Algoritmos de Mineração de Dados para Análise de Evasão na Graduação da Universidade de Brasília

Tiago de Souza Fernandes
Depto. de Ciência da Computação
Universidade de Brasília
tiagotsf2000@gmail.com

***Resumo*—A evasão no Ensino Superior é um problema que preocupa diversas instituições do mundo, e uma solução que busca identificar suas causas e casos é a Mineração de Dados Educacional (EDM). Essa abordagem consiste no treino de algoritmos de aprendizado de máquina a partir de dados institucionais dos alunos. Esses algoritmos são capazes de identificar prematuramente alunos em risco de evasão. Este trabalho propõe aprimorar a abordagem de EDM atualmente desenvolvida na UnB, foram testados: novos atributos acadêmicos como cota e idade; um novo algoritmo de classificação, o CatBoost; e o uso de um método de extração de informações, o SHAP, capaz de melhorar a interpretabilidade dos resultados obtidos. Utilizando a metodologia Knowledge Discovery Process, foi obtido um classificador com sensibilidade de 91,0% e F-Beta de 86,4%, treinado com os dados de quatro cursos ligados a computação. Utilizando o SHAP para extração de informações a respeito do modelo gerado, os atributos acadêmicos como IRA e quantidade de créditos foram identificados como os mais relevantes, e uma relação entre a idade do aluno e a evasão do mesmo foi descoberta, indicando que alunos mais velhos possuem mais chances de evadir em cursos de computação na Universidade de Brasília.**

***Index Terms*—mineração de dados, evasão estudantil, aprendizado de máquina, educação.**

I. Introdução

As Instituições de Ensino Superior (IES) exercem um papel fundamental no desenvolvimento socioeconômico e científico-tecnológico de um país, pois são responsáveis por projetos de impacto na sociedade e pela formação de profissionais qualificados [1]. A evasão estudantil é um grande problema para essas instituições, pois impede a formação completa de profissionais, fazendo com que os recursos humanos e financeiros investidos não tenham nenhum retorno para a sociedade [2]. A evasão é resultado de uma combinação de fatores sociais, econômicos e pessoais [3], e no ensino superior em específico, existe uma maior dificuldade em identificar suas origens e causas, pois o contexto do aluno de ensino superior é complexo e sua trajetória pode tomar diversos caminhos [4].

A evasão estudantil ocorre em diversas IES no Brasil e no mundo [2]. A Universidade de Brasília (UnB), que registrou em 2016 taxas de 24,54% de evasão, reforçou no Plano de Desenvolvimento Institucional 2018-2022 (PDI) [5] a importância e o interesse que a instituição tem em mitigar os casos de evasão, entender as causas e origens do problema e propor melhores medidas institucionais de acompanhamento dos estudantes. A própria instituição possui algumas estratégias para tentar lidar com esse problema [5], alguns métodos usados para entender essas causas utilizam pesquisas baseadas em perguntas e formulários, contudo são métodos custosos de serem implementados e não podem ser generalizados para outras instituições [5], [6].

A *Educational Data Mining* (EDM), ou Mineração de Dados Educacional [7], é uma área de pesquisa crescente nos últimos anos que utiliza dados institucionais para obter conhecimento. A Mineração de dados (MD) é definida como o processo de descoberta de padrões a partir de dados [8], utilizando métodos estatísticos [9] ou de algoritmos de aprendizagem de máquina [10]. O uso dessa abordagem vêm tomando espaço como solução eficiente para a identificação prematura de casos e causas de evasão no ensino superior, por ser uma solução automatizada, eficaz, e que pode ser generalizada e adaptada para outras instituições (e.g. [6], [11], [12]).

Com o intuito de comprovar a viabilidade da abordagem de EDM na Universidade, alguns estudos de cunho exploratório foram realizados, utilizando dados institucionais de cursos específicos, para a predição de casos de evasão (e.g. [13], [14], [15], [16]), ou seja, para tentar classificar se um aluno vai ou não evadir antes que isso ocorra. Essa abordagem permite que a coordenação possa intervir e acompanhar o aluno, na tentativa de evitar sua evasão, ao identificar e lidar de forma prematura com as possíveis causas desse problema. O trabalho mais recente [16] propõe um preditor treinado utilizando dados acadêmicos e sociodemográficos dos estudantes do curso de Engenharia Mecatrônica, obtendo acurácia de 86.6% e sensibilidade de 88.1% na predição com o melhor modelo.

Contudo, um problema recorrente na Mineração de Dados, e presente no trabalho citado, é a falta de interpretabilidade dos modelos gerados. A utilização de algoritmos de aprendizado de máquina complexos produz modelos de difícil interpretabilidade e compreensão humana [17]. Por outro lado, algoritmos mais simples, apesar de serem mais transparentes e compreensíveis, não são capazes de identificar padrões complexos nos dados. No problema de predição de evasão, a compreensão do modelo gerado pelo algoritmo é de suma importância, não só para validar os resultados, mas para que seja possível extrair informações relevantes a respeito das causas de evasões por parte dos alunos.

O SHAP (Shapley Additive Explanations) é um método para explicações locais, baseado em conceitos da área de

teoria dos jogos, capaz de extrair informações a respeito do processo de predição de um modelo complexo [17]. Esse é um método estado da arte para interpretabilidade de modelos, e está sendo utilizado em trabalhos recentes relacionados ao uso de MD no problema da evasão (e.g. [18], [19]). Essa abordagem oferece uma compreensão das causas de evasão identificadas pelos modelos, de forma geral e individual para cada aluno, informação relevante para a tomada de decisões da gestão de Universidades.

Além disso, o trabalho [16] realiza a análise de forma restrita a apenas um curso de graduação, e não utiliza de alguns atributos não acadêmicos disponíveis, como a cota e a idade, que são informações importantes no contexto da evasão na graduação. A utilização de uma base de dados com diferentes cursos de graduação permite validar a generalização do modelo gerado, e a análise de novos atributos não-acadêmicos utilizando o SHAP pode revelar informações importantes para uma melhor compreensão do problema de evasão na UnB.

