da MDE ao utilizar uma abordagem da previsão a partir dos percursos realizados, que caracterizam atributos previsores dinâmicos e continuadamente recentes. A cada período letivo esse percurso é atualizado desenhando novas características às probabilidades de desempenho de cada aluno, ao invés da utilização de atributos estáticos de pré-ingresso como dados demográfios e socioeconômicos, por exemplo, que usualmente são utilizados em aplicações de MDE com esta finalidade.

2. Trabalhos Relacionados

São inúmeros os estudos que implementam técnicas de MDE em abordagens preditivas, os quais vão desde a sua utilização para apoiar admissão de alunos [MENGASH, 2020], até a previsão do potencial de empregabilidade de egressos [PABREJA, 2017]. Todavia, são predominantes os estudos para previsão dos resultados da trajetória acadêmica (diplomação, evasão e retenção) e do desempenho.

Manhães e Cruz (2019) propõem uma arquitetura para prever o desempenho acadêmico dos estudantes de graduação e identificar aqueles que estão em risco de evadir do sistema de ensino. Os autores utilizam exclusivamente dados acadêmicos que incluem o coeficiente de rendimento, quantidade de matrículas e aprovações, estados de progressão, dentre outros. Em diferentes configurações do modelo de predição, são destacadas as previsões do desempenho em aprovação ou reprovação nos cinco primeiros semestres letivos para os cursos de graduação de base matemática; situação de conclusão ou não dos cursos; e o progresso a ser obtido pelos alunos, considerando como progresso positivo quando seu desempenho em um determinado semestre do curso está acima de um padrão mínimo.

Souza e Cazella (2022), utilizando algoritmos de Regressão, realizaram a previsão do desempenho de alunos em um conjunto de dados públicos divididos em duas bases de dados menores, uma referente à disciplina de Português e outra com dados relativos à disciplina de Matemática. Os atributos destas bases apresentam 4 segmentos: atributos pessoais, comportamentais, antecedentes acadêmicos e antecedentes econômicos; totalizando 33 atributos. Os autores buscam avaliar a precisão dos algoritmos utilizados e indicar quais os principais atributos preditores para o desempenho dos alunos. Todos os algoritmos utilizados apresentaram a Acurácia em torno de 80%, com destaque para os melhores desempenhos das Árvores de decisão[1] e Random Forest[2] em ambas as disciplinas. Além disso, constatou-se que atributos relacionados às atividades escolares são mais preditores para o desempenho dos alunos.

Alturki (2021) utiliza a MDE para conceber modelos para a predição dos conceitos finais dos alunos (categorizadas nos conceitos Excelente, Muito Bom, Bom, Aceitável e Insuficiente) e identificar os alunos ‘honorários’ (com desempenho promissor) em um estágio inicial. É utilizado como atributo pré-matrícula o percentual de desempenho no ensino médio, e como atributos pós matrículas são utilizados a média cumulativa do aluno dos 4 primeiros semestres letivos, carga horária, número de reprovações e notas em disciplinas básicas do curso. O classificador Naive Bayes[3] teve um desempenho melhor do que os modelos baseados em árvore na previsão do desempenho acadêmico dos alunos em geral. No entanto, o Random Forest superou o Naive Bayes na previsão de alunos honorários. Destacou-se que as principais características que podem prever o desempenho acadêmico dos alunos são a média cumulativa do aluno do aluno durante os primeiros quatro semestres, o número de cursos reprovados durante os primeiros quatro semestres e as notas dos três principais cursos.

Miguéis et al. (2018) combinam técnicas de agrupamento e de classificação para a previsão do desempenho acadêmico. Inicialmente utilizando dados de desempenho do primeiro ano do curso, os alunos são agrupados em 5 classes de desempenho (A, B, C, D e E), posteriormente são agregados a esta informação atributos sociodemográficos, socioeconômicos, do ensino médio, notas de admissão, coeficiente de rendimento e número total de créditos matriculados neste período. O melhor resultado foi obtido com o algoritmo Random Forest, imprimindo uma precisão acima de 95%. Destacou-se a média dos exames do ensino médio e média das notas admissionais como atributos mais relevantes na previsão.

[1] Algoritmo baseado em entropia, onde cada atributo no conjunto de dados é tratado como um nó na árvore de decisão.
[2] Conhecida como ensemble learning essa abordagem utiliza várias Árvores de Decisão.
[3] Algoritmo baseado no teorema de Bayes e fundamentado no princípio de independência de recurso.

