Uso de Aprendizado de Máquina para detecção de risco de evasão no curso de Licenciatura em Computação

Felipe Neves[1] , Fernanda Campos[2] , Mário Dantas[3], José Maria David[4], Regina Braga[5], Victor Ströele[6]

Resumo

A evasão escolar é um verdadeiro desafio para os especialistas em educação. Os cursos de educação a distância lidam com o desengajamento e abandono dos alunos, o que resulta em impactos sociais e econômicos. Fatores comportamentais, cognitivos e demográficos podem estar associados à evasão escolar precoce. Este artigo propõe uma arquitetura baseada em um comitê de classificadores, capaz de prever o desengajamento dos alunos durante o curso. São feitas notificações aos professores e tutores, permitindo-lhes intervir de forma eficaz e tornar o sucesso dos alunos possível. Para avaliar a proposta, um estudo de caso foi conduzido no curso de Licenciatura em computação, modalidade a distância da UFJF e os resultados indicam a viabilidade da solução e do uso de suas tecnologias. Os resultados apontaram um aumento significativo de acerto na identificação de alunos com risco de evasão, chegando a 93% de precisão.
Keywords: Licenciatura em Computação, aprendizado de máquina, modelo preditivo, comitê de classificadores, evasão escolar, educação à distância

Abstract

Dropping out of school is a real challenge for education specialists. Distance education courses deal with student disengagement and abandonment, which results in social and economic impacts. Behavioral, cognitive, and demographic factors may be associated with early school dropout. This article proposes an architecture based on an ensemble model capable of predicting student disengagement during the course. Notifications are made to teachers and tutors, allowing them to intervene effectively and make student success possible. To evaluate the proposal, a case study was conducted in the Degree in Computing course, distance mode at UFJF, and the results indicate the feasibility of the solution and the use of its technologies. The results showed a significant increase in the accuracy of identifying students at risk of dropping out, reaching 93% accuracy.
Keywords: Degree in Computing, machine learning, predictive modeling, ensemble model, school dropout, distance education

1 Programa de Pós-graduação em Ciência da Computação Universidade Federal de Juiz de Fora felipe.neves.braz@ice.ufjf.br
2 Departamento de Ciência da Computação Programa de Pós-graduação em Ciência da Computação Universidade Federal de Juiz de Fora fernanda.campos@ufjf.edu.br
3 Departamento de Ciência da Computação Programa de Pós-graduação em Ciência da Computação Universidade Federal de Juiz de Fora mario.dantas@ice.ufjf.br
4 Departamento de Ciência da Computação Programa de Pós-graduação em Ciência da Computação Universidade Federal de Juiz de Fora jmaria.david@ice.ufjf.br
5 Departamento de Ciência da Computação Programa de Pós-graduação em Ciência da Computação Universidade Federal de Juiz de Fora regina.braga@ufjf.edu.br
6 Departamento de Ciência da Computação Programa de Pós-graduação em Ciência da Computação Universidade Federal de Juiz de Fora regina.braga@ufjf.edu.br

1. Introdução

A educação mudou de um modelo de transferência de conhecimento para um modelo autodirigido ativo e colaborativo pela influência disruptiva da tecnologia (Bagheri e Movahed, 2016). A aprendizagem e as tecnologias de mídias sociais influenciaram muitos aspectos da educação, do papel do professor ao envolvimento do aluno, da inovação à avaliação do aluno, da interação personalizada e única às questões de segurança e privacidade (Neves et al. 2021).

As características comportamentais dos alunos durante o processo educacional é um recurso importante para prever seu desempenho. No contexto da educação assistida, o maior desafio não é apenas enviar recomendações e conteúdos acadêmicos aos alunos, mas prever problemas de aprendizagem e enviar notificações de alerta para professores, administradores, alunos e familiares.

Os sistemas de e-learning permitem que professores e alunos interajam de forma virtual e fornecem um número exponencialmente crescente de serviços educacionais. Quando consideramos a educação online massiva, como os populares Massive Open Online Courses (MOOCs), o desempenho dos alunos, bem como as faltas e o abandono podem ser especialmente problemáticos (Daradoumis et al., 2013).

A evasão escolar resulta de um longo processo de desligamento da escola e das aulas, e tem profundas consequências sociais e econômicas para os alunos, suas famílias e a comunidade em geral (Márquez-Vera et al., 2016). Fatores comportamentais, cognitivos e demográficos podem estar associados à evasão escolar precoce.

A evasão escolar é uma preocupação genuína nas instituições privadas e públicas do ensino superior devido ao seu impacto negativo no bem-estar dos alunos e da comunidade em geral. Ser capaz de prever esse comportamento precocemente pode melhorar o desempenho dos alunos, bem como minimizar suas faltas e desinteresse (Neves et al. 2021).

A Mineração de Dados Educacional (MDE) e o aprendizado de máquina, do inglês Machine Learning (ML), usam métodos e técnicas diferentes, como classificação, regressão, clustering e mineração de relacionamento (Kumari, Jain e Pamula, 2018), capazes de prever fatores que influenciam o índice de evasão dos alunos. Instituições de ensino superior, principalmente as aulas online, podem obter a vantagem da previsão antecipada do desempenho do aluno (Diego et al., 2020; Young e Sunbok, 2019).

Os comitês de classificadores de Aprendizado de Máquina usam mais de um modelo de forma combinada e podem melhorar a precisão e o desempenho da previsão. Eles fornecem precisão de classificação agregando os resultados de vários classificadores. Esses métodos constroem um conjunto de classificadores básicos a partir dos dados de treinamento e realizam a classificação ao votar nas previsões feitas por cada classificador.

A utilização destes sistemas no contexto educacional é de grande relevância, uma vez que permite realizar um diagnóstico prévio de abandono pelo aluno ou possibilidade de evasão escolar precoce (Young e Sunbok, 2019; Kumari, Jain e Pamula, 2018). A associação de modelos preditivos com Sistemas de Recomendação (SR) permite notificar os professores e tutores acerca deste diagnóstico para que os mesmos tentem motivar esses alunos enviando mensagens personalizadas. Os sistemas de recomendação têm se tornado cada vez mais populares ao antecipar necessidades e gerar sugestões personalizadas para os usuários (Nicola Capuano et al., 2019).

