UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE ENGENHARIA E GESTÃO DO CONHECIMENTO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO CONHECIMENTO

Sérgio Nicolau da Silva

Modelo de Engenharia do Conhecimento para a evasão no ensino superior

Florianópolis
2021

Sérgio Nicolau da Silva

Modelo de Engenharia do Conhecimento para a evasão no ensino superior

Dissertação submetida ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina para a obtenção do título de Mestre em Engenharia do Conhecimento.
Orientador: Prof. Fernando Álvaro O. Gauthier, Dr.

Florianópolis
2021

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

da Silva, Sérgio Nicolau
Modelo de Engenharia do Conhecimento para a evasão no ensino superior / Sérgio Nicolau da Silva ; orientador, Fernando Álvaro Ostuni Gauthier, coorientador, Rogério Cid Bastos, 2021.
124 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2021.

Inclui referências.

1. Engenharia e Gestão do Conhecimento. 2. Evasão. 3. Engenharia do Conhecimento. 4. Extração do Conhecimento. 5. Ontologia. I. Gauthier, Fernando Álvaro Ostuni. II. Bastos, Rogério Cid. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. IV. Título.

Sérgio Nicolau da Silva

Modelo de Engenharia do Conhecimento para a evasão no ensino superior

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Profa. Lia Caetano Bastos, Dra.
Universidade Federal de Santa Catarina

Prof. João Bosco da Mota Alves, Dr.
Universidade Federal de Santa Catarina

Prof. Antônio Pereira Cândido, Dr.
Instituto Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Engenharia e Gestão do Conhecimento.

ICP-Edu
Documento assinado digitalmente
Roberto Carlos dos Santos Pacheco
Data: 17/02/2021 11:07:56-0300
CPF: 439.543.830-04

Prof. Roberto Carlos dos Santos Pacheco, Dr.
Coordenador do Programa

ICP-Edu
Documento assinado digitalmente
Fernando Alvaro Ostuni Gauthier
Data: 17/02/2021 10:59:41-0300
CPF: 395.442.220-49

Prof. Fernando Álvaro O. Gauthier, Dr.
Orientador

Florianópolis, 2021.

Este trabalho é dedicado aos meus pais que sempre acreditaram no poder transformador da Educação e a minha esposa, parceira e amiga Luciana.

AGRADECIMENTOS

A Deus, pelo dom da vida.

Ao Instituto Federal de Santa Catarina, possibilitando que esta dissertação fosse concluída por meio de apoio à capacitação e disponibilização de informações cruciais para os estudos realizados.

Ao PPGEGC por ter me acolhido e me mostrado o quão importante é a união da diversidade de pensamentos e conhecimentos para a ciência.

Ao orientador Prof. Fernando Gauthier pela paciência e orientação, direcionando as ações para que fosse possível chegarmos até aqui.

RESUMO

A evasão é fenômeno e objeto de estudo desde os anos 50. Hoje no Brasil, o MEC e as IES buscam entender e mitigar tal fenômeno que, quando ocorre, afeta diretamente a sociedade: sem o devido retorno social, acadêmico e econômico, além de frustações pessoais. Este trabalho avaliou as publicações de 2015 até 2019 que apontam possíveis causas de evasão e as variáveis associadas. A partir destes, é proposto um modelo de Engenharia do Conhecimento suportado por uma ontologia que auxilia as instituições de ensino superior a direcionar as suas análises sobre a evasão em suas bases de dados estruturadas. Como prova de conceito (PoC), aplicou-se ao modelo o processo KDD – *Knowledge-Discovery in Databases* para extrair conhecimentos sobre a correlação entre variáveis e, por meio de *machine learning* para classificação, a avaliação da capacidade de predição de comportamento futuro de evasão. Como resultado, obteve-se o nível de correlação entre variáveis que direcionam as ações para detalhamento da análise da evasão. No que tange à classificação, dentre os algoritmos de classificação submetidos, redes neurais se mostrou uma escolha promissora para a instituição analisada, com área sobre a curva ROC de 0,91, acurácia de 0,86, precisão 0,86, *recall* 0,79 e *f-score* 0,81. Tais conhecimentos extraídos na PoC – correlação entre variáveis e métricas de capacidade de predição – realimentaram o modelo proposto por meio de uma extensão da ontologia do modelo. A partir dos resultados alcançados com a PoC, constata-se que o conjunto de causas e variáveis propostas – identificadas no arcabouço de publicações – direciona a fase de seleção de dados no processo KDD e promove nível de confiança para correlação e predição. Isto torna o modelo um bom direcionador para a análise de evasão. Como evolução, pretende-se projetar e implantar um SBC baseado no modelo para a instituição de ensino avaliada, auxiliando-a na Gestão do Conhecimento sobre o fenômeno e a identificar discentes com tendência de evasão a fim de mitigar os riscos.