Todos os trabalhos aqui relacionados norteiam e fundamentam as metodologias do presente estudo, que busca trazer abordagens diferenciadas e contribuir para este campo. Assim como em Manhães e Cruz (2019), este estudo fundamenta-se em informações pós-ingresso, contemplando o percurso curricular dos alunos. Esta abordagem vai ao encontro da concepção de Alturski (2021), que descreve que o uso de informações pós-ingresso para prever o desempenho acadêmico dos alunos pode maximizar a precisão da predição, pois esses recursos representam a situação atual dos alunos no curso. Ademais, objetiva-se avaliar esta proposição sob a perspectiva de que dados pré-ingresso, isoladamente, não contemplam a evolução dos alunos durante sua trajetória acadêmica. Mesmo um rico histórico de perfil não pode determinar que o aluno que cursa a disciplina de Cálculo I no segundo semestre de um curso de computação vai perdurar com as mesmas características de desempenho quando a mesma disciplina for cursada no 5º semestre, por exemplo. Utilizar as informações acerca do percurso realizado busca considerar a evolução da experiência do aluno durante sua permanência no curso para a previsão do seu desempenho.

Os trabalhos de [SOUZA e CAZELLA, 2022; ALTURKI, 2021; MIGUÉIS *et al*., 2018] utilizam dados pós-ingresso como o coeficiente de rendimento, média cumulativa e carga horária matriculada ou cursada. Porém, estas informações podem não retratar satisfatoriamente as experiências dos alunos durante sua trajetória, pois, o coeficiente do rendimento ou média cumulativa não detalham sobre quais componentes curriculares essas notas foram obtidas. Da mesma forma, a carga horária não detalha esta informação. Já o percurso curricular, por sua vez, contempla as interações dos alunos com cada componente curricular e qual foi o resultado obtido, atualizando esta informação a cada período letivo do aluno, retratando a situação atual de sua vivência acadêmica. Além disso, essa abordagem atende a flexibilidade do currículo no ensino superior, uma vez que o aluno pode matricular-se em uma disciplina em qualquer etapa de sua trajetória, desde que se respeite os pré-requisitos. Assim, a previsão de desempenho em uma disciplina considera o histórico do aluno até o exato momento em que ele for cursá-la.

3. Metodologia

Este estudo compõe uma pesquisa cujo objetivo é desenvolver um ambiente no qual é possível aos professores de quaisquer disciplinas verificar, em sua lista de matriculados, no início de um período letivo, a probabilidade de cada aluno em ser aprovado ou reprovado em sua disciplina e, dessa maneira, planejar intervenções pedagógicas que amparem alunos com previsão de reprovação, para que assim eles obtenham a aprovação.

Desse modo, este estudo configura uma etapa inicial de validação de modelos preditivos e tem como principal objetivo avaliar a utilização de modelos preditivos na previsão de desempenho acadêmico em disciplinas a partir do percurso curricular realizado previamente pelo aluno. Este percurso é concebido neste estudo como a interação dos alunos com os componentes curriculares, e nos resultados destas interações. Assim, quando o aluno se matricula em uma disciplina X na sua 3º fase letiva em um curso, por exemplo, seu percurso curricular é retratado por todas as disciplinas cursadas em sua 1ª e 2ª fase letiva (período anterior à matrícula na disciplina) e nos respectivos resultados, definidos em Reprovação ou Aprovação. Desse modo, deseja-se obter, por meio de um modelo preditivo, a previsão deste aluno ser aprovado ou reprovado na disciplina X. Obviamente, nas fases letivas posteriores esse percurso é atualizado.

3.1. Obtenção dos Dados

Este estudo foi realizado a partir de dados reais da Universidade do Estado de Mato Grosso[4]. Como a abordagem é baseada no currículo de cada curso, não é possível gerar um Dataset para toda instituição e sim um para cada curso. Assim, para este experimento, foram contemplados os dados dos cursos de Computação e de Matemática do campus de Barra do Brugres. Para ambos os cursos, foram contemplados alunos de turmas ingressantes de 2013/2 a 2017/1 (8 turmas) acompanhando suas interações com o currículo do curso até o período letivo de 2019/2 (último período pré-pandemia COVID19). Para o curso de Computação, foram contemplados 297 alunos que geraram o total de 7.677 matrículas em disciplinas neste período. Para o curso de Matemática foram contemplados 247 alunos que geraram 5.189 matrículas em disciplinas. A Tabela 1 retrata a caracterização dos dados extraídos. Observa-se a alta taxa de reprovação desses cursos. Ao considerar-se os dois cursos juntos, é apresentado um total de 52,60% de reprovações, o que é refletido em outros resultados negativos, como baixa diplomação (4,7%) e alta evasão (72,06%).

