Identificação de Atributos Relevantes na Evasão no Ensino Superior Público Brasileiro

Bruno Claudino Pereira de Brito, Gabriel Alves de Albuquerque Júnior, Rafael Ferreira Leite de Mello

Programa de Pós-Graduação em Informática Aplicada (PPGIA)
Universidade Federal Rural de Pernambuco (UFRPE)

Abstract. *The dropout is a frequent problem in Brazilian universities. In addition to the loss of the student in the academic environment, it may cause inefficient university expenses since the university budget is unable to achieve its goal of training the student. Data mining techniques with the aid of algorithms can assist in an identification of the possible evasion student. Several algorithms can be used, including Random Forest. This work aims to identify, with the help of this algorithm, the relevance of personal characteristics in students in association with demographic characteristics in the dropout process. The work was carried out on data from a federal public university in conjunction with data from the Brazilian demographic census. The results achieved, evaluated between the two databases, allow to verify which characteristics are associated with the dropout students at this university.*

Resumo. A e*vasão é um problema frequente nas universidades brasileiras. Além da perda do aluno no meio acadêmico, pode provocar gastos universitários ineficientes uma vez que orçamento universitário não consegue atingir seu objetivo de formar o aluno. Técnicas de mineração de dados com auxílio de algoritmos podem auxiliar na identificação do possível aluno em evasão. Diversos algoritmos podem ser utilizados entre eles o Random Forest. Esse trabalho visa identificar, com a ajuda desse algoritmo, a relevância de características pessoais nos alunos em associação com características demográficas no processo de evasão. O trabalho foi realizado nos dados de uma universidade pública federal em conjunto com dados do censo demográfico brasileiro. Os resultados alcançados, avaliados entre as duas bases de dados, permitem verificar quais características estão associadas aos alunos evadidos nessa universidade.*

1. INTRODUÇÃO

A evasão é um problema recorrente nas universidades no mundo, tornando-se um dos maiores desafios pelas quais elas enfrentam [Litoiu 2018]. No Brasil, diversas ações foram realizadas com o objetivo de democratizar e ampliar o acesso às universidades, como a oferta de novas vagas, programas de financiamento e sistemas de cotas. Contudo, não houve uma quantidade proporcional de ações voltadas para a permanência do estudante nessas instituições. A facilitação no acesso aliada à insuficiência de políticas de assistência estudantil levaram a um aumento da evasão proporcionalmente maior que o aumento das vagas ofertadas. Assim, a taxa da evasão no ensino superior brasileiro vem aumentando ao longo dos últimos anos.

No caso das universidades públicas, dadas as restrições cada vez maiores de recursos, é imprescindível que os recursos aplicados nas políticas de assistência sejam

DOI: 10.5753/cbie.sbie.2020.1032

efetivos. Ou seja, que o aluno contemplado por essa assistência seja de fato aquele aluno que terá maior risco de evasão. Nesse contexto, o investimento em programas de assistência estudantil se torna crítico, uma vez que pode ser capaz, quando a evasão é evitada, de diminuir consideravelmente o custo por aluno formado. Criar mecanismos de identificação do provável aluno a ser evadido constitui-se, também, de uma forma eficiente de gasto universitário. Dessa forma, técnicas de mineração de dados podem ajudar na identificação de características ou atributos que levam a uma maior probabilidade do estudante se evadir. Especialmente no caso das universidades públicas, esse tipo de análise pode auxiliar no sentido de direcionar seus recursos para esses estudantes com maior risco de evasão.

Diante disso, o presente trabalho tem como objetivo identificar atributos relevantes na evasão escolar usando técnicas de mineração de dados e aprendizagem de máquina. Serão utilizados dados de uma universidade em conjunto com dados demográficos abertos. Conceitos como Mineração de Dados, Classificadores e aprendizagem de máquina servem de base teórica para a pesquisa em questão.

Mineração de Dados é parte de uma ciência maior chamada *Knowledge Discovery in Database* [Goldschmidt et al. 2005]. KDD atua em grandes quantidades de dados tendo como objetivo encontrar conhecimento válido e útil que possa ser aplicado à solução de problemas. KDD pode ser tradicionalmente dividido em 5 passos: seleção, pré-processamento, transformação, mineração de dados e interpretação/avaliação [Goldschmidt et al. 2005].