Palavras-chave: Evasão. Engenharia do Conhecimento. Ontologia. KDD.

ABSTRACT

Dropping out university has been a phenomenon and object of study since the 1950s. Nowadays, in Brazil, the Ministery of Education (MEC) and the Higher Education Institutions (IES) try to understand and mitigate such a phenomenon which, when occurring, directly affects society: without the expected social, academic and economic return, besides personal frustrations. This work has evaluated some papers published from 2015 to 2019 that identify the possible causes of the dropout and indicate the variables associated to it. Based on those works, a Knowledge Engineering model is proposed, and supported by an ontology that helps the higher education institutions to direct their analysis about dropping out in their structured databases. As a proof of concept (PoC), the KDD - Knowledge-Discovery in Databases – process was applied to the model to extract knowledge about the correlation among variables and, through machine learning classify and evaluate the ability to predict future possibilities of dropping out. As a result, the level of correlation among the variables that direct the actions to detail the dropping out analysis was obtained. Regarding the classification, among the classification algorithms submitted, neural networks proved to be a promising choice for the analyzed institution, with an area of 0.91 on the ROC curve, 0.86 accuracy, 0.86 precision, 0.79 recall and 0.81 f-score. Such knowledge extracted in the PoC - correlation among variables and metrics of predictive capacity - fed back the proposed model through an extension of the model's ontology. From the results achieved with the PoC, it is possible to say that the set of causes and variables proposed - identified in the publications framework - directs the phase of data selection in the KDD process and promotes a confidence level for correlation and prediction. This makes the model a good driver to analyze the dropping out process. As an upgrade it is intended to design and implement a SBC based on the model for the institution evaluated, assisting it in the Knowledge Management about the phenomenon and also to identify students who tend to drop in order to mitigate the risks.

Keywords: Dropping out. Knowledge Engineering. Ontology. KDD.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

EC – Engenharia do Conhecimento

EGC – Engenharia e Gestão do Conhecimento

EBTT – Ensino Básico, Técnico e Tecnológico

ETL – *Extract, Transformation, and Load*

GC – Gestão do Conhecimento

IDEB – Índice de Desenvolvimento da Educação Básica

IES – Instituição de Ensino Superior

IFSC – Instituto Federal de Santa Catarina

INEP – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

KDD – *Knowledge-Discovery in Database*

MEC – Ministério da Educação

PPGEGC – Programa de Pós-graduação em Engenharia e Gestão do Conhecimento

PoC – *Proof of Concept*

OECD – *Organization for Economic Co-operation and Development*

RUF – Ranking Universitário Folha

SINAES – Sistema Nacional de Avaliação da Educação

TIC – Tecnologia da Informação e Comunicação

WoS – *Web of Science*

SUMÁRIO

1 INTRODUÇÃO

O fenômeno da evasão escolar vem sendo percebido pela comunidade acadêmica nacional e internacional há anos. Vários pesquisadores e instituições têm buscado compreender melhor tal fenômeno e atuar para mitigá-lo. Esse trabalho aborda tal situação, apresentando dados e informações relevantes sobre o tema, que corrobora com a sua importância para as IES. A partir de um breve histórico sobre o ensino superior brasileiro e relacionando-o com evasão, este trabalho aborda o tema sob a visão da Engenharia do Conhecimento como norteador para as instituições avaliarem esta ocorrência, bem como na descoberta de conhecimento sobre os dados de uma instituição pesquisada.