Tabela 1. Caracterização da Amostra de Dados

	Sexo		Resultado		Trajetória Acadêmica		
Curso	M	F	Reprovação	Aprovação	Diplomados	Matriculados	Evadidos
Computação	242	55	4.296	3.381	12	66	219
Matemática	101	146	2.472	2.717	14	60	173

A extração dos dados para geração dos Datasets foi realizada no sistema acadêmico da instituição, especificamente a partir do histórico escolar dos alunos contemplados no estudo e dos componentes curriculares cadastrados para cada curso. Os dados foram centralizados em um banco de dados MySql separado do sistema acadêmico.

3.2. Tratamento dos Dados

A partir dos históricos extraídos do sistema acadêmico, foi necessário implementar uma aplicação para gerar os percursos curriculares realizados em um Dataset para treinamento dos modelos preditivos. Implementada na linguagem de programação Java e ambiente de desenvolvimento Eclipse, essa aplicação inicia percorrendo cada componente curricular de cada curso. Em cada componente são recuperadas todas as matrículas realizadas no componente dentro do período analisado. Em cada matrícula é extraída a fase letiva em que o aluno está cursando o componente, o resultado obtido (esta será a classe de predição). Também é recuperado o sexo e data de nascimento do aluno. A partir das datas de nascimento, foram calculadas as idades dos alunos e discretizadas em faixas etárias baseadas da segmentação realizada pelo Instituto Brasileiro de Geografia e Estatística. Por fim, são extraídas todas as disciplinas cursadas nas fases letivas anteriores à fase da referida matrícula. A Figura 1 ilustra o vetor de características gerado para cada instância de matrícula. Observa-se que todos os componentes curriculares compõem o vetor de características, independentemente se ele foi cursado ou não pelo aluno. Quando o aluno não cursou um componente curricular até a data analisada é atribuído o parâmetro "?", caso contrário, quando ele já cursou o componente, pode ser atribuído o parâmetro "APROVADO" ou "REPROVADO".

[4] Pesquisa aprovada no comitê de ética na Plataforma Brasil, sob número CAAE 50431321.0.0000.5347

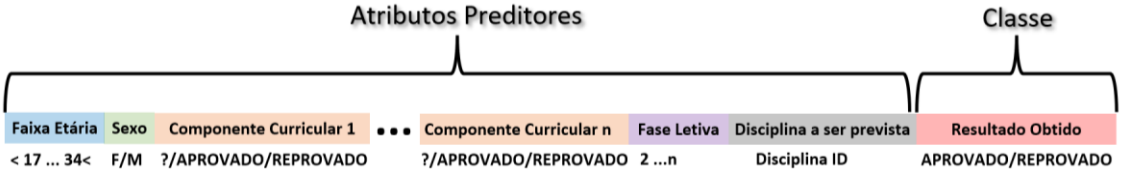

Figura 1. Vetor de Características para Predição

Cabe destacar que é considerada a última vez que o aluno cursou a disciplina. Isto é, tome-se, por exemplo, que se esteja analisando a matrícula de um aluno em sua 5ª fase letiva, ele pode ter cursado uma disciplina Xn em sua 2ª fase letiva (obtendo reprovação) e a ter cursado novamente em sua 4ª fase letiva (agora obtendo aprovação). Será registrado em seu vetor de característica a última vez que ele cursou esta disciplina, ou seja, será indicado que ele já obteve a aprovação neste componente curricular.

O atributo "Disciplina a ser prevista" corresponde ao componente curricular sobre o qual se pretende fazer a predição. Ou seja, quando um aluno X se matrícula em uma disciplina Y (Disciplina a ser prevista) é pretendido prever se o "Resultado Obtido" (classe) será 'Aprovação'' ou 'Reprovação. O atributo "Fase letiva" corresponde ao momento da trajetória acadêmica do aluno em que ele irá cursar o componente curricular (Disciplina a ser prevista), no qual pretende-se prever se ele será aprovado ou não. Este "momento" é baseado no período letivo de ingresso do aluno. Ou seja, quando um aluno ingressa em um curso, ele inicia em sua 1ª fase letiva, ao concluir um período letivo esta fase é incrementada e ele passa para a 2ª fase letiva de sua trajetória, e assim sucessivamente. O Resultado Obtido é descrito como "APROVAÇÃO" ou "REPROVAÇÃO", e esta é a classe previsora para a qual é esperado que uma instância de matrícula seja classificada.