Na etapa de Mineração de Dados podem ser aplicados diversos algoritmos e ferramentas para identificação de padrões. Entre os algoritmos utilizados pela pesquisa estão os Classificadores. Por conseguinte, Classificadores são um grupo de algoritmos cuja tarefa se dá em classificar um conjunto de dados através da construção de um modelo. Esse modelo rotula o conjunto de dados em uma ou mais classes [Tan 2009]. Dentre os Classificadores mais conhecidos está o *Random Forest*. Adicionalmente, *Random Forest* pode ser considerado um dos melhores algoritmos classificadores [Delgado et al. 2014] e o uso desse classificador será discutido nas sessões posteriores.

A presente pesquisa utilizou o *Random Forest* para extração de informações em um grande volume de dados acadêmicos de discentes de uma instituição universitária em conjunto com dados demográficos abertos, tais como dados relativos à água e esgoto das cidades, taxa de fecundidade, IDH, entre outros.

2. TRABALHOS RELACIONADOS

O estudo da evasão em ambientes educacionais não é recente. O fenômeno de evasão já é tema constante nos últimos anos nos anais da SBIE (Simpósio Brasileiro de Informática na Educação). Vários trabalhos publicados no SBIE possuem a palavra evasão no próprio título do artigo, o que ressalta a importância do tema para a comunidade científica.

Quando se trata de artigos que utilizaram o *Random Forest,* os trabalhos de [Queiroga et al. 2017] e [Santos et al. 2017], únicos encontrados na pesquisa exploratória, possuem o termo *Random Forest* em seu conteúdo. Em [Queiroga et al. 2017], a pesquisa foi realizada em 4 cursos técnicos à distância, em que teve como objetivo predizer a evasão baseado em contagens de interações no ambiente virtual. A pesquisa teve dois cenários para treinamento e teste. O primeiro cenário de predição se deu apenas em um curso, sendo realizado treinamento e teste nesse único curso. O segundo cenário consistia de treinamento com 3 cursos e teste sendo aplicado aos demais cursos restantes. Na pesquisa citada, foram utilizados diversos outros algoritmos de mineração de dados; contudo, o algoritmo *Random Forest* obteve os melhores índices de acurácia na predição de evasão em ambos os cenários.

Em [Santos et al. 2017] a pesquisa se deu na construção e avaliação de um protótipo de avaliação de alunos em ambiente virtual. Esse continha, entre outras funcionalidades, a

classificação do aluno baseado em interações no ambiente virtual a fim de identificar um possível aluno com possibilidades de evasão. A ferramenta proposta utilizava o *Random Forest* como algoritmo classificador e obteve bons resultados quando avaliado na pesquisa.

Nesse contexto, vale destacar como trabalho relacionado o de [Pascoal et al. 2016] em que foi analisado a importância de atributos que influenciam na evasão. Esse trabalho tem relação ao presente trabalho na medida em que foram incluídos dados socioeconômicos para análise. No artigo mencionado, foram incluídos dados acadêmicos e sociais dos alunos, submetidos ao algoritmo de classificação *Naive Bayes,* a fim de obter uma ordem de importância dos atributos para determinação da evasão escolar. Obteve um índice de acurácia grande, contudo, não compreendeu grandes bases de dados, limitando-se apenas a um curso de graduação.

Podemos destacar também o trabalho de [Carrano et al. 2019], em que foi realizado um trabalho de predição de evasão em grande base de dados, com diversos cursos, incluindo dados sociais dos alunos. Ao passo que tentou prever a quantidade de evasão do ano seguinte, também tentou identificar quais atributos tinham mais relevância nessa previsão.

Pode-se constatar pelas referências acima citadas que o uso do *Random Forest* em ambiente educacional obteve sucesso. A presente pesquisa pretende expandir o uso do algoritmo utilizando-o em uma base dados de uma universidade brasileira com mais de 56 cursos de graduação, entre dados de cursos presenciais e a distância, junto com dados demográficos do censo brasileiro de 2010. Diferentemente dos trabalhos anteriores, esse artigo não visa apenas a classificação, mas também obter as características mais relevantes na identificação da evasão.

3. METODOLOGIA

Como citado nesse artigo, a pesquisa se deu em duas bases de dados: uma, acadêmica, onde contém informações de formação dos alunos, e outra, demográfica, onde contém informações demográficas e socioeconômicas sobre as cidades do Brasil[1]. A base demográfica contém dados do último censo brasileiro realizado em 2010. As bases foram unidas de forma a relacionar cada aluno a sua cidade, com seus respectivos indicadores demográficos e socioeconômicos.