1.1 CONSIDERAÇÕES INICIAIS

No que se refere às questões do direito básico à educação, especialmente no ensino superior, há uma preocupação com a qualidade dos serviços ofertados. Isto motivou a busca por estruturas e mecanismos para aferição e garantia da qualidade do ensino. A partir dos anos de 1990, diversos países latino-americanos criaram seus organismos de avaliação de cursos superiores (DIAS-SOBRINHO, 2006). Como resultados obtidos por estes organismos, "os sistemas de avaliação oferecem subsídios para que as universidades busquem adotar procedimentos formais de melhoria de desempenho, pautados no autoconhecimento e na organização dos processos" (GUERRA, 2019, p. 291).

No Brasil, o governo federal atribuiu a responsabilidade de monitorar e avaliar a qualidade do ensino ao Sistema Nacional de Avaliação da Educação Superior – SINAES do INEP (DIAS-SOBRINHO, 2006). Paralelamente, iniciativas independentes surgiram neste sentido, como o Ranking Universitário Folha – RUF[1], que classifica as melhores universidades do Brasil com base em cinco indicadores: pesquisa científica, qualidade de ensino, internacionalização, mercado de trabalho e inovação.

[1] Ranking Universitário Folha http://ruf.folha.uol.com.br/

Entretanto, mesmo com tais ações voltadas à qualidade do ensino, ainda é um desafio político, social e educacional oferecer, por meio de políticas públicas, maior oportunidade de acesso e garantir o êxito (DE GUIMARÃES, *et al.*, 2019).

Sendo assim, a busca de garantia de qualidade do ensino por si só não garante que todo aluno ingressante em um curso o conclua com sucesso, conhecida como "permanência e êxito" nos casos de concluintes ou "evasão" para os não concluintes. O fenômeno da evasão (não êxito) tem deixado em estado de atenção o MEC[2] e as Instituições de Ensino Superior no Brasil. Em 2016 o então Ministro da Educação, Mendonça Filho enfatiza – baseado nos dados de 2010 a 2014 do INEP[3], órgão também responsável pelo SINAES – a dificuldade que reflete num acréscimo desordenado nas taxas de desistências nos cursos de ingresso (MEC, 2016).

Todavia, a preocupação tem relação mais ampla que o simples fato de concluir um curso superior. Devem ser considerados os efeitos produzidos pela educação superior, que vão além das questões profissionais: os conhecimentos técnicos adquiridos promovem melhorias na economia, na qualidade de vida, bem como na cidadania (DIAS-SOBRINHO, 2006) (SILVA FILHO, *et al.*, 2007) (DE LIMA; ZAGO, 2018) (CAMPOS, *et al.*, 2017). Sem atingir os objetivos, os efeitos são reduzidos. Ou seja, pode-se afirmar que há uma perda do investimento – não só financeiro – para a formação acadêmica deste aluno e retorno à sociedade. Isto ocorre porque o propósito de ter, ao final do curso, um profissional capacitado e habilitado para a função em determinada área, não é alcançado em sua totalidade quando o aluno evade.

Aprimorando a análise sobre a evasão – Tabela 1 – e possibilitando uma visão mais detalhada sobre os impactos desta, apresenta-se o Censo da Educação Superior de 2017 do MEC com a relação entre ingressos e concluintes de 2011 a 2017.