É importante ressaltar que só foram contempladas a predição para as disciplinas previstas para serem cursadas a partir da 2ª fase letiva, de acordo com o currículo do curso. Pois, ao se matricular no curso ele já é matriculado nas disciplinas previstas para a 1ª fase. E, também, porque ainda não há percurso gerado quando o aluno inicia seu primeiro período letivo. Para testar diferentes configurações da base de treinamento, além do Dataset com todos os atributos descritos na Figura 1, foram gerados outros três conjuntos de treino: Sem_Faixa_Etaria (excluindo a faixa etária), Sem_Sexo (excluindo o atributo sexo) e Sem_Sexo_Faixa_Etaria (excluindo os atributos sexo e faixa etária).

3.3. Aplicação dos Algoritmos de Mineração de Dados

Os algoritmos foram aplicados por meio da biblioteca Java do Weka 3.8.5. Foi implementada uma aplicação em Java para aplicar todos os algoritmos selecionados nas 4 bases de treino. Os algoritmos de classificação utilizados foram o Naive Bayes, IBK[5], JRIP[6], J48 (Árvore de Decisão), Random Forest e MultiLayer Perceptron[7]. Estes algoritmos foram escolhidos pois os mesmos são amplamente utilizados em artigos que relatam pesquisas focadas em MDE. Os algoritmos foram aplicados às bases em sua configuração natural. Posteriormente, estas bases foram submetidas a métodos de balanceamento de carga e aplicados todos os algoritmos novamente. Foram realizados

[5] Algoritmo supervisionado não paramétrico, é baseado na distância euclidiana entre as instâncias.
[6] Técnica baseada no uso algoritmos de cobertura sequencial para criar listas de regras ordenadas.
[7] Implementa camadas de neurônios treinados com o algoritmo de aprendizado de retropropagação.

testes com os métodos Class Balancer e SMOTE. Utilizou-se o particionamento das bases de dados, por meio de validação cruzada com 10 conjuntos (10-fold *cross validation*).

3.3.1. Resultados obtidos

Para avaliação dos algoritmos, foram utilizados como parâmetros duas métricas principais: a Acurácia e o Recall, este para reprovados. A Acurácia descreve a proporção de amostras estimadas corretamente em relação ao número de todas as amostras. Ou seja, o teste é a taxa de diagnósticos corretos totais. É uma das métricas mais utilizadas no contexto de avaliação de modelos de aprendizagem supervisionada e classificação. Já o Recall corresponde à taxa de acerto dos exemplos positivos classificados corretamente entre todos os exemplos positivos da base. É uma métrica por meio da qual é possível avaliar isoladamente o desempenho da predição específica para cada classe.

Para conjectura deste estudo, a métrica Recall tem uma significativa relevância, pois, embora a Acurácia seja a métrica mais utilizada em estudos de mineração de dados, ela deve ser utilizada e interpretada com atenção, caso contrário pode induzir a um equívoco na avaliação do modelo, principalmente em casos de classes desbalanceadas. Uma alta taxa de Acurácia pode estar sendo baseada nos acertos das classes majoritárias, negligenciando classes minoritárias. Neste estudo, é preferível que os modelos de previsão tenham um excelente desempenho ao identificar alunos em risco de reprovação, visto que os benefícios de realizar uma previsão destes resultados residem na possibilidade de realizar uma intervenção precoce em alunos com risco de resultados negativos, de forma a alterar suas trajetórias, transformando possíveis tendências negativas em resultados positivos [GARCIA *et al.*, 2022].

Neste documento, são apresentados resultados principais com os melhores desempenhos obtidos. Nesta perspectiva, destaca-se que os testes com as bases de treinamento balanceadas com o método Class Balancer não apresentaram melhorias e tiveram resultados inferiores às demais bases. Sendo assim, seus resultados não serão retratados. Quanto às 4 bases de treinamento geradas na seção 3.2, os melhores resultados foram obtidos com a base sem o sexo e sem a faixa etária, contendo somente o percurso curricular. A Figura 2 apresenta os resultados das Acurácias e Recall (da classe REPROVADO) obtidas para o curso de Matemática para a base, com somente o percurso curricular (sem os atributos sexo e faixa etária) em sua configuração original e após o balanceamento com o método SMOTE. A Figura 3 apresenta as mesmas informações contemplando o curso de Computação.