Os dados acadêmicos foram retirados do sistema acadêmico de uma universidade brasileira com 59 cursos de graduação, dentre cursos presenciais e a distância. Foram selecionados os alunos cuja última situação acadêmica tenha sido Formado a partir de 2010. Ou seja, alguns alunos podem ter se matriculado antes de 2010, entretanto, a consulta se deu em alunos cuja conclusão no curso tenha sido dada a partir de 2010. No total, foram retornados 51175 registros de alunos. Assim como, visando atingir o anonimato dos dados acadêmicos, de forma que não se possa identificar os alunos, foram omitidos na consulta seus dados pessoais, tais como nome, endereço, CPF, deixando apenas sua situação acadêmica, cidade natal, gênero, estado civil e o ano do seu ingresso.

As situações acadêmicas encontradas são citadas a seguir: *matriculado, matrícula vínculo, trancamento, formado, complementação curricular, desligamento, transferência externa, desvinculado, transferência interna, trancamento de semestre anterior, titulado, matriculado sub judice, desistência, reintegração, intercâmbio, excluído, integralizado, mobilidade estudantil*. Contudo, o trabalho tem foco em evasão acadêmica. Então, essas situações foram agrupadas em apenas dois grupos: *Evadido* e *Não Evadido*. Foram considerados alunos Evadidos todos aqueles que não concluíram o curso com sucesso [INEP 1998]. Situação de *desistência*, *desligamento* e *excluído* foram agrupados como Evadidos. Para esse grupo foram encontrados 19693 registros.

1 http://www.atlasbrasil.org.br/2013/pt/consulta/.

Ao passo que para o grupo do Não Evadido, foram considerados os demais alunos das situações restantes, citadas a seguir: *matriculado*, *matrícula vínculo, trancamento, formado, complementação curricular, transferência externa, transferência interna, trancamento de semestre anterior, titulado, matriculado sub judice, reintegração, intercâmbio, integralizado, mobilidade estudantil*. Nesse grupo foram categorizados 31482 registros. Essas situações consideram alunos que ainda possuem vínculo com a universidade estudada ou que possuem o curso concluído com sucesso, por isto, considerados Não Evadidos [INEP 1998].

Os dados demográficos se referem ao site do Atlas do Desenvolvimento Humano no Brasil e traz índices como o Índice de Desenvolvimento Humano Municipal (IDHM) e outros mais de 200 indicadores de demografia, educação, renda, trabalho, habitação e vulnerabilidade para as cidades brasileiras[2]. Essa base foi colocada em uma tabela cuja coluna principal é o nome da cidade, em que determina semanticamente as demais colunas contendo os respectivos índices. Uma vez os dados estando no mesmo esquema, fez-se uma consulta entre ambos os dados com a junção se dando na cidade natal do aluno e sua unidade federativa com a cidade do atlas brasileiro e sua unidade federativa. Gerando-se, consequentemente, um conjunto de dados cuja tupla contém informações do grupo acadêmico, evadido ou não evadido, com seus respectivos índices demográficos. Vale ressaltar, mais uma vez, que os índices são relativos aos dados do último censo brasileiro realizado em 2010.

Nesse ponto, os dados foram divididos por área de conhecimento. Ou seja, foram agrupados os cursos pelas suas áreas, quer sejam áreas de ciências humanas, ciências sociais, exatas e da terra, etc[3]. Esse agrupamento se fez necessário para dar maior granularidade à pesquisa, dando maior especificidade aos resultados obtidos. Dessa forma, foram formados 6 grupos de dados representando a área de conhecimento dos cursos afins. Estas áreas estão citadas a seguir: ciências sociais, ciências agrárias, ciências biológicas, engenharias, ciências humanas e exatas e da terra.

Uma vez os dados agrupados e disponíveis para uso, esses foram tratados para a aplicação do *Random Forest*, que inclui tratamentos de valores de string para números e remoção de valores em branco e colunas duplicadas. Um destaque pode ser feito nesse momento em relação à motivação do uso do *Random Forest*. Esse algoritmo é bastante popular e, para alguns autores [Tibshirani 2008], ele é considerado o mais acurado, rápido para grandes bases de dados, assim como funciona notavelmente com poucos ajustes necessários. Esse classificador pode ser considerado um conjunto de classificadores, chamados *árvores de decisão*. Desse conjunto de árvores advêm o nome “floresta” [Han et al. 2011]. No entanto, definir a importância de uma característica é complexa na medida em que uma característica pode perder importância quando interagida com outras características. Para isso, o algoritmo estima a importância de uma variável verificando o quanto de erros de predição aumentam quando um valor para aquela variável é permutada enquanto os outros dados permanecem imutáveis [Liaw 2002]. Assim, os cálculos obtidos são transformados em árvores na medida em que o *Random Forest* é construído [Liaw 2002]. Durante a classificação, cada árvore desse conjunto vota e, então, a classe mais votada é retornada [Han et al. 2011]. Sendo, desse modo, definido o objetivo do algoritmo predizer qual classe, isto é, grupo acadêmico, as tuplas deveriam ser: Evadido ou Não evadido.