Nesse contexto, este trabalho propõe aprimorar a abordagem de EDM realizada no trabalho [16]. Os objetivos consistem na utilização do SHAP para extração de informações interpretáveis dos modelos gerados; na análise de dois novos atributos sociodemográficos, a cota e a idade, com base nas informações extraídas pelo SHAP; e por fim, na utilização de dados de diferentes cursos para o treinamento do algoritmo, para verificar o poder de generalização dos modelos gerados;

Os assuntos abordados neste relatório estão organizados da seguinte forma: a Seção II possui uma revisão dos trabalhos passados, a Seção III contém uma revisão da teoria abordada no trabalho, a Seção IV apresenta uma descrição dos processos realizados, a Seção V mostra os resultados e métricas obtidas, e por fim, a Seção VI apresenta conclusões alcançadas e futuras melhorias para o projeto.

II. Revisão da Literatura

Com a crescente procura por soluções que identifiquem os casos e causas de evasão no ensino superior, diversos trabalhos foram realizados utilizando a Mineração de Dados Educacional, propondo diferentes abordagens para o problema.

Alguns dos trabalhos correlatos definem a evasão estudantil de forma binária, *evadido* ou *não evadido*, e no geral, classificam estudantes como evadidos caso deixem de pagar as taxas do curso, comuniquem oficialmente sua saída do curso a Universidade ou sejam desligados por não alcançarem o desempenho acadêmico esperado [11], [12], [20]. Um dos trabalhos procura refinar a definição utilizando mais classes para definir evasão, como risco alto, médio e baixo [21], contudo não obteve bons resultados, dificilmente alcançando 50% de predições corretas, pois apesar de uma maior quantidade de classes parecer atrativo, possui uma grande desvantagem: a quantidade de dados por categoria diminui de forma a se tornar insuficiente para uma predição confiável [11].

Foi também analisado nos trabalhos passados quais atributos foram utilizados pelo algoritmo para treinar os modelos. Alguns dos autores afirmam que a maioria dos estudantes evade logo após o primeiro ano de curso [6], [11], [22], e utilizam dos dados de desempenho acadêmico até o primeiro ano da graduação na análise, como notas nas disciplinas, para que seja possível predizer evasão antes que a maioria dos alunos já tenha evadido. Existe uma discordância entre os trabalhos quanto ao uso de dados demográficos e sociais na análise, enquanto alguns afirmam que esses dados possuem pelo menos uma relevância parcial em relação evasão [20], [23], outros afirmam que seu uso produz resultados menos acurados e mais enviesados [12], [6], ou até mesmo discutem sobre uma visão positiva de que alguns alunos podem aumentar suas chances de sucesso apesar de possíveis dificuldades relacionadas a informações sociodemográficas [21].

Grande parte dos trabalhos analisados utilizam do desempenho acadêmico dos estudantes como informação mais relevante para a predição (e.g. [6], [11], [12], [22]), e um dos autores também afirma que o índice de rendimento acadêmico (*IRA*) do estudante, conhecido como *GPA* em universidades do exterior, possui alta correlação com evasão [22].

Em relação ao processamento da base da dados, alguns trabalhos utilizam de classificadores conhecidos, como as *Florestas Aleatórias*, *Árvores de Decisão*, *XGBoost*, entre outros, e comparam os resultados obtidos por esses algoritmos, obtendo até 83% de acurácia nos experimentos [6], [11]. Algoritmos como o XGBoost, baseados na ideia de "boosting", se tornaram referência para processamento de dados tabulares, como são os utilizados no problema da evasão estudantil. Um dos trabalhos tenta obter resultados próximos aos obtidos pelo XGBoost utilizando uma abordagem de aprendizado profundo, alcançando métricas semelhantes, com AUC de 0.771 [19].

A importância das métricas de precisão e sensibilidade é discutida em um dos trabalhos, já que é desejado em um preditor de evasão minimizar o número de falsos positivos (alunos que evadiram classificados como não evadidos) [6]. O SHAP, técnica estado da arte na interpretação de modelos, foi recentemente utilizado em trabalhos relacionados a predição de evasão [18], [19]. Os autores utilizam da técnica para compreender iterações e inferir relações entre os atributos e a evasão dos estudantes, e baseiam grande parte da análise nos dados extraídos pelo SHAP.

No contexto da Universidade de Brasília, [16] recentemente propôs um preditor de evasão que utiliza dos dados institucionais de estudantes do curso de Engenharia Mecatrônica da UnB. No trabalho, a evasão foi definida de forma binária, foram utilizados dados sociodemográficos e acadêmicos, até o primeiro ano, pelos alunos, testando três modelos diferentes de engenharia de atributos e utilizando seis diferentes algoritmos de classificação. Os melhores resultados encontrados utilizaram o modelo 3 e o algoritmo de *Random Forest*, com sensibilidade de 88,1% e acurácia de 86,6%.

Tendo em vista os trabalhos correlatos, para aprimorar a última abordagem explorada na Universidade de Brasília, neste trabalho é utilizada da técnica SHAP para extração de informações interpretáveis a respeito da predição dos modelos, para analisar impacto de atributos acadêmicos e não-acadêmicos na predição da evasão. Assim como em trabalhos correlatos, é utilizada a definição de evasão binária para a

classificação dos estudantes, e são utilizados apenas os dados acadêmicos do primeiro ano de curso do aluno na análise.

III. Conceitos Teóricos

A. Classificação Binária

O primeiro passo para o uso de uma abordagem de Mineração de Dados é identificar o tipo de problema que se deseja resolver: o problema da evasão pode ser visto como uma *Classificação Binária*, ou seja, onde se deseja classificar cada um dos elementos em dois grupos [24], para o propósito do trabalho, em evadido ou não-evadido. O processo utiliza um conjunto de atributos chamados *Variáveis de Entrada*, para estimar um atributo binário chamado *Variável de Saída*, esses dados são entregues aos algoritmos, para que identifiquem relações entre os dados de entrada e saída e sejam capazes de gerar sozinhos estimativas de saída para novas variáveis de entrada [24].

B. Métricas

As métricas são um conjunto de medidas utilizadas para analisar o resultado de um processo [24], e na Mineração de Dados são essenciais para medir a qualidade da predição dos algoritmos de classificação. Na classificação binária existem quatro possíveis resultados para uma predição: *Verdadeiro Positivo* (VP), onde a classe principal é prevista corretamente; *Falso Positivo* (FP), onde a classe principal é prevista incorretamente; *Verdadeiro Negativo* (VN), onde a classe secundária é prevista corretamente; e *Falso Negativo* (FN), onde a classe secundária é prevista incorretamente. A partir dos resultados obtidos para cada predição é possível calcular as seguintes métricas usadas neste trabalho.