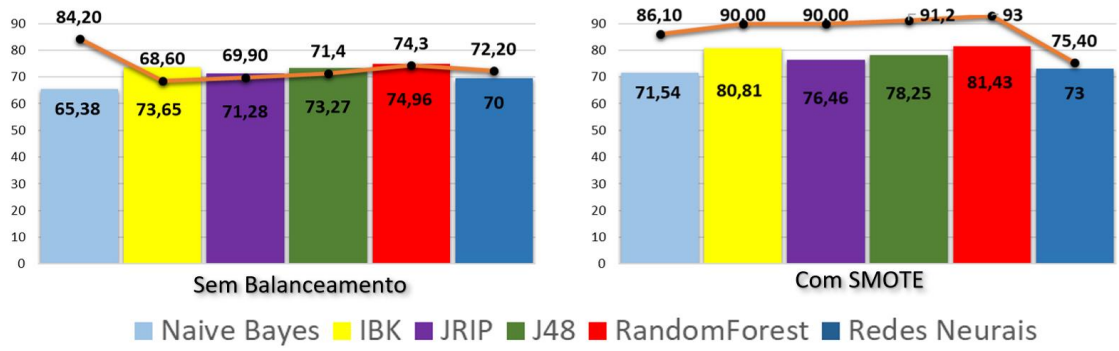

Figura 2. Acurácia e Recall para o Curso de Matemática

As barras retratam a Acurácia, e a linha laranja indica a taxa de Recall para a classe "REPROVADO" para cada modelo preditivo. Observa-se que os melhores resultados

foram obtidos com o algoritmo Random Forest, em ambos os cursos, tanto na base original quanto na base balanceada com o método SMOTE. Este algoritmo também apresentou os melhores resultados nos estudos de [SOUZA e CAZELLA, 2022; ALTURKI, 2021; MIGUÉIS *et al.*, 2018]. Destaca-se aqui, que mesmo com resultados inferiores, os algoritmos IBK e J48 apresentaram taxas aproximadas do Random Forest. Como já destacado anteriormente, para os objetivos desse trabalho, foi importante obter uma boa taxa de acerto principalmente para a previsão de alunos em risco de reprovação. Para o curso de matemática, o que se observa é que os algoritmos que apresentaram o melhor Recall para classe REPROVADOS são os mesmos com a melhor Acurácia, com exceção dos resultados obtidos com o Naive Bayes para a base sem balanceamento. Observa-se ainda uma considerável melhoria nos resultados após a aplicação do método SMOTE, chegando a apresentar um incremento de até 7% na Acurácia, e 21% no Recall.

Já para o curso de computação, os algoritmos Naive Bayes e Multilayer Perceptron (Redes Neurais) apresentam os melhores resultados de Recall, porém com as menores taxas de Acurácia. E, embora, ocorra a melhoria da Acurácia após o a aplicação do método SMOTE, há uma redução significativa do Recall, o que pode inviabilizar o uso do método de balanceamento para este curso.