Como a tarefa de classificação é baseada em construir um modelo. Esse modelo, portanto, precisa ser treinado, ou seja, o algoritmo precisará trabalhar com classes já conhecidas, gerando um modelo baseado em dados de treinamento. Tendo o modelo construído, esse será usado para dados desconhecidos, isto é, dados de teste [Han et al. 2011].

2 *http://www.atlasbrasil.org.br/2013/pt/consulta/.*

3 https://www.capes.gov.br/avaliacao/documentos-de-apoio/91-conteudo-estatico/avaliacao-capes/6831-tabela-de-areas-de-conhecimentoavaliacao

Na prática, os dados devem ser divididos de forma que o modelo tenha dados de treino suficientes para gerar um modelo eficiente. Por não haver uma divisão formalizada para mineração de dados entre dados de treinamento e teste, utilizou-se, para cada grupo de dados, a divisão estratificada, em 70% para dados de treinamento e 30% para dados de teste.

Assim então, aplicando o algoritmo para cada grupo de dados especificado acima. O objetivo central será, então, medir o nível de importância de cada características dos grupos de dados utilizados pelo algoritmo para a escolha do grupo acadêmico. Foi, também, medido a acurácia do algoritmo quanto à predição de cada grupo escolhido. Por conseguinte, a acurácia de um classificador pode ser definida como a porcentagem da correta classificação do conjunto de dados de testes por um classificador [Han et al. 2011]. Esses valores de acurácia são importantes para a pesquisa uma vez que demonstra o quanto de acerto os resultados possuem.

Adicionalmente, nessa etapa, para cada grupo aplicado ao algoritmo, foram realizados 3 experimentos. O primeiro experimento foi realizado contendo todas as características do conjunto de dados. Esse cenário avalia mais de 200 características, como mostrado na Tabela 1. Esse primeiro experimento realiza uma análise dos dados completos, sem nenhuma remoção ou seleção. O segundo experimento foi realizado removendo-se a característica ANO_ADMISSAO que teve uma influência muito grande na classificação. Por fim, o terceiro experimento foi realizado contendo um total de 8 características. Um número baixo a fim de fazer um balanço mais equânime entre as características pessoais e demográficas.

Tabela 1. Resumo do conjunto completo de dados.

Grupo de características	Quantidade	Exemplos
Características internas da instituição	04	Gênero, ano de admissão, estado civil e cor/raça.
Dados demográficos municipais	237	Indicadores de demografia, educação, renda, trabalho, habitação dos municípios brasileiros tais como expectativa de vida, taxa de fecundidade, renda entre outros.

No experimento 3, foram selecionadas todas as características relacionadas ao indivíduo e um subconjunto das características relacionadas às condições socioeconômicas de sua cidade. Dos índices demográficos foram escolhidos a Taxa de Fecundidade Total (FECTOT), Esperança de Vida ao Nascer (ESPVIDA), Expectativa de Anos de Estudo (E_ANOSESTUDO) e a porcentagem de pessoas em domicílios com abastecimento de água e esgotamento sanitários inadequados (AGUA_ESGOTO). Optou-se por escolher essas características por conta da disponibilidade dos dados e por sua relevância nos indicadores sanitários, educacionais e humanos. Esses índices acima citados não estão restringidos apenas a faixas de valores, como os demais índices demográficos estão nesses referidos temas.

Como esse trabalho tem o objetivo de prever o risco de abandono do aluno ingressante, foram utilizadas apenas características que já estão presentes no momento em que o estudante é vinculado ao curso. Assim, características como a média em disciplinas específicas, ou o rendimento acadêmico não foram consideradas, uma vez que são características adquiridas no decorrer do curso.

4. RESULTADOS

Essa seção mostra os resultados obtidos após a execução dos experimentos. Como já mencionado, os resultados estão divididos em três experimentos diferentes e o principal foco

é identificar as características mais relevantes para cada contexto analisado, não apenas o resultado final do classificador.