A *Acurácia* é a métrica mais simples, que representa a proporção média de acertos das previsões.

$$A = \frac{VP + VN}{VP + FP + VN + FN}$$

A *Sensibilidade* é a métrica mais importante para este trabalho, pois busca responder a proporção da classe principal que foi corretamente predita, ou seja, a porcentagem de alunos evadidos que foram identificados.

$$S = \frac{VP}{VP + FN}$$

A *Precisão* é a proporção de identificações positivas que foram realmente previstas de forma correta, ou seja, a porcentagem dos alunos classificados como evadidos que evadiram de fato.

$$P = \frac{VP}{VP + FP}$$

A última métrica utilizada neste trabalho é a *F-Score*, dada pela média harmônica entre as métricas de *Sensibilidade* e *Precisão*. Para possuir um maior controle da proporção entre as duas métricas, foi utilizada uma versão mais geral da *F-Score*, a F_β, que utiliza um parâmetro β informando quantas vezes a *Sensibilidade* é considerada mais importante que a *Precisão*.

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot S}{(P \cdot \beta^2) + S}$$

C. SHAP

O SHAP[1] (SHapley Additive exPlanations) é um método capaz de extrair informações a respeito de modelos de aprendizado de máquina, que fundamenta-se no conceito de valores de Shapley presente na teoria dos jogos. Essa abordagem testa, de forma eficiente, permutações dos atributos, permitindo inferir a importância que a presença/ausência de um determinado atributo afetou a qualidade dos resultados. O SHAP possui uma implementação específica para modelos baseados em árvore, que permite que a análise das permutações de atributos seja realizada em tempo polinomial, e não exponencial.

As informações extraídas são locais, e dizem respeito a predição de um único elemento. Contudo, é possível unir informações locais e para produzir informações globais em relação ao modelo gerado. O SHAP também conta com ferramentas de visualização dos dados, que apresentam as informações extraídas de forma mais compreensível a um humano.

D. CatBoost

Existem diversos algoritmos de aprendizado de máquina utilizados em problemas de classificação, que usam diferentes abordagens para identificar os padrões nos dados. Uma dessas abordagens, é o *boosting*, que utiliza da combinação de um conjunto de modelos ruins (*weak learners*), para gerar um modelo bom (*strong learner*), assimilando os modelos ruins e ajustando erros cometidos de forma iterativa e construtiva.

O CatBoost é um algoritmo open source, mantido pela empresa Yandex[2], baseado na ideia de boosting em árvores de decisão. Esse algoritmo superou as métricas de diversos outros algoritmos similares em testes, como o XGBoost [25], foi utilizado em soluções campeãs de algumas competições de mineração de dados [26], e possui diversos processos auxiliares já implementados de forma nativa, como o processamento de atributos categóricos, uso de validação cruzada e ajuste de parâmetros, entre outros. Além disso, o CatBoost é integrado com a implementação para algoritmos baseados em árvore do SHAP, permitindo a extração de informações interpretáveis do modelo de forma eficiente.

IV. Metodologia

A metodologia utilizada neste trabalho é a KDD (Knowledge Discovery in Databases) [27], que define etapas sequenciais para o processo de extração de conhecimento em bases de dados. Nesta seção é apresentada as etapas realizadas no trabalho, primeiramente indicando as tecnologias utilizadas, detalhando e realizando a limpeza da base de dados, apresentando o processamento dos dados e os atributos gerados, e

[1] https://github.com/slundberg/shap

[2] https://yandex.com/

por fim, preparando a base e o algoritmo para a realização dos experimentos.

A. Tecnologias Utilizadas

Todos os processos foram implementados na linguagem de programação Python 3[3] devido a simplicidade de sua sintaxe, e a variedade de bibliotecas de manipulação, visualização de dados e aprendizado de máquina disponíveis (Pandas[4], Seaborn[5], CatBoost, SHAP, etc). Foi utilizado o Jupyter Notebook[6] para organizar e executar o código principal, facilitando a visualização e execução dos processos e resultados. Todas as funções de processamento da base foram armazenadas em arquivos separados, para aumentar a modularização e abstração do código principal. Todos os códigos utilizados e resultados obtidos neste trabalho estão disponíveis no GitHub[7].

B. Seleção e Extração dos Dados

A Universidade de Brasília possui diversas informações a cerca dos seus estudantes, como dados sociodemográficos e de desempenho nas disciplinas. A base de dados utilizada neste trabalho é referente aos quatro cursos ligados ao Departamento de Ciência da Computação da UnB: Ciência da Computação, Computação Licenciatura, Engenharia de Computação e Engenharia Mecatrônica. A base de dados extraída inicialmente possui 6745 alunos únicos, ingressos entre 1991 a 2021.

Dentre os atributos da base que possuem informações sobre o aluno, temos: um identificador único anonimizado, o *Sexo*, o *Tipo de Escola* frequentada, o *Curso*, a *Data de Nascimento*, o *Tipo de Cota*, o *CEP* da residência, a *Forma* e o *Período* de Ingresso, a *Forma* e o *Período* de Saída da Universidade. Cada linha da base representa o desempenho de um aluno em uma disciplina em um determinado período letivo, e dentre os atributos referentes ao desempenho na disciplina, temos: o *Período* em que a disciplina foi cursada, o *Código*, o *Nome* e a *Quantidade de Créditos* da disciplina, e por fim, e a *Menção* obtida pelo aluno.

C. Limpeza da Base

Para que os dados possam ser utilizados no treinamento dos algoritmos, eles devem passar por um processo de limpeza, removendo ruídos e tratando atributos com dados incorretos ou ausentes. Inicialmente são removidos os alunos ativos e que vieram a falecer, já que não podem ser classificados como *evadido* e *não-evadido*. Alguns dos *CEP*s da base de dados estão incorretos, mal formatados, ou informam o CEP da residência do aluno em outro estado, e para facilitar o uso desse atributo, os alunos com o *CEP* incorreto foram removidos da base.