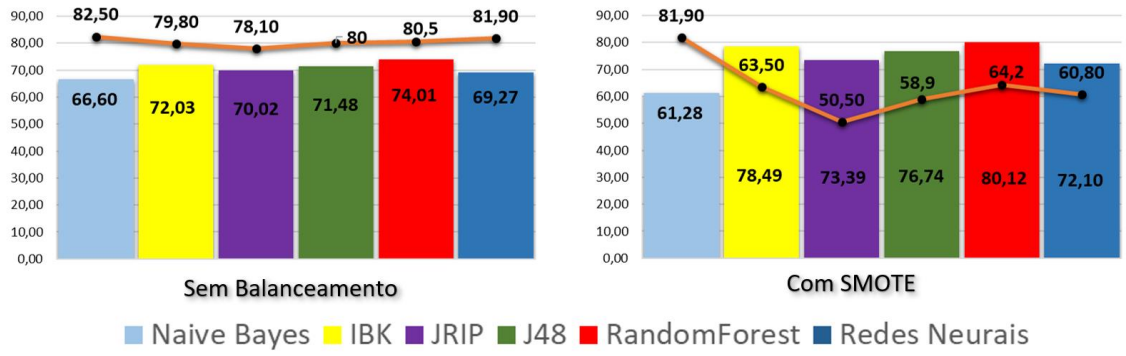

Figura 3. Acurácia e Recall para o Curso de Computação

As bases de treinamento descritas e testadas até o momento contemplaram todos os dados de cada curso. Porém, realizou-se também testes em subconjuntos destas bases de dados em busca de melhores resultados. Dessa forma, contendo os mesmos atributos previsores descritos na Figura 1, foram construídas bases por período letivo, destacando somente matrículas em disciplinas disponíveis para cada fase letiva, de acordo com o currículo do curso. Nos testes destes conjuntos, o algoritmo Random Forest também se destacou com os melhores resultados após aplicação do método SMOTE para o curso de Matemática e sem balanceamento para o curso de Computação. A Figura 4 demonstra a Acurácia e o Recall obtidos com este algoritmo para os cursos de Matemática (balanceado com SMOTE) e Computação (sem balanceamento).

São trazidos para discussão somente os resultados obtidos do 2º ao 5º período letivo, pois, devido à amostragem, os subconjuntos após esse período possuem poucas instâncias, o que prejudicou a classificação. É possível verificar uma melhoria superficial na Acurácia e Recall dos classificadores em relação aos dados de todo o curso. Nota-se que as taxas tendem a cair a partir da 5ª fase. Todavia, isso pode ser justificado devido às características da amostra em que, no recorte realizado, os períodos mais avançados apresentam uma menor quantidade de instâncias. Isso pode prejudicar a capacidade de treinamento do classificador.

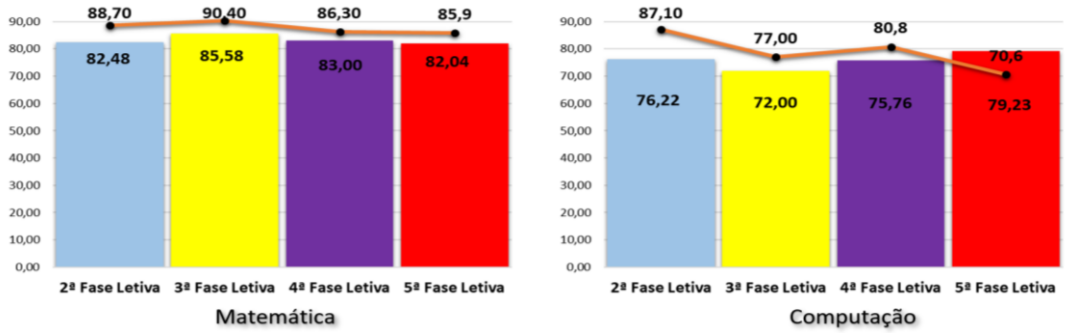

Figura 4. Acurácia e Recall por Período Letivo para Matemática e Computação

Neste documento, foram selecionadas para a apresentação as informações consideradas mais relevantes. Entretanto, os resultados completos dos testes contemplando todos os algoritmos, método Class Balancer e as diferentes configurações das bases de treinamento podem ser obtidos pelo link: resultados_testes**.** Estão contidos ali, inclusive, testes realizados em bases com dados exclusivamente de disciplinas específicas, que, pela quantidade baixa de instâncias ao se realizar essa segmentação, apresentaram taxas inferiores para Acurácia e Recall. Considera-se que, para avaliar o treinamento de modelos específicos para cada disciplina é necessário um recorte temporal maior para a coleta, o que não foi possível no momento.

4. Considerações Finais

Os resultados alcançados pelos testes realizados demonstraram a viabilidade de se utilizar o percurso curricular como atributos previsores, visto que as taxas alcançadas se aproximam dos resultados dos estudos relacionados. Ainda que a metodologia não empregue dados pré-ingresso, não é descartado o seu valor e nem a possibilidade de agregá-los às informações dos percursos curriculares realizados. Até porque, figura-se como limitação deste estudo a impossibilidade a previsão de desempenho na primeira fase letiva. O que se almejou, aqui, foi avaliar uma metodologia alternativa para MDE e contribuir com esse campo de estudo, considerando a trajetória realizada. Esta abordagem busca captar a experiência do aluno ao longo de sua trajetória, como também pode refletir os resultados das intervenções realizadas.