Inicialmente, a Figura 1 mostra o resultado geral da acurácia por área do conhecimento avaliada nesse trabalho. É possível constatar nos resultados que o nível de acurácia gira em torno 70%. Esse valor é importante, pois revela que é possível identificar os possíveis casos de evasão usando a abordagem proposta. O classificador conseguiu alcançar uma taxa de acurácia maior para a área de engenharia enquanto a área de ciências sociais e ciências biológicas foram as que obtiveram os menores resultados.

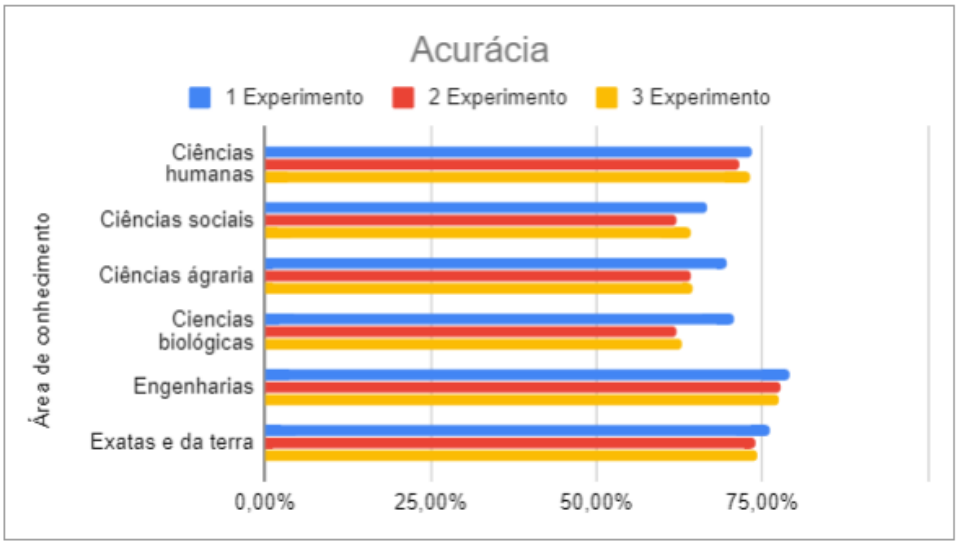

Figura 1. Acurácia por área do conhecimento.

Devido às restrições de espaço, os resultados foram agrupados e disponibilizados em formato de tabelas listando as 8 primeiras características. O resultado da aplicação do algoritmo no primeiro experimento pode ser observado na Tabela 2.

Tabela 2. Resultado da aplicação do algoritmo no 1º experimento

Feature Importance											
1º Experimento											
Ciências Sociais		Ciências Agrárias		Ciências biológicas		Engenharias		Ciências Humanas		Exata e da Terra	
Acurácia: 0.66685		Acurácia: 0.69737		Acurácia: 0.70656		Acurácia: 0.78989		Acurácia: 0.73514		Acurácia: 0.76194	
ANO ADMISSAO	0.5150	ANO ADMISSAO	0.5437	ANO ADMISSAO	0.5414	ANO ADMISSAO	0.4440	ANO ADMISSAO	0.3471	ANO ADMISSAO	0.5255
COR_RACA	0.1421	COR_RACA	0.1228	COR_RACA	0.1273	COR_RACA	0.2035	COR_RACA	0.0836	COR_RACA	0.1096
EST_CIVIL	0.0782	GENERO	0.0650	GENERO	0.0642	EST_CIVIL	0.0805	EST_CIVIL	0.0638	EST_CIVIL	0.0733
GENERO	0.0719	EST_CIVIL	0.0493	EST_CIVIL	0.0504	GENERO	0.0692	GENERO	0.0358	GENERO	0.0574
MULH0A4	0.0041	PREN20RICOS	0.0101	PREN80	0.0105	T_ATIV1824	0.0045	MULH15A19	0.0191	REN3	0.0050
PESO13	0.0040	RDPC10	0.0066	R2040	0.0068	T_FLPRE	0.0041	HOMEM20A24	0.0156	T_ANALF18A24	0.0045
PESO1824	0.0028	PREN60	0.0066	PREN20RICOS	0.0067	T_ATIV18M	0.0038	PESO610	0.0155	T_FREQ25A29	0.0043
PREN80	0.0027	T_FLMED	0.0058	PREN60	0.0058	T_FUND12A14	0.0037	PESOM15M	0.0153	CORTE1	0.0043

O resultado da aplicação do algoritmo no segundo experimento é observado na Tabela 3. Lembrando que nesse segundo experimento foi removido a característica ANO_ADMISSAO.