A principal questão de pesquisa é a seguinte: É possível usar um comitê de classificadores de Aprendizado de Máquina para identificar alunos com desistência ou possibilidade de evasão escolar precoce em uma disciplina específica? A proposta é a arquitetura de predição DPE-PRIOR baseada em um comitê de classificadores. A solução é capaz de sincronizar e gerenciar vários métodos de Aprendizado de Máquina que, combinados, alcançam um resultado final com maior precisão. A arquitetura inclui uma camada de pré-processamento capaz de limpar e estruturar os dados provenientes do Ambiente Virtual de Aprendizagem (AVA). Por fim, avaliamos a arquitetura em turmas do curso de Licenciatura em Computação da Universidade Federal de Juiz de Fora.

Este artigo está organizado da seguinte forma: a seção 2 descreve a fundamentação teórica, com os conceitos fundamentas para o entendimento deste trabalho. A seção 3 apresenta os trabalhos relacionados que preveem o desempenho dos alunos e a possibilidade de evasão. A seção 4 descreve a arquitetura DPE-PRIOR e o modelo do comitê de classificadores. Na seção 5, apresentamos uma análise de evasão no curso de licenciatura em computação, utilizando a solução proposta. Finalmente, na seção 6, resumimos nossas contribuições e apresentamos os trabalhos futuros.

2. Fundamentação teórica

A evasão escolar é considerada quando um aluno matriculado em uma instituição de ensino decide abandonar voluntariamente o curso ou turma. A evasão do aluno decorre de fatores acadêmicos e não acadêmicos. (Diego et al., 2020) destacam alguns desses fatores: desempenho acadêmico, hábitos institucionais, interação social, restrições financeiras, motivação e personalidade. O abandono precoce causa perdas monetárias e, também, custos sociais. Por outro lado, (Amelec e Bonerge, 2019) mostram que o bem estar do aluno, as relações interpessoais e a frequência às aulas contribuem positivamente para sua adaptação ao curso.

Esta seção apresenta uma visão geral dos principais conceitos e definições de sistemas de recomendação, modelos preditivos de aprendizagem de máquina e estilos de aprendizagem que serão os fundamentos da abordagem proposta nesta pesquisa.

2.1. Sistemas de recomendação

Um Sistema de Recomendação pode ser "qualquer sistema que produza recomendações individualizadas como saída ou tenha o efeito de orientar o usuário de forma personalizada para objetos interessantes ou úteis em um grande espaço de opções possíveis" (Pereira et al., 2018). Esses sistemas têm como objetivo recomendar conteúdos aos usuários com base em seus perfis e contexto, considerando suas preferências, necessidades e interesses (Burke, 2002; J. et al., 2013). Técnicas de recomendação são usadas para ajudar a caracterizar o perfil e o contexto do usuário, alocá-los em grupos com necessidades semelhantes, localizar recursos que atendam às suas necessidades e projetar estratégias para recomendar de forma mais eficaz.

A primeira etapa em um Sistema de Recomendação é a extração de dados, que é responsável por extrair dados do perfil do usuário e do contexto ao qual ele pertence. Essa é a primeira camada a ser ativada no processo de recomendação. É importante extrair os dados relevantes dos perfis dos usuários para caracterizar suas preferências. A filtragem é a segunda etapa, responsável por filtrar as informações, correlacionando as preferências dos usuários. A terceira etapa está relacionada ao modelo do sistema, onde o algoritmo usado no processo de recomendação opera com o contexto e os dados de entrada para fornecer as recomendações ao usuário. Os Sistemas de Recomendação usam esses modelos para escolher o conteúdo mais apropriado para um determinado usuário. Na etapa final, os recursos selecionados são apresentados ao usuário como recomendação.

Este processo é contínuo, ou seja, quando ocorre uma recomendação e o usuário indica que está satisfeito com o recurso recebido, o resultado positivo é utilizado pelo sistema para agregar a sua preferência para realizar outras recomendações.

2.2. Modelos preditivos

O uso de abordagens de comitê de classificadores de Aprendizado de Máquina é amplamente explorado na literatura. O objetivo original de usar sistemas de decisão baseados em comitês é melhorar a assertividade de uma decisão. Os estudos indicam que ao considerar várias opiniões e combiná-las por meio de algum processo inteligente para chegar a uma decisão final pode reduzir a variação dos resultados.

Selecionamos os modelos clássicos de Aprendizado de Máquina da literatura e os mais aderentes ao problema de classificação supervisionada (Han, Pei e Kamber, 2011; Breiman, 2001; Braz et al., 2019):

Árvore de Decisão: é uma estrutura abstrata caracterizada por uma árvore em que cada nó denota um teste em um valor de atributo, cada ramo representa um resultado do teste e as folhas da árvore representam classes ou distribuições de classes.

K-Nearest Neighbours (KNN): é um algoritmo de aprendizado supervisionado. É um classificador no qual o processo de aprendizagem se baseia na "semelhança" de dois elementos. O treinamento consiste em vetores n-dimensionais, que medem a distância entre uma determinada tupla de teste com tuplas de treinamento e os compara.

SVM: é um algoritmo de classificação que trabalha com dados lineares e não lineares. Ao usar kernels, este modelo transforma os dados originais em uma dimensão superior, de onde é possível pesquisar e encontrar um hiperplano para os dados lineares ideais usando tuplas de treinamento chamadas vetores de suporte.

Floresta aleatória: é um classificador que consiste em uma coleção de classificadores estruturados em árvore (árvore de decisão) que ajusta as sub-árvores de decisão a várias subamostras do conjunto de dados e usa a média para melhorar a precisão preditiva. É um comitê de árvores de decisão.

Regressão Logística: é uma abordagem linear generalizada que modela a probabilidade de algum

evento ocorrer e é modelada como uma função linear de um conjunto de variáveis preditoras.

Multi-Layer Perceptron: Este modelo é representado por uma rede neural que contém neurônios para passar os dados através dela. O modelo pode adotar um aproximador de função não linear para classificação ou regressão.