Os resultados distintos dos algoritmos e métodos para os cursos de Matemática e Computação reforçam o entendimento de que cada curso possui suas características e dinâmicas próprias, que são refletidas no perfil e comportamento de seus alunos. Dessa maneira, eles reagem de forma diferente aos percursos curriculares, resultando em bases de treinamento com características distintas, de forma que se alterem os algoritmos que possam apresentar os melhores resultados. Assim, para todos os cursos da instituição devem ser repetidos todos esses processos, para se criar um modelo de predição específico e adequado a cada curso.

Ressalta-se, ainda, para trabalhos futuros, a necessidade de explorar estes dados sob uma perspectiva descritiva da MDE, explorando padrões que podem ser destacados pelos algoritmos J48 ou JRIP, por exemplo. Dessa maneira, podem ser identificados insights relevantes, tão úteis aos gestores educacionais quanto informações preditivas. Ainda tratando-se de trabalhos futuros, é identificada a necessidade de realizar outros testes com um maior detalhamento do percurso realizado, pois, neste estudo se tratou apenas dos resultados de aprovação ou reprovação. É possível, em testes futuros, detalhar estes resultados de acordo com a normatização acadêmica como em reprovação por faltas,

reprovação direto por nota, reprovação na prova final, aprovação direta e aprovação na prova final, por exemplo. Testes realizando a predição do desempenho utilizando os modelos preditivos treinados sobre dados de alunos após o período contemplado na base de treinamento, podem ratificar a eficácia dos modelos.

Referências

Akmeşe, Ö. F., Kör, H., and Erbay, H. (2021). "Use Of Machine Learning Techniques For The Forecast Of Student Achievement In Higher Education". Information Technologies and Learning Tools, 82(2), https://doi.org/10.33407/itlt.v82i2.4178

Alturki, S., e Alturki, N. (2021). "Using educational data mining to predict students' academic performance for applying early interventions". Journal of Information Technology Education: Innovations in Practice, 20. https://doi.org/10.28945 /4835

Anoopkumar, M. and Zubair Rahman, A. M. J. Md. (2018). "Bound Model of Clustering and Classification (BMCC) for Proficient Performance Prediction of Didactical Outcomes of Students". International Journal of Advanced Computer Science and Applications 9(11), http://dx.doi.org/10.14569/IJACSA.2018.091133

Garcia, L. M. L. S., Lara, D. F., e Antunes, F. (2020). "Análise da Retenção no Ensino Superior: um Estudo de Caso em um Curso de Sistemas de Informação". Revista da Faculdade de Educação 34:15-38. https://doi.org/10.30681/21787476.2020.34.1538.

Garcia, L. M. L. S.; Lara, D. F.; Gomes, R. S.; e Cazella, S. C. (2022). "The Discovery of Knowledge in Educational Databases: A Literature Review with Emphasis on Preprocessing and Postprocessing". The Turkish Online Journal of Educational Technology (TOJET), v. 21, p. 75-87, 2022.

Manhães, L. M. B., Cruz, S. M. S. (2019) "Predição do Desempenho Acadêmico de Alunos da Graduação Utilizando Mineração de Dados". In: XIX Simpósio de Pesquisa Operacional e Logística Marinha. Rio de Janeiro-RJ. Novembro de 2019.

Mengash, H. A. (2020)."Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," in IEEE Access, vol. 8, pp. 55462-55470, doi: 10.1109/ACCESS.2020.2981905.

Miguéis, V. L., Freitas, Ana., Garcia, Paulo J.V. Silva, André. (2018). "Early segmentation of students according to their academic performance: A predictive modelling approach". Decision Support Systems, Volume 115, Pages 36-51, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2018.09.001.

Pabreja, K. (2017). "Comparison of Different Classification Techniques for Educational Data." IJISSS vol.9, no.1: pp.54-67. http://doi.org/10.4018/IJISSS.2017010104

Romero, C., Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIRES Data Mining and Knowledge Discovery, 10(3), 3. https://doi.org/10.1002/ widm.1355 doi:10.1002/widm.1355.

Souza, V. F., e Cazella, S. C. (2022). "Mineração De Dados Educacionais Com Algoritmos De regressão: Um Estudo Sobre a predição Do Desempenho". Revista Educar Mais 6 :183-98. https://doi.org/10.15536/reducarmais.6.2022.2691.

Tinto, V. (2017). Through the eyes of students. Journal of College Student Retention: Research, Theory & Practice, 19(3), 254–269.