---

[2] Ministério da Educação
[3] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

Tabela 1 - Ingressos e egressos por ano

Área Geral do Curso	Ingressantes para cada 10.000 habitantes								Concluintes para cada 10.000 habitantes							
	Total OCDE 2014	Brasil							Total OCDE 2014	Brasil						
		2011	2012	2013	2014	2015	2016	2017		2011	2012	2013	2014	2015	2016	2017
Ciências sociais, negócios e direito	21,7	50,3	59,7	56,4	61,3	55,8	56,0	60,6	23,2	22,0	23,1	21,8	22,1	24,3	23,7	23,0
Educação	5,1	23,3	24,8	23,6	28,0	25,9	29,0	31,5	5,4	12,2	11,3	10,0	10,7	11,6	11,6	12,3
Saúde e bem estar social	9,8	14,3	16,4	17,0	20,4	19,7	21,1	24,4	9,8	7,8	8,2	7,0	6,7	7,7	7,8	8,5
Engenharia, produção e construção	11,5	14,8	19,0	20,2	22,7	20,8	18,4	17,4	9,1	3,3	3,8	4,0	4,4	5,2	6,1	6,8
Ciências, matemática e computação	5,9	8,2	9,1	8,9	9,3	8,9	8,8	9,4	5,7	2,9	3,0	2,7	2,8	3,0	3,0	3,0
Agricultura e veterinária	1,2	2,3	2,7	2,8	3,3	3,3	3,4	3,6	1,1	1,0	1,0	1,0	1,0	1,1	1,2	1,3
Humanidades e artes	10,9	3,0	3,4	3,3	3,3	3,4	3,3	3,7	11,4	1,3	1,4	1,4	1,4	1,4	1,5	1,6
Serviços	4,8	3,4	3,9	4,2	4,1	4,1	3,9	4,3	4,8	1,5	1,6	1,4	1,6	1,9	1,9	1,6

Fonte: MEC (2018, p. 46)

Como é percebido, a relação entre ingressos e egressos ao longo dos anos não tem seguido o ritmo esperado no que tange a reflexos do retorno destes como profissionais capacitados à sociedade. Ao passo que a quantidade de ingressos vem aumentando, a de egressos está praticamente estável. Para De Lima e Zago (2018, p. 368), isto caracteriza uma baixa taxa de sucesso (êxito) anual. Analisando um mesmo ano, os dados corroboram com a constatação anterior, o que carece de atenção por parte das IES. Por exemplo, para a área "Ciências sociais, negócios e direito" em 2017 foram 60,6 ingressantes contra 23,0 egressos para cada 10.000 habitantes.

Confrontando esses dados com o universo de 8.450.755 alunos matriculados (INEP, 2019b), tem-se uma boa percepção do volume de evasão que ocorre no Brasil. Este comportamento mostra que a evasão é algo desafiador às IES na busca de encontrar meios para aumentar a taxa de sucesso (reduzir a evasão) dos ingressantes, de tal forma que a diferença entre quantidade de ingresso e egresso seja minimizada ao máximo possível.

1.2 IDENTIFICAÇÃO DO PROBLEMA

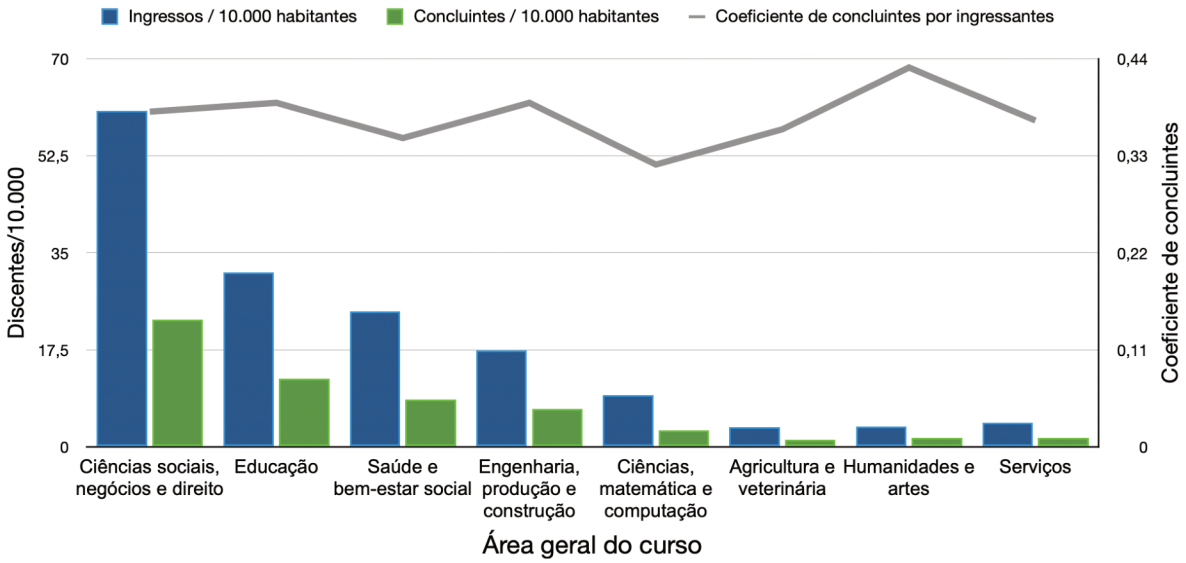
O fenômeno da evasão impacta negativamente o social, o acadêmico e o econômico do país, levando estudiosos a buscar pela compreensão de tal fenômeno e, em alguns estudos, propor formas para mitigá-lo.