Como são utilizados apenas os dados acadêmicos do primeiro ano da graduação, todas as disciplinas que foram realizadas após esse período são retiradas da base. As disciplinas de verão são consideradas como cursadas no primeiro semestre do mesmo ano. São também removidos os alunos que já evadiram antes do período analisado, pois não existe proposito em predizer um aluno que já evadiu, e alunos que vieram de transferência, por não ser possível ter acesso a suas disciplinas antes da transferência. Além disso, foram mantidos na base apenas os aluno ingressos entre 2004 e 2019. Essa decisão foi tomada pois os perfis dos alunos antigos, e dos alunos durante o ensino remoto na pandemia são diferentes, e para realizar uma análise não desbalanceada em relação ao atributo cotas, foi definido o ano de 2004 como limiar, pois foi o ano em que a politica de cotas entrou em funcionamento na Universidade.

Algumas correções específicas da base de dados são realizadas, como remoção de alunos *outliers*, correção de uma mesma disciplina com diferentes nomes, entre outros. Por fim, são realizados alguns ajustes finais, como a remoção de acentos, espaços e letras maiúsculas, e a renomeação alguns dados e atributos, com o propósito de facilitar a legibilidade e interpretabilidade dos resultados. Após a limpeza dos dados, a base possui 3042 alunos restantes, sem dados ausentes, pronta para a etapa de transformação.

D. Processamento da Base

O processamento da base consiste em utilizar de conhecimento do domínio para projetar e gerar atributos que sejam úteis para a identificação de evasão e compatíveis com os algoritmos utilizados [8]. Os atributos envolvidos no processo de mineração de dados podem ser divididos em dois grupos: a *Variável de Saída*, que se deseja prever, e que pode ser estimada através das *Variáveis de Entrada*.

Para o problema de evasão, a *Variável de Saída* é o atributo *Forma de Saída*, que informa como o aluno saiu da Universidade, situação que se deseja estimar. Neste trabalho é utilizada a definição de evasão binária, classificando um aluno como evadido caso o atributo *Forma de Saída* informe que o aluno não se formou no curso de origem. Desconsiderando alunos ativos e falecidos, os únicos alunos considerados como *não-evadido* são aqueles com o valor *"Concluído"* nesse atributo. Os demais alunos são classificados como *evadido*, incluindo alunos desligados por desempenho acadêmico, como reprovar a mesma disciplina três vezes, alunos que realizaram transferência do curso/habilitação, alunos desligados por abandono, entre outros.

As *Variáveis de Entrada* neste trabalho podem ser divididas em dois grupos de atributos: os *Atributos Acadêmicos* e os *Atributos Não-Acadêmicos*, descritos a seguir.

Os *Atributos Acadêmicos* são aqueles referentes ao desempenho acadêmico do aluno, como as menções obtidas nas disciplinas cursadas. Para transformar as informações presentes na base em atributos relevantes, é utilizada a abordagem que obteve os melhores resultados no trabalho [16]: Cada par disciplina/período será transformado em um atributo diferente, contendo uma escala numérica que representa a menção obtida pelo aluno naquela disciplina naquele período. Por exemplo, o atributo "*2_estruturas_de_dados*" representa a menção do aluno no segundo semestre na disciplina de Estrutura de

[3]https://www.python.org/
[4]https://pandas.pydata.org/
[5]https://seaborn.pydata.org/
[6]https://jupyter.org/
[7]https://github.com/gnramos/trj-academica

Dados. Para menções de aprovação (MM/CC, MS e SS), a escala consiste de um valor numérico de 1 até 3, e para menções de reprovação (MI, II e SR), de um valor numérico de -1 até -3. Menções nulas representam que o aluno não obteve uma menção de aprovação/reprovação, por exemplo, ao não cursar ou ao realizar trancamento da disciplina naquele período. Para que disciplinas que o aluno já obteve aprovação não apresentem valores nulos nos períodos seguintes, todos os valores dos atributos representam a menção máxima obtida até esse certo período.

Além dos atributos por disciplina/período, foram também gerados alguns atributos acadêmicos auxiliares. O primeiro deles é o IRA[8] (Índice de Rendimento Acadêmico), métrica utilizada pela UnB que se baseia em uma média ponderada das menções nas disciplinas com suas cargas horárias. É também gerado um atributo que informa a quantidade de créditos aprovados no período. São gerados duas versões desse atributos, uma absoluta, e outra relativa, em relação a quantidade de créditos cursados no período, isso pois alguns cursos possuem mais créditos por período na grade de disciplinas.

Os *Atributos Não-Acadêmicos* são aqueles referentes a informações sociodemográficas dos alunos e que não englobam o desempenho acadêmico. Os atributos categóricos (não numéricos), como o *Tipo de Cota*, ou a *Forma de Entrada*, tiveram 3 ou 4 dos seus valores mantidos, e o restante foram agrupados com o valor "outros", devido a baixa frequência na base. Além disso, duas versões da base foram geradas, uma mantendo esses atributos como categóricos, por recomendação da documentação do algoritmo a ser utilizado[9], e outra utilizando a técnica *one hot encoding*, para transforma-los em atributos binários, possibilitando uma análise posterior desses atributos.

O atributo *CEP* é utilizado para gerar o atributo *distância da residência à Universidade*. É utilizado o serviço *web* ViaCEP[10] para extrair informações dos valores de *CEP*, como o endereço, e posteriormente é utilizado o banco de dados geográfico do serviço de localização Google Maps[11] para calcular a distância do endereço até a UnB, utilizando meios rodoviários. Os alunos com mais de 50 quilômetros de distância da Universidade são considerados de fora do DF e são removidos da base. O atributo *Data de Nascimento* é transformado no atributo *Idade*, que informa a idade que o aluno tinha quando ingressou na Universidade. Os demais atributos, como *Sexo* (Masculino ou Feminino) e *Tipo de Escola* (Particular ou Pública), por já serem de natureza binária, foram apenas transformados em atributos binários.

E. Treinamento do Algoritmo

Esta etapa consiste no treinamento do algoritmo, que utiliza os dados recém processados para aprender a identificar e classificar alunos com risco de evasão.