No entanto, existem alguns desafios nas abordagens preditivas. Identificar o contexto semântico ao prever as preferências e necessidades de um indivíduo é um desafio que apresenta grande dificuldade em identificar de forma assertiva o significado semântico de tais necessidades e preferências. É importante notar que a identificação do contexto do indivíduo é a própria base preditiva (Gao et al., 2018).

Alcançar maior precisão por meio de maior acurácia dos modelos preditivos é outro desafio (Anam et al., 2017). Os resultados podem ser influenciados pelo volume de dados, sendo que dados antigos geram resultados menos precisos e, quando a quantidade de dados é muito grande, distorções e valores ausentes podem influenciar as previsões (H. et al., 2017).

2.3. Estilos de aprendizagem

Um Ambiente Virtual de Aprendizagem (AVA) contém informações valiosas sobre o estilo de aprendizagem do aluno. (Leonardo et al., 2018) definem como exemplos dessas informações o Learner Learning Trail (LLT), que é a sequência de interações entre os alunos e o ambiente virtual, e o Learner Learning Style (LLS), que está associado ao comportamento e escolhas do aluno durante o processo de aprendizagem.

Durante uma experiência de aprendizagem, é importante considerar a personalidade dos alunos a fim de encontrar e entregar o melhor recurso disponível (Chi, Chen, & Tsai, 2014 aput Nicola Capuano et al., 2014). Identificar estilos de aprendizagem não significa rotular os alunos e adaptar instruções para se adequar às suas preferências, mas propor recomendações cada vez mais aderentes ao seu perfil e contexto. Um grande desafio na área de Sistemas de Recomendação Educacional é a personalização de recomendações.

Vários modelos descrevem as classificações de um aluno em um estilo de aprendizagem específico (Buiar, Andrey e Oliveira, 2017). Os Modelos de Estilo de Aprendizagem classificam os alunos de acordo com escalas predeterminadas. Os modelos mais citados na literatura são os de Kolb, Felder e Silverman e Vark (Nascimento et al., 2017; Valaski, Malucelli e Reinehr, 2011; Carvalho et al., 2017).

Essa pesquisa se concentra em combinar os benefícios do contexto de educação a distância com o poder dos Sistemas de Recomendação, com o objetivo de prever a evasão precoce de alunos. O abandono escolar é um problema sério para os alunos, a sociedade e as instituições. Por meio dos sistemas de e-learning, podemos notificar automaticamente professores e tutores e os próprios alunos. Para os alunos, ter uma notificação personalizada (por exemplo, vídeos ou mensagens), com base no seu estilo de aprendizagem, pode motivar a volta e o reengajamento às disciplinas.

3. Trabalhos aelacionados

A previsão da evasão escolar por meio de técnicas estatísticas e de Aprendizado de Máquina tem ganhado cada vez mais atenção (Diego et al., 2020). Inúmeros estudos têm sido feitos nos Ambientes Virtuais de Aprendizagem, enfocando o comportamento dos alunos em uma disciplina ou curso. Essas técnicas têm sido utilizadas, principalmente, para classificar os alunos com base em suas atividades de aprendizagem. Selecionamos alguns trabalhos de pesquisa que utilizam modelos de Aprendizado de Máquina em uma abordagem de comitê de classificadores para evitar a evasão escolar no contexto de disciplinas ou cursos on-line.

O trabalho desenvolvido por (Márquez-Vera et al., 2016) apresenta uma forma de prever antecipadamente a evasão escolar. A metodologia usa regras para definir a probabilidade de evasão do aluno e considera faixas de tempo ao longo do curso para prever essa probabilidade. Os autores ressaltam que não é necessário esperar até o final do curso para prever e tomar uma decisão para reagir e prestar ajuda específica aos alunos que estão apresentando desengajamento e com risco de evasão.

Outro trabalho que teve como objetivo predizer o desempenho dos alunos por meio de dados educacionais foi apresentado por (Barbosa et al., 2017). Os autores propuseram um método baseado na Análise de Componentes Principais para a identificação de padrões relevantes em relação às suas características. O objetivo principal é interpretar padrões em conjuntos de dados educacionais e reduzir a dimensionalidade em tarefas de predição. No contexto educacional, a pesquisa busca compreender os fatores que afetam o desempenho dos alunos no ambiente educacional, bem como prever seus desempenhos.

(Leonardo et al., 2018) propõem um modelo capaz de integrar dados gerados a partir do comportamento dos alunos na Educação a Distância

com aspectos cognitivos, tais como seus Estilos de Aprendizagem, cruzando a curva de aprendizagem com o estilo de aprendizagem do aluno. O estudo com 202 alunos avaliou se os estilos de aprendizagem são capazes de explicar aspectos do comportamento do aluno. A dimensão do estilo de aprendizagem Sequencial/Global foi capaz de explicar o abandono mais do que as outras dimensões.

Em (Kumari, Jain e Pamula, 2018), os autores propuseram um modelo para avaliar o impacto das características comportamentais de aprendizagem do aluno com base em seu desempenho acadêmico. A tarefa de análise de desempenho é conduzida usando a Classificação como uma técnica de mineração de dados. Quatro classificadores foram usados: ID3, Naive Bayes, K-Nearest Neighbor (KNN), Support Vector Machine (SVM). Para melhorar o desempenho dos classificadores e a precisão do modelo de desempenho do aluno, os autores utilizaram os métodos ensemble Bagging, Boosting e Voting.

De acordo com (Young e Sunbok, 2019) "a modelagem preditiva usando Aprendizado de Máquina tem um grande potencial no desenvolvimento de sistemas de alerta para identificar antecipadamente os alunos em risco de evasão e ajudá-los". O estudo usa o modelo de Floresta Aleatória para prever os alunos em risco de abandono no ensino médio.

No trabalho desenvolvido por (Cerezo et al., 2020), os autores propuseram um algoritmo para descobrir a aprendizagem autorregulada dos alunos durante um curso a distância usando técnicas de Mineração de Processos. Os autores aplicaram um novo algoritmo no domínio educacional denominado Inductive Miner na plataforma Moodle. A técnica foi capaz de descobrir modelos ótimos em termos de aptidão para alunos aprovados e reprovados, bem como modelos com um certo nível de granularidade.