Neste documento, considera-se evasão escolar "a situação do aluno que abandonou a escola ou reprovou em determinado ano letivo e que, no ano seguinte, não efetuou a matrícula

para dar continuidade aos estudos" (QEDU, 2019). No caso dos cursos de ciclos diferentes de ano, como os semestrais, considera-se tal ciclo.

Com tal conceito delineado, analisando a Tabela 1, mais especificamente o ano de 2017, e calculando o coeficiente pela fórmula: concluintes / ingressantes. Tem-se os seguintes valores entre ingressantes e concluintes no mesmo ano / área:

Figura 1 – Coeficiente de concluintes por ingressantes 2017



Fonte: Do autor

Tomando como um coeficiente ideal de 1, no qual a quantidade de egressos é igual a de ingressos, depreende-se da

Figura 1 que em todas as áreas a ocorrência de evasão possui um coeficiente de concluintes por ingressante abaixo de 0,45 pontos. Isto torna o tema relevante para a educação no país. Sob a ótica financeira, segundo o INEP (2015) o investimento anual em um aluno de graduação é de R$ 23.215,00 (vinte e três mil, duzentos e quinze reais); ou seja, em caso de evasão tal investimento por aluno evadido não trará, na sua totalidade, o retorno desejado à sociedade.

Não obstante a dificuldade de conclusão no ensino superior, para o MEC (2016) a falta de orientação vocacional no ensino médio também tem sua parcela de contribuição quanto ao insucesso. Isto deve ser revisto para contribuir com a escolha mais acertada do curso pelo ingressante e, consequentemente, a permanência no ensino superior. O que caracteriza tal necessidade é o fato de o discente ingressar no ensino superior e, durante o curso, perceber que

sua escolha de curso não é “a sua vocação”. Constatado isso, ele parte para um novo processo seletivo para ingresso em outro curso e, em caso de sucesso, abandonando o anterior. É o que mostra levantamentos do (INEP, 2018), nos quais 21% dos que ingressaram em um curso superior em 2017 realizaram nova prova do ENEM no ano corrente. Este fato vai ao encontro da necessidade de orientação vocacional no ensino médio com foco na escolha do curso de ensino superior mais adequado às expectativas dos alunos.

Paralelo a este cenário de evasão que se apresenta, as instituições de ensino superior públicas, no Brasil, vêm sofrendo com contingenciamentos e bloqueios de seu orçamento (MEC, 2019). Isto reflete na dificuldade em executar as práticas acadêmicas/pedagógicas, o que poderá desmotivar ainda mais os alunos, podendo elevar a tendência de evasão. Em épocas de baixa disponibilidade de recursos financeiros, conter a escalada da evasão é também uma forma de melhor aplicar os recursos escassos disponíveis.

Sendo assim, na busca da compreensão de tal fenômeno, pesquisadores têm aplicado esforços no entendimento, mapeamento e proposição de soluções. Em termos de publicações nos anos de 2015 até 2019, e utilizando-se das bases científicas mais significativas na área pesquisada, os dados mais relevantes foram encontrados nas bases *Scielo*, *Scopus* e *Web of Science* (WoS). Utilizando-se os termos “evasão” e “ensino superior” ou “dropout” juntamente com “undergraduate degree” ou “higher education”, foi identificado o volume de publicações apresentado na Tabela 2.