Tabela 3. Resultado da aplicação do algoritmo no 2º experimento

Feature Importance											
2º Experimento											
Ciências Sociais		Ciências Agrárias		Ciências Biológicas		Engenharias		Ciências Humanas		Exata e da Terra	
Acurácia: 0.62106		Acurácia: 0.64296		Acurácia: 0.62162		Acurácia: 0.77777		Acurácia: 0.71447		Acurácia: 0.73975	
COR_RACA	0.3253	COR_RACA	0.3414	COR_RACA	0.3019	COR_RACA	0.3123	COR_RACA	0.1316	COR_RACA	0.2610
EST_CIVIL	0.1410	GENERO	0.1161	GENERO	0.1407	GENERO	0.1507	EST_CIVIL	0.1049	EST_CIVIL	0.1398
GENERO	0.1241	EST_CIVIL	0.1008	EST_CIVIL	0.1107	EST_CIVIL	0.1173	GENERO	0.0565	GENERO	0.1199
PREN20RICOS	0.0107	PREN60	0.0130	TRABPUB	0.0159	T_FORA4A5	0.0074	HOMEM20A24	0.0282	T_RMAXIDOSO	0.0081
R1040	0.0085	PPOB	0.0126	RDPC5	0.0140	PREN40	0.0069	PEA1517	0.0238	PEA1014	0.0076
PREN10RICOS	0.0083	RDPC10	0.0124	CORTE4	0.0133	T_ATIV	0.0067	PEA1014	0.0237	THEIL	0.0074
PREN80	0.0061	PREN20RICOS	0.0114	RDPCT	0.0123	T_ATIV2529	0.0063	MULH30A34	0.0235	RAZDEP	0.0074
HOMEM40A44	0.0053	R1040	0.0113	PREN10RICOS	0.0089	T_LIXO	0.0059	PESOTOT	0.0229	REN5	0.0073

Por fim, A Tabela 4 apresenta o resultado do terceiro experimento. Ressaltando que, nesse experimento, foram selecionadas características predeterminadas para a análise conforme descrito na seção de metodologia.

Tabela 4. Resultado da aplicação do algoritmo no 3º experimento

Feature Importance											
3º Experimento											
Ciências Sociais		Ciências Agrárias		Ciências Biológicas		Engenharias		Ciências Humanas		Exata e da Terra	
Acurácia: 0.64167		Acurácia: 0.64605		Acurácia: 0.62741		Acurácia: 0.77373		Acurácia: 0.73062		Acurácia: 0.74238	
COR_RACA	0.2788	COR_RACA	0.3125	COR_RACA	0.2629	COR_RACA	0.2914	AGUA_ESGOTO	0.2052	COR_RACA	0.1898
EST_CIVIL	0.1291	FECTOT	0.1279	FECTOT	0.1381	GENERO	0.1300	FECTOT	0.1803	FECTOT	0.1533
FECTOT	0.1122	AGUA_ESGOTO	0.1078	AGUA_ESGOTO	0.1230	EST_CIVIL	0.1204	ESPVIDA	0.1792	AGUA_ESGOTO	0.1347
ESPVIDA	0.1121	ESPVIDA	0.1018	EST_CIVIL	0.1058	E_ANOSESTUDO	0.1013	COR_RACA	0.1201	ESPVIDA	0.1280
AGUA_ESGOTO	0.1117	E_ANOSESTUDO	0.0991	GENERO	0.1036	ESPVIDA	0.0957	E_ANOSESTUDO	0.0959	EST_CIVIL	0.1233
GENERO	0.0857	E_ANOSESTUDO	0.0903	ESPVIDA	0.0947	FECTOT	0.0890	EST_CIVIL	0.0933	E_ANOSESTUDO	0.0963
MUNICIPIO	0.0856	MUNICIPIO	0.0853	E_ANOSESTUDO	0.0886	AGUA_ESGOTO	0.0875	MUNICIPIO	0.0813	MUNICIPIO	0.0952
E_ANOSESTUDO	0.0845	GENERO	0.0748	MUNICIPIO	0.0827	MUNICÍPIO	0.0843	GENERO	0.0444	GENERO	0.0790

Para melhor destacar a importância das características analisadas no terceiro experimento, a Figura 2 apresenta as sete principais características encontradas nesse experimento divididas por área do conhecimento.