(Diego et al., 2020) aplicam um modelo Uplift ao problema da prevenção da evasão escolar. A modelagem Uplift é uma abordagem para estimar o efeito incremental de uma ação ou tratamento no nível individual. Eles propuseram uma abordagem para maximizar a eficácia dos esforços no combate a retenção em instituições de ensino superior, ou seja, a melhoria do desempenho acadêmico por meio da oferta de tutoriais. Seus resultados demonstram as vantagens da modelagem Uplift na adaptação dos esforços de retenção no ensino superior em relação às abordagens convencionais de modelagem preditiva.

3.1. Análise comparativa

Como em nossa proposta, (Kumari, Jain e Pamula, 2018; Young e Sunbok, 2019) utilizaram comitês para melhorar o desempenho dos classificadores e a precisão do modelo. (Cerezo et al., 2020) usaram técnicas de mineração, enquanto nesta pesquisa e em (Young e Sunbok, 2019) foram utilizados modelos de Aprendizado de Máquina. A plataforma Moodle também foi utilizada para avaliar a proposta e identificar os alunos com dificuldades. (Young e Sunbok, 2019) também usaram métricas de desempenho, assim como foi feito nesta pesquisa. (Diego et al., 2020) apresentam outra abordagem para o problema da prevenção da evasão escolar, mas também com foco nas instituições de ensino superior.

Prever a evasão escolar de forma precoce é o foco de (Márquez-Vera et al., 2016) e, como na proposta deste trabalho, permitem que os professores ajudem os alunos em risco de abandono. (Babosa et al., 2017) lidam com os fatores que afetam o desempenho dos alunos no ambiente educacional. Seus resultados contribuem para o nosso estudo sobre o perfil e contexto dos alunos. Os trabalhos de (Carvalho et al., 2017) e (Leonardo et al., 2018) tratam dos estilos de aprendizagem e sua categorização, que foram adotados neste trabalho para enviar uma notificação personalizada.

Motivados por estes estudos e pelos resultados anteriores do nosso grupo de pesquisa em Sistemas de Recomendação Educacional e assistência em sistemas e-Learning (Pereira et al., 2018), este trabalho propõe prevenir a evasão escolar por meio da concepção de um modelo preditivo. Foram adotadas seis técnicas clássicas de aprendizado de máquina na construção do Comitê de Classificadores. A proposta foi avaliada com o objetivo de garantir um maior nível de confiança e destacando a importância de notificar professores, tutores e alunos dos resultados das previsões.

4. DPE-PRIOR: comitê de classificadores para predição de evasão

Esta seção descreve a arquitetura conceitual da pesquisa, os aspectos de desenvolvimento e as tecnologias adotadas. Propomos um modelo de Comitê de Classificadores baseado em modelos clássicos de Aprendizado de Máquina usados como núcleo por um

Sistema de Recomendação. O modelo pode prever se um aluno apresenta risco de abandono da disciplina e o sistema pode notificar os professores e tutores e os próprios alunos sobre essa possibilidade.

No contexto da identificação do desempenho e do perfil do aluno, a arquitetura é composta por seis camadas principais (Figura 1).

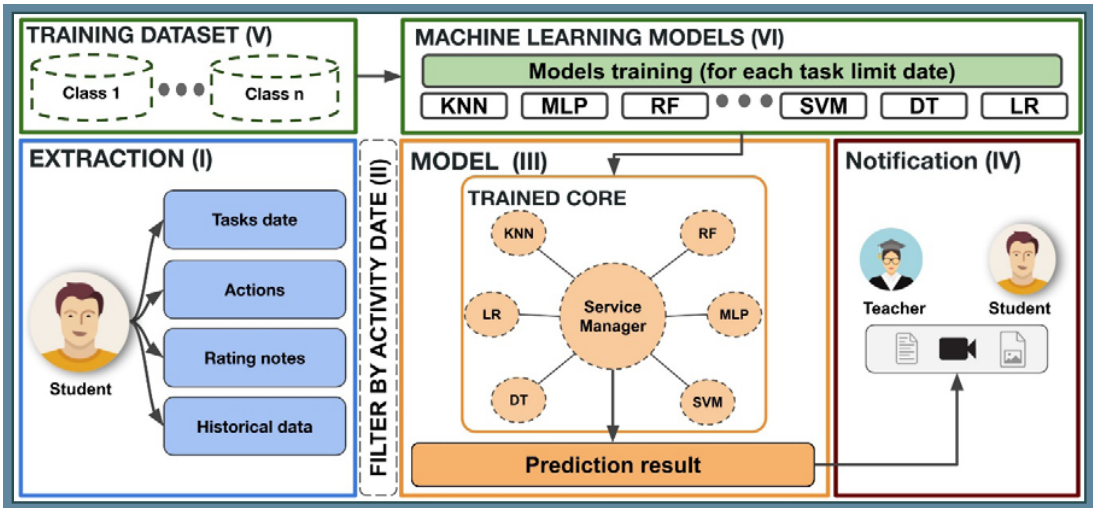

Figura 1 - Arquitetura DPE-PRIOR

A Camada de Extração é responsável por extrair todas as informações do aluno necessárias para compor seu perfil de atividades. No contexto educacional, informações como ações do aluno e notas de avaliação são importantes para entender a trajetória do discente, bem como seu desempenho na disciplina; sendo que suas preferências definirão seu estilo de aprendizagem e o envio de mensagens personalizadas.

Como pretendemos prever a probabilidade de evasão do aluno, é necessário filtrar o seu desempenho ao longo da disciplina. Essa etapa é feita pela Camada de Filtragem que filtra todas as atividades dos alunos por data limite, pré-estabelecida pelo professor.