[8]http://www.saa.unb.br/acompanhamento-academico/17-avaliacao-de-desempenho-academico

[9]https://catboost.ai/en/docs/features/categorical-features

[10]https://viacep.com.br/

[11]https://developers.google.com/maps

Para que seja possível produzir um modelo geral para todos os cursos, e modelos específicos para cada curso, são geradas diferentes versões da base de dados, uma com todos os dados, e outras separadas por curso. Além disso, cada base de dados é dividida em dois grupos: de treino, responsável pelo aprendizado do algoritmo e que possui 70% dos dados base, e de teste, responsável por testar a capacidade de generalização do algoritmo treinado, com 30% dos dados. A divisão é feita de forma estratificada, para que a proporção de elementos *evadido* e *não-evadidos* continue a mesma nos dois grupos. A distribuição da quantidade de alunos evadidos e não-evadidos em cada base pode ser visto na Tabela I. Para uma análise mais criteriosa em relação ao uso de atributos não-acadêmicos na predição, foram também gerados bases de dados somente com atributos acadêmicos, somente com atributos não-acadêmicos, e com os dois tipos de atributos, para uma posterior comparação dos resultados obtidos.

Tabela I: Distribuição da quantidade de alunos por classe em cada base.

	Evadido	Não-Evadido
Ciência da Computação	589	339
Computação Licenciatura	677	218
Eng. de Computação	345	191
Eng. Mecatrônica	359	224
Todos	1970	1072

O algoritmo escolhido para gerar os modelos neste trabalho é o CatBoost, por ter obtido bons resultados na literatura, por ser um algoritmo baseado na ideia de *boosting*, que obteve bons resultados no trabalho passado [16], e que possui diversos recursos já implementados para uso, e podem ser utilizados para gerar modelos menos enviesados. É utilizada a técnica *Validação Cruzada 5-fold* e a técnica *Randomized Search* para a otimização dos hiper-parâmetros do algoritmo, para que seja possível identificar a configuração que gera os melhores resultados. Foram testados diferentes valores para os parâmetros `learning_rate`, `depth` e `l2_leaf_reg`. O uso da validação cruzada e o ajuste desses parâmetros é essencial para o treinamento de um bom modelo, visto que em problemas complexos como o da evasão, os algoritmos tendem a gerar modelos enviesados com os dados de treino, sofrendo de sobre-ajuste e tendo dificuldade de generalizar a predição para alunos de teste.

A métrica otimizada pelo algoritmo é a F_β, utilizando o parâmetro $\beta = 1,2$, indicando que a métrica sensibilidade é $1,2$ vezes mais importante que a precisão para este trabalho. Esse valor é escolhido com base no entendimento do problema da evasão, onde existe uma preferência em identificar verdadeiros positivos, ou seja, alunos evadidos que foram corretamente identificados, ao custo de uma maior quantidade de falsos negativos, alunos não-evadidos identificados como evadidos. Isso se justifica uma vez que a prioridade é identificar alunos com risco de evasão para que eles possam

ser acompanhados e aconselhados, mesmo que ao custo de identificar de forma incorreta alunos que não o necessitam. Para otimizar o F_β, são extraídas as probabilidades para cada aluno, utilizando o método `predict_proba`, e é escolhido um limiar de classificação que maximiza essa métrica. Essa abordagem permite a alteração do limiar quando necessário.

Após o treinamento dos algoritmos, eles são utilizados para classificar os dados de teste, e são extraídas as métricas comparando os resultados obtidos com os resultados esperados. São geradas quatro métricas, a *Acurácia*, a *Precisão*, a *Sensibilidade*, e por último, a $F_{1,2}$-*Score*. O CatBoost é integrado com o SHAP, sendo então possível extrair informações locais e globais em relação as predições do modelo gerado. Com essas informações, é possível identificar e visualizar quais atributos possuem maior impacto na evasão de um estudante em específico, ou dos estudantes no geral.

V. Experimentos e Resultados

Foram geradas diferentes versões da base de dados no processamento. Primeiramente, a base foi dividida em cinco outras bases: uma contendo todos os cursos juntos, e outras quatro contendo apenas alunos de cada curso em específico. Depois, foram geradas 3 versões para cada uma dessas bases: a primeira apenas com os atributos acadêmicos, a segunda apenas com os atributos não-acadêmicos, e a última, com todos os atributos. O algoritmo é utilizado para gerar modelos para cada uma dessas bases, de onde são extraídas as quatro métricas e a interpretação da predição identificada pelo SHAP. Os valores das quatro métricas, para cada um dos cursos, podem ser vistos nas Tabelas II, III e IV, para a base apenas com atributos acadêmicos, não-acadêmicos, e com todos os atributos, respectivamente.

Tabela II: Resultados utilizando apenas atributos acadêmicos.

	Acurácia	Precisão	Sensibilidade	$F_{1.2}Beta$
Ciência da Computação	76,53%	83,61%	78,06%	80,24%
Computação Licenciatura	80,19%	80,57%	97,44%	89,73%
Eng. de Computação	75,98%	79,53%	85,59%	83,00%
Eng. Mecatrônica	75,82%	74,50%	84,09%	79,87%
Todos	79,73%	81,71%	88,81%	85,75%

Tabela III: Resultados utilizando apenas atributos não-acadêmicos.

	Acurácia	Precisão	Sensibilidade	$F_{1.2}Beta$
Ciência da Computação	62,02%	63,50%	95,08%	78,98%
Computação Licenciatura	72,40%	75,77%	93,36%	85,25%
Eng. de Computação	59,88%	64,89%	81,73%	73,87%
Eng. Mecatrônica	58,77%	57,42%	80,91%	69,29%
Todos	65,10%	66,06%	94,89%	80,49%

Para a primeira base de dados, que possui apenas atributos acadêmicos, é possível perceber que as métricas alcançadas são altas, alcançando F_β acima dos 80%. Contudo na base de dados contendo apenas atributos não-acadêmicos, o modelo não é capaz de alcançar as mesmas métricas, dificilmente obtendo valores altos de precisão. É importante destacar o curso de Computação Licenciatura, que obteve boas métricas nos dois modelos, indicando uma maior facilidade em prever evasão para estudantes desse curso, tanto com atributos acadêmicos quanto com não-acadêmicos. Além disso, a base de dados com os alunos de todos os cursos obteve métricas altas, possivelmente pela maior quantidade de dados disponíveis e do uso de diferentes alunos para o treino do modelo, uma evidência da capacidade de generalização da abordagem de EDM em dados não restritos a um único curso.