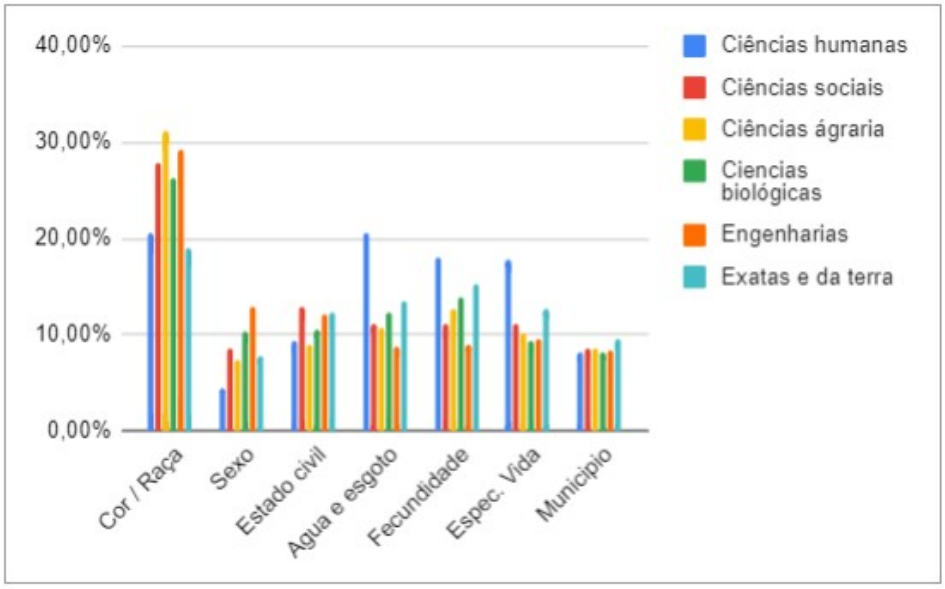

Figura 2. Principais características do terceiro experimento

5. DISCUSSÃO

É possível constatar no primeiro experimento que o ano de admissão do aluno possui grande influência na determinação do grupo acadêmico. Acredita-se que esse fato decorra da grande quantidade de um determinado valor para um grupo acadêmico, o que pode ter ocasionado um desbalanceamento no algoritmo durante o experimento. O ano de admissão ficou acima de mais de 50% de influência em muitos grupos, sendo preponderante na determinação no grupo acadêmico. Ressalte-se nesse primeiro ensaio que, seguido do ano de admissão, as características próprias dos alunos, tais como cor e gênero seguem tendo mais influência em relação às características demográficas e, de certa medida, em todos os grupos, com distanciamento e influência maior entre características do aluno sobre características de sua cidade. Destaca-se aqui que a característica de cor e raça estão presentes em todos os grupos como a segunda mais influente.

No segundo experimento, destaca-se a influência forte de cor e raça ao se remover o ano de admissão. Há, nesse ponto, um fator a ser enfatizado. A característica de ano de admissão é uma característica que não há como se repetir para alunos novos ou calouros. É uma característica do momento passado, não sendo válidas para o momento presente, assim como são valores que não são levados em consideração para programas de auxílio estudantil. Dessa forma, removê-la trará mais coerência à análise dos dados. Sendo assim, nesse segundo experimento mostra-se uma equidade maior entre as características do aluno, mantendo-se a preponderância sobre as características demográficas. Ainda nesse experimento, mantêm-se uma influência maior para a cor e raça contendo valor numérico maior sobre os demais em todos os grupos.

É possível perceber, nesse momento, que não há nenhuma característica demográfica que seja relevante ou que se mantenha constante nas primeiras medições. Ao passo que, no terceiro experimento, ao se colocar características demográficas selecionadas, o

resultado passa a ter um valor diferenciado, como mostrado na Figura 2.

No terceiro experimento, a característica de cor e raça se mantém constante como a de maior influência na determinação do grupo acadêmico. Adicionalmente, todas características ainda conservam uma equidade entre os valores, não possuindo um sobressalente ou que se distancie bastante sobre os demais. Contudo, há aqui um ponto de destaque a ser realizado. Outras características pessoais perdem espaço para características demográficas. Características como Água e Esgoto ou Taxa de fecundidade total tornam-se relevantes na determinação do grupo, acima de características pessoais como gênero ou estado civil.

Com base nos resultados apresentados, nossas principais conclusões são:

1. Característica de cor e raça é determinante em todos os cenários. O que sugere que esses elementos de cor, raça e etnia estão intrinsecamente ligados a fatores de evasão na universidade estudada.
2. Características relacionadas ao gênero do indivíduo não possuem tanta relevância, ficando atrás, em alguns grupos, de características como o estado civil. O que sugere que ser do gênero masculino ou feminino tem pouca influência na determinação da evasão na universidade estudada.
3. Características socioeconômicas da cidade influenciam na determinação da evasão, a depender de quais características serão levadas em consideração para análise. As características intencionalmente selecionadas nessa pesquisa revelaram-se mais importantes que características como gênero ou estado civil.