O núcleo do modelo é composto por um Comitê de Classificadores que combina diferentes modelos clássicos de Aprendizado de Máquina, proporcionando um resultado mais preciso e com maior confiança. Cada modelo é um serviço autônomo capaz de atender às solicitações e prever a probabilidade de evasão do aluno. No núcleo do Comitê, definimos seis modelos diferentes de Aprendizado de Máquina para compor os principais serviços autônomos. Visando sincronizá-los e combiná-los, utilizamos um sétimo serviço, que funciona como coordenador do Comitê proposto. Todos os modelos adotados são supervisionados, visto que estamos lidando com um problema de classificação, que consiste em indicar se um aluno pode abandonar o curso ou não.

A decisão do Comitê de Classificadores é feita por meio de um processo de votação, por meio da média das previsões positivas. Esta abordagem visa minimizar a diferença entre a previsão dos modelos e maximizar a assertividade do comitê (Kumari, Jain e Pamula, 2018). Uma notificação é enviada ao professor responsável pela turma pela Camada de Notificação, propiciando uma intervenção por parte do professor ou tutor para evitar que o aluno abandone a turma. Também pode enviar mensagens aos alunos como motivação para evitar o seu desligamento.

As outras duas camadas são responsáveis pelo treinamento dos modelos de Aprendizado de Máquina com dados de outras turmas anteriores. Este processo está descrito em detalhes na próxima seção.

4.1. Processo de predição

O Módulo Preditor é composto por modelos clássicos de Aprendizado de Máquina destinados a funcionar em conjunto, capazes de resolver problemas de classificação de forma autônoma. É responsável por prever os recursos e sua aderência aos usuários. Pode ser realizado por um modelo baseado em memória, um modelo baseado em modelo ou mesmo um modelo híbrido, representando uma combinação de ambos. A Figura 2 mostra as duas camadas principais que contêm o processo de fluxo de dados: o modelo de treinamento e o próprio modelo preditivo.

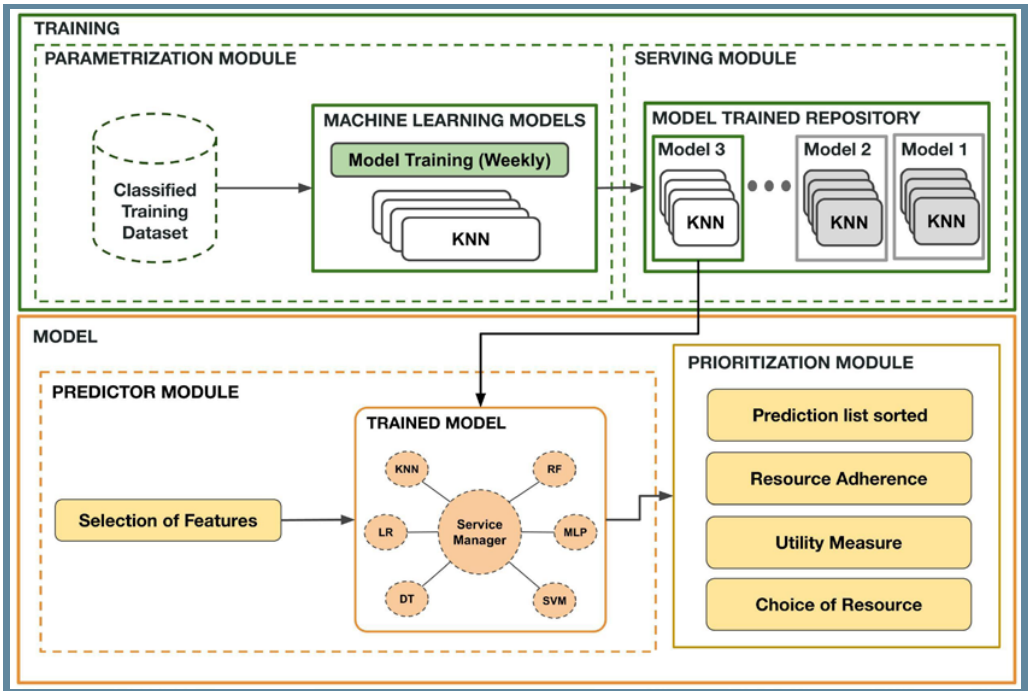

Figura 2 - Processo de predição

A Camada de Treinamento é responsável por armazenar os dados e treinar os modelos individuais. Cada modelo deve ser treinado separadamente com os mesmos dados, gerando diferentes preditores. Cada modelo é testado com vários parâmetros visando maximizar sua abordagem de predição. Uma vez selecionados os melhores parâmetros e treinados, todos os modelos são armazenados como serviços para ficarem disponíveis para a camada de modelo.

O Comitê de Classificadores foi projetado para ser composto por métodos que combinam diferentes modelos clássicos de Aprendizado de Máquina, implementados como serviços autônomos de software, com reatividade, inteligência e características sociais. Cada serviço, composto por um modelo de aprendizado de máquina, pode manipular de forma autônoma os dados de entrada e oferecer um resultado de saída.

A camada de modelo é responsável por fazer predições em tempo real. O módulo preditor possui um gerenciador de serviços que coordena cada modelo de Aprendizagem de Máquina como um serviço. Com o objetivo de sincronizar todos os modelos, o gerenciador de serviços aciona todos os preditores em paralelo para calcular a entrada. Cada modelo calcula os recursos selecionados e, pela porcentagem de precisão, um Método de Votação (Polikar, 2012) é aplicado para avaliar os resultados que a maioria dos modelos têm em comum para obter a previsão final com maior precisão.

O Módulo de Priorização classifica os resultados e obtém os objetos de recomendação que são mais aderentes a cada caso de priorização. Depois disso, ele pode apresentar o objeto selecionado ao usuário.

5. Análise de evasão no Curso de Licenciatura em Computação

A disciplina Fundamentos de Sistemas de Informação é oferecida para a maioria dos cursos da área de Ciência da Computação. Seu principal objetivo é preparar os alunos para reconhecer a importância dos Sistemas de Informação em diferentes organizações e identificar diferentes possibilidades para sua implementação e uso. Na Universidade Federal de Juiz de Fora, Brasil, essa é uma disciplina do currículo de curso de Licenciatura em Computação na modalidade a distância.