Tabela IV: Resultados utilizando todos os atributos.

	Acurácia	Precisão	Sensibilidade	$F_{1.2}Beta$
Ciência da Computação	79,21%	83,24%	84,19%	83,79%
Computação Licenciatura	81,41%	85,92%	90,15%	88,36%
Eng. de Computação	72,67%	74,59%	87,50%	81,70%
Eng. Mecatrônica	78,05%	73,33%	91,67%	83,15%
Todos	79,85%	80,42%	91,03%	86,36%

A base de dados que utiliza de todos os atributos possui métricas melhores que as duas bases utilizadas anteriormente, alcançando cerca de 90% de sensibilidade na predição. Isso indica que alguns atributos não-acadêmicos foram importantes para a predição. É possível perceber que as métricas são menores para os cursos de engenharia, indicando uma maior dificuldade em predizer evasão de alunos desses cursos, quando comparados com o curso de licenciatura por exemplo, que alcançou 90,15% de sensibilidade e 85,92% de precisão.

Esses são resultados esperados comparando-se com a literatura: os atributos acadêmicos são os mais importantes para a predição de evasão, e são suficientes para gerar modelos com boas métricas, contudo, alguns atributos não-acadêmicos também possuem um impacto relevante na evasão de estudantes. Para identificar quais atributos foram os mais importantes para predição, e realizar uma análise global em relação a causas de evasão dos estudantes, foram extraídos os *SHAP Importances*, primeiramente para os alunos de todos os cursos utilizando todos os atributos, que pode ser visto na Figura 1.

Os atributos acadêmicos de quantidade relativa e absoluta de créditos aprovados do segundo semestre foram os atributos mais importantes para a predição, é possível perceber que valores baixos desses atributos (em azul no gráfico) possuem um impacto positivo no modelo, tendendo a classificação para verdadeiro, ou seja, evadido, e o contrário acontece para valores altos desses atributos (em vermelho no gráfico), que tendem a classificação para não-evadido. Outros atributos acadêmicos também aparecem entre os mais relevantes, como o IRA, o desempenho do aluno no primeiro semestre em Algoritmos e Programação de Computadores, e no segundo semestre em Estrutura de Dados. Essas são as disciplinas iniciais de programação para os cursos de computação, e

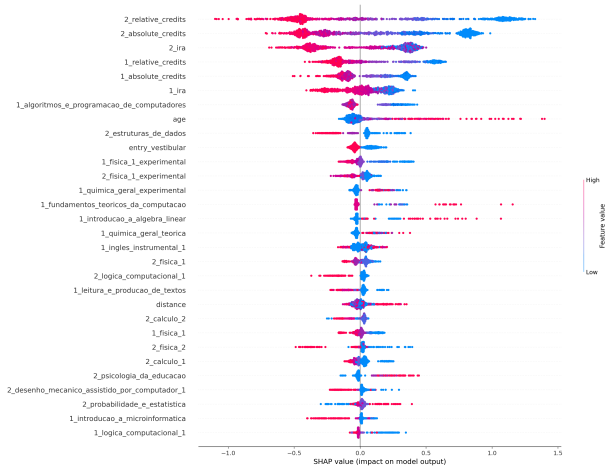

Figura 1: Importância global dos atributos para base com todos os cursos.

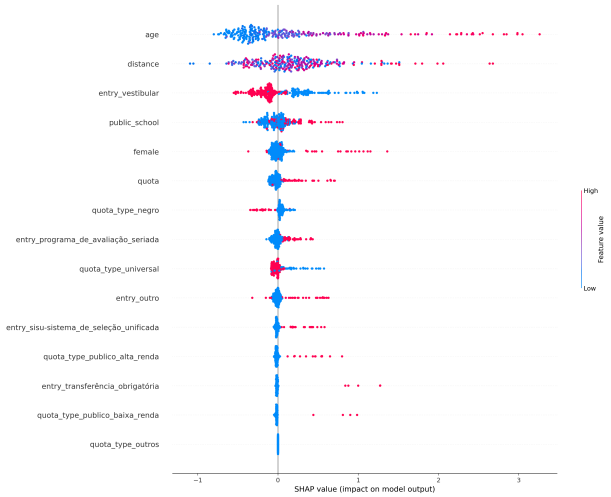

Figura 2: Importância global dos atributos não-acadêmicos do curso de Computação Licenciatura.

esses resultados indicam que um mal desempenho nessas disciplinas está relacionado com a evasão do aluno. A presença de atributos acadêmicos no topo do gráfico confirma sua maior importância para predição de evasão quando comparados com os atributos não-acadêmicos.

Contudo, dois atributos não-acadêmicos aparecem entre os atributos mais importantes: a idade do aluno, e se a forma de entrada dele foi pelo vestibular tradicional. Para o atributo idade, o gráfico mostra uma dispersão de pontos vermelhos (alunos mais velhos) para valores positivos de impacto, indicando que estudantes mais velhos que entram na universidade possuem maiores chances de evadir. O segundo atributo indica que alunos que entraram na universidade pelo vestibular tradicional da UnB possuem menos chances de evadir.

Para uma análise mais específica nos atributos não-acadêmicos, foi gerado o gráfico de *SHAP Importances* da base de dados que utiliza apenas atributos não-acadêmicos, apenas com os dados do curso de Computação Licenciatura, por ser o único com métricas boas utilizando apenas atributos não-acadêmicos. O gráfico pode ser visto na Figura 2.

Os quatro atributos mais importantes nessa análise, foram a idade, a distância da residência até a UnB, se o aluno entrou pelo vestibular, e se ele veio de escola pública no ensino médio. A distribuição do atributo de distância da residência é complexa de ser analisada de forma individual, mas é possível perceber que alguns alunos que moram longe da universidade estão mais espalhados para direita, indicando que possuem uma maior chance de evadir. O atributo de escola pública também aparece, e possui os pontos vermelhos (alunos de escola pública) espalhados para a direita, também indicando maiores chances de evadir.

Além das análises globais dos atributos, o SHAP é capaz de realizar análises locais, ou seja, identificar os valores e atributos que impactaram na predição do modelo para aquele aluno. Dois exemplos dessa análise podem ser vistos nas Figuras 3 e 4.