Essas conclusões entram em sintonia com as políticas afirmativas de cotas raciais nas universidades públicas. Existem diversos estudos que atuam nesse tópico analisando desde aspectos jurídicos, sociais e históricos [Silva 2012]. Por fim, os resultados desse artigo trazem, através da mineração de dados, outro aspecto que pode apoiar trabalhos nesse tema.

6. CONSIDERAÇÕES E TRABALHOS FUTUROS

A presente pesquisa apresentou uma análise de dados entre duas bases de dados em busca de fatores que pudessem predizer a evasão. Para isto, foram colhidos os dados dessas duas bases de dados. Uma, acadêmica, de uma universidade de 59 cursos de graduação e outra, demográfica, de índices socioeconômicos, de renda, educacionais, sanitários e de vulnerabilidade dos municípios brasileiros. Esses dados foram agrupados por áreas de conhecimento dos cursos e submetidos ao algoritmo *Random Forest*.

O objetivo seria identificar fatores que pudessem predizer a evasão. Os dados foram disponibilizados e analisados seus principais pontos de destaque, o que sugere que a característica de cor e raça está relacionada diretamente à evasão escolar, características demográficas possuem relevância moderada, não sendo fundamentais na predição do grupo, assim como demais características pessoais da pesquisa em questão não demostraram fundamental atuação na evasão escolar.

Como trabalho futuro, está em acrescentar outras características pessoais, tais como pai ou mãe registrado, estudante de escola pública ou privada, faixa de idade, faixa de renda entre outras, mantendo-se, ainda, as características demográficas para a análise, como também adicionando outras características de outras bases de dados, tais como relativas ao Sistema de Seleção Unificada (SISU), como notas e médias. Como citado na pesquisa, sempre com o objetivo de predizer a evasão no seu início; por isso, ainda não serão consideradas características como coeficiente de rendimentos ou matrículas em disciplinas ou dados obtidos no decurso do curso.

7. REFERÊNCIAS

CARRANO, Davi et al. Combinando Técnicas de Mineração de Dados para Melhorar a Detecção de Indicadores de Evasão Universitária. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, [S.l.], nov. 2019.

DELGADO, M. F. et al. (2014) "Do we need hundreds of classifiers to solve real world classification problems?", J. Machine Learning Research, v. 15, n. 1, p. 3133 – 3181.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining : um guia prático**. 4a Reimpressão. Rio de Janeiro: Elsevier, 2005.

HAN, J., Kamber, M., and Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

INEP. **Informe estatístico do MEC revela melhoria do rendimento escolar**[1998?]. Disponível em:**http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/informe-estatistico-do-mec-revela-melhoria-do-rendimento-escolar/21206** . Acesso em: 06 de Jul. 2020.

LIAW, A.; wiener, M. Classification and regression by random forest. R News, v.2, p.18-22, 2002.

LIŢOIU, N., & Oproiu, G. C. (2018). Dropout Rate in Universities–a Critical Issue of Today Education. **In The International Scientific Conference eLearning and Software for Education** (Vol. 3, pp. 77-83). " Carol I" National Defence University.

PASCOAL, Tulio et al. Evasão de estudantes universitários: diagnóstico a partir de dados acadêmicos e socieconômicos. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, [S.l.], p. 926, nov. 2016.

QUEIROGA, Emanuel; CECHINEL, Cristian; ARAÚJO, Ricardo. Predição de estudantes com risco de evasão em cursos técnicos a distância. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, [S.l.], p. 1547, out. 2017. ISSN 2316-6533.

SANTOS, Daniel Cirne Vilas-Boas dos; FALCÃO, Taciana Pontual. Acompanhamento de alunos em ambientes virtuais de aprendizagem baseado em sistemas tutores inteligentes. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, [S.l.], p. 1267, out. 2017. ISSN 2316-6533.

SILVA, Regina Maria Ferreira da. Ações afirmativas e direito fundamental à educação. Uma análise à luz das cotas raciais universitárias. **Revista Jurídica da Presidência,** v. 14, n. 104, out. 2012.

TAN, P.; Steinbach, M. K. V. (2009). **Introdução ao data mining, Mineração de Dados**. Ciência Moderna.

TIBSHIRANI, Robert; FRIEDMAN, Jerome (2008). **The Elements of Statistical Learning** (2nd ed.). Springer. ISBN 0-387-95284-5. Hastie, Trevor