Selecionamos duas turmas do curso de Licenciatura em Computação, ofertado na modalidade à distância, 2018 e 2019, para avaliação da proposta. A disciplina utiliza o Ambiente Virtual de Aprendizagem Moodle e as interações do professor e tutor e dos alunos são baseadas em mensagens, fórum, chat, wiki, sendo que os alunos podem escolher seus respectivos grupos para desenvolvimento das atividades. A programação das aulas inclui várias atividades a serem realizadas pelos alunos e tarefas de avaliação.

O tratamento dos dados fixou os valores faltantes com zero (0) e substituiu todos os caracteres especiais por caracteres válidos. Após o pré-processamento, os dados foram divididos por data, gerando um arquivo

para cada data limite de entrega de tarefa, classificando os resultados como ativos, que indicam aprovação do aluno e inativos caso contrário.

Com o objetivo de obter a melhor configuração para o nosso problema, combinamos as diferentes configurações do modelo e geramos um arquivo CSV com um total de 36.000 combinações. Em seguida, treinamos os dados para cada combinação até obtermos alta precisão em cada uma. Este processo gerou um total de 37 configurações diferentes. Definimos cada modelo com a configuração mais representativa e precisa para ele. As técnicas utilizadas foram:

Árvore de Decisão: Com o arquivo CSV gerado usamos a entropia como critério e mantivemos os valores padrão para os demais parâmetros.

KNN: Para este modelo, definimos K = 3, que representa o número de vizinhos a serem comparados. O tipo de distância utilizada foi a distância euclidiana e para os demais parâmetros mantivemos os valores padrão.

SVM: Definimos a função linear como tipo de kernel e valor gama com 100, bem como a regularização igual a 1.

Floresta aleatória: A entropia foi usada como critério. O número de estimadores foi definido para 60.

Regressão logística: A regularização foi definida com o valor 100 e o solver liblinear para lidar com os dados, o que também é recomendado na documentação do framework (Han et al., 2011).

Multi-Layer Perceptron: Usamos 3 camadas ocultas com 6 nós cada e a função de ativação ReLU; uma camada de saída para casos positivos com a função de ativação ReLU, sendo que a camada de entrada tem o número de recursos tratados no conjunto de dados. Além disso, usamos a precisão como métrica, a entropia cruzada binária como função de perda e o otimizador de Adam.

A amostra de teste é responsável por mostrar os resultados das previsões e foi composta pelo conjunto de dados da classe atual. Foram escolhidas as medidas F-measure, Precision e Recall para avaliar a acurácia. Comparamos os modelos preditivos juntos no Comitê de Classificadores, bem como seus resultados individualmente.

Nas Equações 1, 2 e 3, TP representa as classificações positivas verdadeiras, o que significa que a predição teve um resultado positivo e a classificação foi positiva (Sokolova e Lapalme, 2009). O FP representa as classificações falso positivas, o que significa que o resultado foi positivo, mas a classificação correta foi negativa e o FN representa o falso negativo para o caso contrário.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F - measure = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (3)$$

A turma contou com 37 alunos de 12 cidades diferentes. As principais atividades de avaliação incluem a participação em dois fóruns, duas tarefas individuais e uma tarefa final em grupo. A segunda tarefa inclui uma visita a uma escola para identificar as características do seu Sistema de Informação e a atividade de grupo propõe o planejamento de um componente de um Sistema de Informação Escolar. Os grupos eram compostos por um, dois ou três membros. O processo de avaliação também inclui uma revisão por pares da apresentação final da atividade.

A evasão escolar é um verdadeiro desafio e, considerando a educação a distância, temos que lidar com o desengajamento do aluno desde o início, pois, algumas vezes, ele não realiza sequer uma tarefa. Os dados do aluno foram coletados para prever o desempenho na disciplina e a mesma estrutura de dados dos modelos foi mantida. Executamos o experimento e, para cada tarefa entregue, uma previsão foi feita. Quando os resultados indicam entrega inativa, uma notificação é enviada ao professor e ao aluno, como uma mensagem personalizada.

5.1. Resultados dos experimentos e discussão

Avaliamos os resultados usando as Equações 1, 2 e 3 e os comparamos para visualizar e compreender as diferenças entre os modelos.

Os principais resultados da acurácia apresentados na Figura 3 mostram que, individualmente, os modelos de Aprendizado de Máquina apresentam uma estrutura linear em que a acurácia observada não aumentou ao longo do tempo de forma ascendente, apresentando variações. No entanto, em nossa solução, que combina os modelos usando o método de Comitê de Classificadores com votação, a precisão aumenta gradualmente ao longo do tempo conforme a disciplina avança e os alunos têm mais atividades em sua programação.