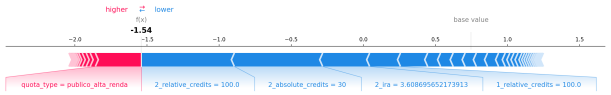

Figura 3: Exemplo de explicação local.

Os atributos em vermelho, são atributos que contribuem de forma positiva para a predição (para classificar um aluno como evadido), já os atributos em azul contribuem de forma negativa (para classificar o aluno como não-evadido). Na primeira figura, é possível perceber que, apesar do atributo do tipo de cota contribuir para a evasão do aluno, por informar que o aluno é cotista de escola pública de alta renda, outros atributos contribuíram de para a não-evasão do aluno, pegar 30 créditos e ser aprovado em 100% deles no segundo semestre, com IRA 3,60. Essas foram as informações de maior relevância para o modelo predizer esse aluno, e o somatório negativo das contribuições indica que esse aluno tem baixas chances de evadir.

Na segunda figura, o aluno possui um IRA alto, de 4,18 no segundo semestre, e isso gerou uma certa contribuição para uma classificação de não-evadido. Contudo, atributos como idade elevada, quantidade baixa de créditos aprovados no segundo semestre, e entrada na Universidade pelo ENEM, contribuíram para uma classificação de evadido, principalmente o atributo idade. Como o somatório das contribuições foi positivo, esse aluno possui chances altas de evadir, de acordo com o modelo. As informações obtidas nessas análises locais contribuem para uma compreensão mais transparente do funcionamento do algoritmo para a predição de evasão, permitindo uma análise mais específica e individual para a condição de cada aluno. Todas as métricas e gráficos obtidos nos resultados, que não foram apresentados no relatório, estão

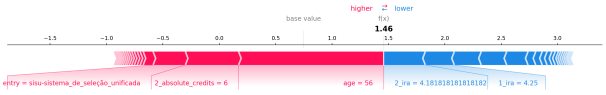

Figura 4: Exemplo de explicação local.

disponíveis no GitHub do projeto[12].

VI. Conclusões

A existência de uma ferramenta que ofereça suporte para a gestão de universidades, auxiliando com o problema da evasão estudantil, é de grande interesse dessas instituições. Neste trabalho, foi utilizado um novo método para melhorar a interpretabilidade de modelos de predição de evasão, o SHAP, bem como a análise de 2 novos atributos, a cota e a idade do aluno, utilizando uma base de dados com quatro cursos ligados a computação.

Foram gerados diferentes modelos, dividindo os alunos por curso, e os atributos em acadêmicos e não acadêmicos. A base de dados com as melhores métricas foi a base que utiliza de ambos os atributos acadêmicos e não acadêmicos, alcançando cerca de 90% de sensibilidade. A análise com todos os cursos juntos obteve precisão de 80,42%, sensibilidade de 91,03% e F-Beta de 86,36%, comprovando a capacidade de generalização da abordagem de EDM em dados não restritos a um único curso. A análise também foi feita para os cursos de forma individual, e para o curso de Computação Licenciatura em específico, a métrica F-beta chegou a 88,36%.

Os atributos acadêmicos foram os mais relevantes para o problema da evasão, quando comparados com os atributos não acadêmicos. As informações de quantidade de créditos aprovados, IRA, e desempenho nas disciplinas iniciais de computação foram as mais relevantes para predizer um aluno como evadido ou não. As bases de dados que utilizam apenas de dados acadêmicos geraram modelos com mais de 80% na métrica F-Beta. Apesar dos modelos treinados apenas com dados não-acadêmicos não alcançarem métricas tão boas, os atributos idade, forma de entrada, distância e tipo de escola também foram relevantes na predição, indicando que alunos mais velhos, que não entraram na UnB pelo vestibular tradicional, que moram mais longe da universidade, e que vieram de escola pública possuem mais chances de evadir. Informações como o sexo do aluno não foram de grande relevância para os modelos.

O SHAP foi utilizado para gerar análises globais e locais utilizadas no trabalho. Esse método foi capaz de extrair informações transparentes a respeito do processo de classificação de um modelo complexo de boosting em árvore de decisão (CatBoost), gerando informações relevantes para o problema de evasão, e que são compreensíveis por humanos. A análise individual por aluno é uma poderosa ferramenta para gestão dos cursos de graduação, visto que possibilita identificar de forma automática e quantitativa as informações mais importantes para a evasão de um aluno específico.

Vale lembrar que as informações extraídas em relação a causas de evasão não representam a realidade diretamente, pois o SHAP apenas interpreta e compreende as decisões tomadas pelo algoritmo utilizado, essa conexão depende da capacidade do modelo gerado de representar a realidade, que é assegurado a partir de boas métricas obtidas em um processo de Mineração de Dados bem executado. Além disso, os fatores identificados como relevantes na análise dos modelos podem ser apenas variáveis relacionadas, e não as causas em si. Essas informações servem de indícios que o aluno vai ou não evadir, mas talvez não sejam a real causa do problema, podendo ser algo externo, ou até mesmo algo não relacionado aos dados disponíveis. Por isso, a Mineração de Dados deve ser utilizada apenas como uma ferramenta para auxiliar as organizações a lidar com o problema, identificando alunos em risco de evasão e fatores relacionados, e não como solução definitiva e autônoma, pois não é possível para uma abordagem de aprendizado de máquina entender com precisão um problema e suas causas caso as respostas não estejam contidas nos dados utilizados, como é o caso problema de evasão estudantil.

Existem diversas formas de continuar este trabalho, a predição de evasão é um problema complexo, vasto e pouco explorado até o momento. Este trabalho se restringiu aos dados até 2019, pois o perfil do aluno no ensino remoto durante a pandemia é diferente, sendo possível realizar uma análise do impacto da ensino remoto na evasão. Apesar da base de dados utilizada possuir mais de um curso de graduação, foram todos cursos ligados a computação, com disciplinas iniciais semelhantes, tornando a análise com cursos de áreas do conhecimento diferentes outro caminho para um trabalho futuro. A análise dos atributos não acadêmicos só pôde ser realizada para o curso de Computação Licenciatura, pois foi o único curso com métricas boas utilizando apenas atributos não-acadêmicos, dito isso, é possível realizar uma análise diferente para tentar entender a importância desses atributos em outros cursos. É possível também construir uma análise em torno do atributo idade, que se mostrou uma informação relevante para a evasão de alunos.

[12]https://github.com/gnramos/trj-academica

Agradecimentos

O presente trabalho foi realizado com apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Referências

[1] F. S. Santos and N. de Almeida Filho, *A quarta missão da universidade: internacionalização universitária na sociedade do conhecimento*. Imprensa da Universidade de Coimbra/Coimbra University Press, 2012.

[2] M. Lobo, "Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. 2012," *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, vol. 25, 2012.

[3] M. M. Braga, M. d. C. L. Peixoto, and T. F. Bogutchi, "A evasÃo no ensino superior brasileiro: O caso da ufmg," *Avaliação: Revista da Avaliação da Educação Superior*, vol. 8, no. 3, 1 1. [Online]. Available: http://periodicos.uniso.br/ojs/index.php/avaliacao/article/view/1237

[4] L. Hagedorn, "How to define retention: A new look at an old problem." in *Transfer and Retention of Urban Community College Students (TRUCCS)*, 2006.

[5] U. de Brasília, "Plano de desenvolvimento institucional 2018-2022 da universidade de brasília," http://www.planejamentodpo.unb.br/index.php?option=com_content&view=article&id=20&Itemid=791, acessado: 19-04-2021.

[6] L. S. Aulck, D. Nambi, N. Velagapudi, J. Blumenstock, and J. D. West, "Mining university registrar records to predict first-year undergraduate attrition," in *EDM*, 2019.

[7] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 40, pp. 601 – 618, 12 2010.

[8] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, "Appendix a - theoretical foundations," in *Data Mining (Fourth Edition)*. Morgan Kaufmann, 2017, pp. 533–552. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128042915000234

[9] R. GOLDSCHMIDT and E. PASSOS, *Data mining: um guia Prático*. Elsevier Editora, 2005. [Online]. Available: https://books.google.com.br/books?id=JJYHNrREwyEC

[10] T. M. Mitchell *et al.*, "Machine learning. 1997," *Burr Ridge, IL: McGraw Hill*, vol. 45, no. 37, pp. 870–877, 1997.

[11] A.-S. Hoffait and M. Schyns, "Early detection of university students with potential difficulties," *Decision Support Systems*, vol. 101, pp. 1–11, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923617300817

[12] F. Jiménez, A. Paoletti, G. Sánchez, and G. Sciavicco, "Predicting the risk of academic dropout with temporal multi-objective optimization," *IEEE Transactions on Learning Technologies*, vol. 12, no. 2, pp. 225–236, 2019.

[13] L. R. S. d. ASSIS, "Perfil de evasão no ensino superior brasileiro: uma abordagem de mineração de dados," *Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação)*, p. 134, 2018. [Online]. Available: https://repositorio.unb.br/handle/10482/32139

[14] G. F. SILVA, "Análise preditiva do desempenho acadêmico de alunos de graduação da unb utilizando mineração de dados," *Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação)*, p. 52, 2017. [Online]. Available: https://bdm.unb.br/handle/10483/17910

[15] W. G. M. GRAMACHO, "Algoritmos de mineração de dados para análise de evasão na graduação da universidade de brasília," *Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação)*, p. 55, 2020. [Online]. Available: https://bdm.unb.br/handle/10483/24527

[16] T. d. S. FERNANDES, "Algoritmos de mineração de dados para análise de evasão na graduação da universidade de brasília," *27º CONGRESSO DE INICIAÇÃO CIENTÍFICA DA UNB E 18º DO DF*, 2021. [Online]. Available: https://conferencias.unb.br/index.php/iniciacaocientifica/27CICUnB18df/paper/view/38944

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[18] S. Dass, K. Gary, and J. Cunningham, "Predicting student dropout in self-paced mooc course using random forest model," *Information*, vol. 12, no. 11, 2021. [Online]. Available: https://www.mdpi.com/2078-2489/12/11/476

[19] M. Baranyi, M. Nagy, and R. Molontay, "Interpretable deep learning for university dropout prediction," in *SIGITE '20: Proceedings of the 21st Annual Conference on Information Technology Education*, ser. SIGITE '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 13–19. [Online]. Available: https://doi.org/10.1145/3368308.3415382

[20] A. O. E. and D. C., "Roads to success in the belgian french community's higher education system: Predictors of dropout and degree completion at the université libre de bruxelles." *Res High Educ*, vol. 54, p. 693–723, 2013. [Online]. Available: https://link.springer.com/article/10.1007%2Fs11162-013-9290-y

[21] J. Vandamme, N. Meskens, and J. Superby, "Predicting academic performance by data mining methods," *Education Economics*, vol. 15, no. 4, pp. 405–419, 2007. [Online]. Available: https://doi.org/10.1080/09645290701409939

[22] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decision Support Systems*, vol. 49, no. 4, pp. 498–506, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923610001041

[23] F. Araque, C. Roldán, and A. Salguero, "Factors influencing university drop out rates," *Computers & Education*, vol. 53, no. 3, pp. 563–574, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360131509000815

[24] J. Han, M. Kamber, and J. Pei, "8 - classification: Basic concepts," in *Data Mining (Third Edition)*, 3rd ed., ser. The Morgan Kaufmann Series in Data Management Systems, J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 327–391. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123814791000083

[25] A. V. Dorogush, A. Gulin, G. Gusev, N. Kazeev, L. O. Prokhorenkova, and A. Vorobev, "Fighting biases with dynamic boosting," *CoRR*, vol. abs/1706.09516, 2017. [Online]. Available: http://arxiv.org/abs/1706.09516

[26] A. Gulin, I. Kuralenok, and D. Pavlov, "Winning the transfer learning track of yahoo!'s learning to rank challenge with yetirank," in *Proceedings of the Learning to Rank Challenge*, ser. Proceedings of Machine Learning Research, O. Chapelle, Y. Chang, and T.-Y. Liu, Eds., vol. 14. Haifa, Israel: PMLR, 25 Jun 2011, pp. 63–76. [Online]. Available: https://proceedings.mlr.press/v14/gulin11a.html

[27] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, pp. 37–37, 1996.