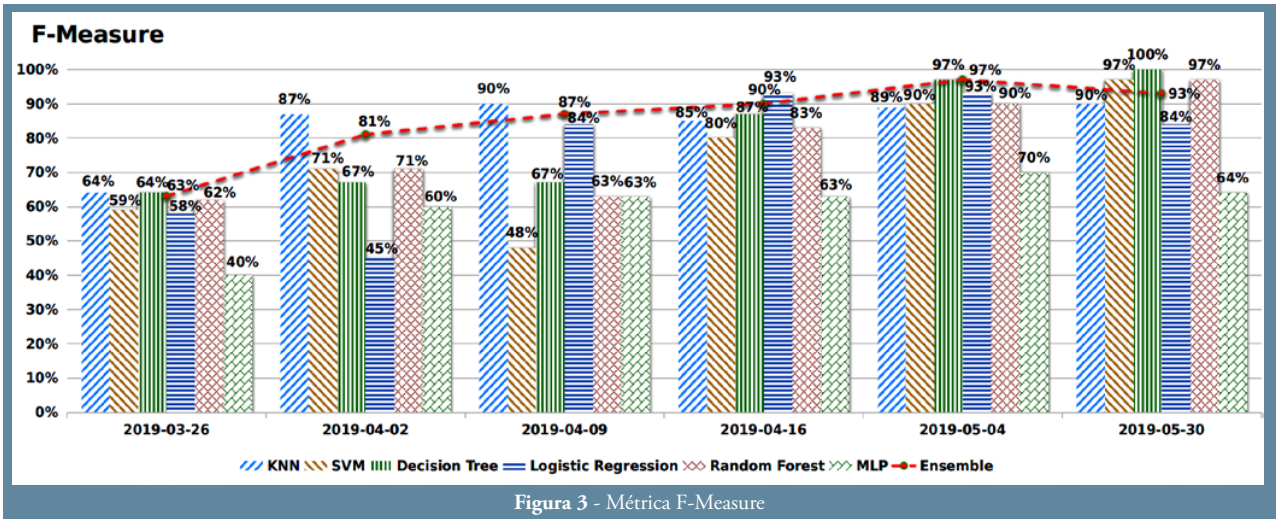

Figura 3 - Métrica F-Measure

Individualmente, alguns modelos apresentam melhor desempenho em métricas específicas, como a F-Measure do modelo Árvore de Decisão. No entanto, o resultado geral mostra um ganho de desempenho e precisão nas previsões por meio do Comitê de Classificadores.

A Tabela 1 mostra, do total de sete tarefas da aula, o número de alunos que não fizeram nenhuma das sete tarefas e relaciona a entrega da tarefa com o desempenho final dos alunos na disciplina, considerando os alunos que foram aprovados ou reprovados

Tabela 1 - Tarefas entregues pelos alunos X desempenho final

Entrega de Tarefas	Número de Alunos	Aprovados	Reprovados
Zero (nenhuma)	11	0	11
Uma	3	0	3
Duas	2	0	2
Três	1		1
Quatro	0	0	0
Cinco	3	2	1
Seis	10	9	1
Sete (todas)	7	7	0
Total	37	18	19

Os resultados mostram que 29,72% dos alunos não realizaram nenhuma tarefa. 16,20% realizaram pelo menos uma tarefa. 8,10% fizeram 3 de 7 tarefas. 27,00% fizeram 6 de sete tarefas, e uma delas faltou. Por fim, 18,91% realizaram todas as tarefas. Podemos dizer que pelo menos 43,24% dos alunos deveriam ter tido atenção e motivação especial e, pelo menos, 8,10% poderiam ser aprovados concluindo mais uma ou duas tarefas.

Considerando os resultados, podemos inferir que o modelo proposto, provavelmente, poderia ajudar a evitar tantas evasões no curso da seguinte forma:

- Alunos que não fizessem nenhuma tarefa seriam notificados desde a primeira tarefa.
- Alunos que não realizassem alguma tarefa seriam notificados após o término da tarefa.
- Todos os alunos em dia com a disciplina não seriam notificados pelo sistema.
- O professor ou tutor seriam notificados sempre quando um aluno não fizesse uma tarefa.

É importante não ter como alvo alunos com desempenho excelente. O Comitê deve identificar aqueles que terão seu desempenho e engajamento afetados positivamente pelas recomendações. Com isso, torna-se possível reverter quadros de evasão e desengajamento.

Finalmente, o sistema de previsão DPI-PRIOR pode ajudar automaticamente na:

- Detecção precoce de alunos com alguma dificuldade, alertando os profissionais da educação sobre isso.
- Possibilidade de professor e tutor atuarem junto com o aluno para solucionar o problema assim que ele surgir.
- Permitir que ações sejam tomadas com base nas necessidades de cada aluno e identificar os tipos preferidos de mensagens dos alunos.

6. Considerações finais

Este estudo aplica um Comitê de Classificadores para prever o desempenho do aluno para evitar a evasão escolar. Essa abordagem pode ajudar os alunos a conhecer seu desempenho, notificando e permitindo que melhorem sua dedicação ao curso. Para os professores e tutores, também permite compreender seus alunos,

além de perceber suas deficiências e necessidades, sua forma de aprender, possibilitando o aprimoramento de sua teoria didático-pedagógica de ensino.

Esta pesquisa teve como principal contribuição uma arquitetura preditiva, visando evitar a evasão escolar em disciplinas ou cursos na modalidade de educação a distância. O objetivo principal era unir os dados dos Ambientes Virtuais de Aprendizagem ao poder preditivo dos modelos de Aprendizado de Máquina. Um Comitê de Classificadores foi adotado para maximizar a precisão da proposta para identificar o desempenho dos alunos. No contexto do curso de Licenciatura em Computação, os resultados com a aplicação da proposta demonstram as vantagens da abordagem utilizada e seus resultados foram significativos.

Considerando a disponibilidade de dados educacionais e o uso de modelos preditivos para cada contexto, os modelos clássicos oferecem alto desempenho, mas as abordagens que combinam esses modelos se destacam em comparação com os modelos individuais. Nossa proposta utiliza um Comitê que permite maior assertividade e confiança em cada resultado.

O potencial deste trabalho permite o desenvolvimento e a continuidade de outras pesquisas sobre o mesmo contexto. Não podemos generalizar os resultados, mas percebemos os benefícios da adoção da proposta na educação a distância.

A grande quantidade de dados educacionais disponíveis tem o potencial de construir novos conhecimentos, notadamente nos que utilizam os Ambientes Virtuais de Aprendizagem. Os processos de mineração de dados podem explorar esses dados e prever o desempenho dos alunos. Em trabalhos futuros, pretendemos fazer recomendações de conteúdos que possam prevenir a evasão escolar, motivando os estudantes a envolverem-se nas aulas e ajudando-os a obter um melhor resultado na disciplina.

Agradecimento

Este trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financiamento 001, Universidade Federal de Juiz de Fora (UFJF), CNPq e FAPEMIG.

7. Referências

Amelec, V. and Bonerge, P. L. O. (2019) "Mixture Structural Equation Models for Classifying University Student Dropout in Latin America," *Procedia Computer Science*. Elsevier BV, pp. 629–634. doi: 10.1016/j.procs.2019.11.036.

Anam, M. *et al.* (2017) "A statistical analysis based recommender model for heart disease patients," *International Journal of Medical Informatics*. Elsevier BV, pp. 134–145. doi: 10.1016/j.ijmedinf.2017.10.008.

Bagheri, M. and Movahed, S. H. (2016) "The effect of the internet of things (Iot) on education business model," in *12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 435–441.

Barbosa, A. *et al.* (2017) "A machine learning approach to identify and prioritize college students at risk of dropping out," in *Brazilian Symposium on Computers in Education - SBIE)*.

Braz, F. *et al.* (2019) "An early warning model for school dropout: a case study in e-learning class," in *Brazilian Symposium on Computers in Education - SBIE*. doi: http://dx.doi.org/10.5753/cbie.sbie.2019.1441.

Breiman, L. (2001) "Random forests," *Machine Learning*. United States, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Buiar, J. A., Andrey, P. and Oliveira, R. (2017) "Identificação de Estilo de Aprendizagem: Um modelo de inferência automatizado baseado no perfil de personalidade identificado nos textos produzidos pelo aluno," in *Brazilian Symposium on Computers in Education - SBIE*. doi: http://dx.doi.org/10.5753/cbie.sbie.2017.1157.

Burke, R. (2002) "Hybrid recommender systems: Survey and experiments," *User Modelling and User-Adapted Interaction*. United States, 12(4), pp. 331–370. doi: 10.1023/A:1021240730564.

Carvalho, V. *et al.* (2017) "OntAES: Uma Ontologia para Sistemas Adaptativos Educacionais Baseada em Objetos de Aprendizagem e Estilos de Aprendizagem," in *Brazilian Symposium on Computers in Education - SBIE*, pp. 1307–1316. doi: http://dx.doi.org/10.5753/cbie.sbie.2017.1307.

Cerezo, R. *et al.* (2020) "Process mining for self-regulated learning assessment in e-learning," *Journal of Computing in Higher Education*. Spain: Springer, 32(1), pp. 74–88. doi: 10.1007/s12528-019-09225-y.

Diego, O. *et al.* (2020) "Uplift Modeling for preventing student dropout in higher education," *Decision Support Systems*. Elsevier BV, p. 113320. doi: 10.1016/j.dss.2020.113320.

Gao, S. *et al.* (2018) "Pairwise Preference over Mixed-Type Item-Sets Based Bayesian Personalized Ranking for Collaborative Filtering," *Proceedings - 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 2017 IEEE 15th International Conference on Pervasive Intelligence and Computing, 2017 IEEE 3rd International Conference on Big Data Intelligence and Computing and 2017 IEEE Cyber Science and Technology Congress, DASC-PICom-DataCom-CyberSciTec 2017*. China: Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.22.

H., C. J. *et al.* (2017) "Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets," *International Journal of Medical Informatics*. Elsevier BV, pp. 71–79. doi: 10.1016/j.ijmedinf.2017.03.006.

HAN, J., PEI, J. and KAMBER (2011) *Data mining: concepts and techniques*.

J., B. *et al.* (2013) "Recommender systems survey," *Knowledge-Based Systems*. Elsevier BV, pp. 109–132. doi: 10.1016/j.knosys.2013.03.012.

Kumari, P., Jain, P. K. and Pamula, R. (2018) "An efficient use of ensemble methods to predict students academic performance," *Proceedings of the 4th IEEE International Conference on Recent Advances in Information Technology, RAIT 2018*. India: Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/RAIT.2018.8389056.

Leonardo, H. *et al.* (2018) "Diagnosis of learner dropout based on learning styles for online distance learning," *Telematics and Informatics*. Elsevier BV, pp. 1593–1606. doi: 10.1016/j.tele.2018.04.007.

Márquez-Vera, C. *et al.* (2016) "Early dropout prediction using data mining: A case study with high school students," *Expert Systems*. Mexico: Blackwell Publishing Ltd, 33(1), pp. 107–124. doi: 10.1111/exsy.12135.

Nascimento *et al.* (2017) "Recomendação de Objetos de Aprendizagem baseada em Modelos de Estilos de Aprendizagem: Uma Revisão Sistemática da Literatura," *Brazilian Symposium on Computers in Education - SBIE*, 28.

NEVES, Felipe; CAMPOS, Fernanda; STROELE, Victor; DANTAS, Mario; DAVID, José Maria, BRAGA, Regina. (2021). Assisted education: Using predictive model to avoid school dropout in e-learning systems. In book: Intelligent Systems and Learning Data Analytics in Online Education (pp.153-178).

Nicola Capuano, Matteo Gaeta, Pierluigi Ritrovato, Saverio Salerno, Elicitation of latent learning needs through learning goals recommendation, Computers in Human Behavior, Volume 30, 2014, Pages 663-673, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2013.07.036.

Nicola Capuano, Francisco Chiclana, Enrique Herrera-Viedma, Hamido Fujita, Vincenzo Loia, Fuzzy Group Decision Making for influence-aware recommendations, Computers in Human Behavior, Volume 101, 2019, Pages 371-379.

Pereira, C. K. *et al.* (2018) "BROAD-RSI – educational recommender system using social networks interactions and linked data," *Journal of Internet Services and Applications*.

Polikar, R. (2012) "Ensemble learning," in *Ensemble Machine Learning: Methods and Applications*. United States: Springer US, pp. 1–34. doi: 10.1007/9781441993267_1.

Sokolova, M. and Lapalme, G. (2009) "A systematic analysis of performance measures for classification

tasks," *Information Processing and Management*. Canada, 45(4), pp. 427–437. doi: 10.1016/j.ipm.2009.03.002.

T. Daradoumis, R. Bassi, F. Xhafa, and S. Caball´e. A review on massive e-learning (mooc) design, delivery and assessment. In 2013 Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, pages 208–213, 2013.

Valaski, J., Malucelli, A. and Reinehr, S. (2011) "Revisão dos Modelos de Estilos de Aprendizagem Aplicados à Adaptação e Personalização dos Materiais de Aprendizagem," in *Brazilian Symposium on Computers in Education - SBIE. Aracaju*, pp. 844–847.

Young, C. J. and Sunbok, L. (2019) "Dropout early warning systems for high school students using machine learning," *Children and Youth Services Review*. Elsevier BV, pp. 346–353. doi: 10.1016/j.childyouth.2018.11.030.

Recebido em: 28/08/2021
Aceito em: 26/10/2021