Tabela 2 – *String* de pesquisa e bases analisadas

***String* de pesquisa nos anos de 2015 até 2019**	**Resultados obtidos**		
	WoS	Scopus	Scielo
("drop out" and ("undergraduate degree" or "higher education")) OR ("evasão" and "ensino superior")	165	161	35

Fonte: do Autor.

Analisando os resultados da WoS (com o maior número de publicações), identificou-se as características apresentadas a seguir.

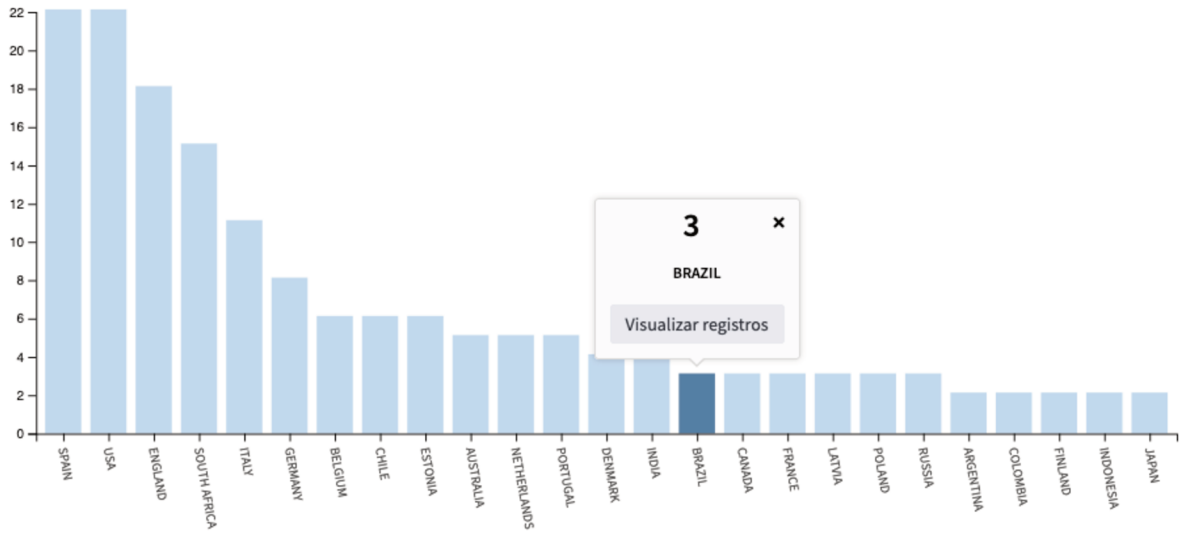
Tabela 3 – Publicações dos últimos 5 anos

Anos de publicação	Quantidade de registros	% de 165
2019	23	13,94 %
2018	42	25,45 %
2017	43	26,06 %
2016	31	18,79 %
2015	26	15,76 %

Fonte: Web of Science. Data da extração 19/11/2019

Ano a ano existe uma preocupação crescente com a evasão nas IES de todo o mundo, evidenciado nas publicações por ano (Tabela 3). As publicações por países evidenciam que Estados Unidos (22 registros) e Espanha (22 registros) apresentam a maior quantidade de publicações. O Brasil apresenta apenas três trabalhos registrados, conforme Figura 2.

Figura 2 – Publicações por país



Fonte: Web of Science. Data da extração 19/11/2019

Para a seleção das publicações a serem utilizadas como base desta pesquisa, realiza-se a leitura dos resumos, metodologias e resultados das publicações encontradas na base. Como resultado foram selecionados 20 artigos que tratam especificamente dos fatores de evasão e, em alguns casos, propõem ferramentas, métodos ou processos para mitigá-la.

1.3 PERGUNTA DE PESQUISA

Diante das pesquisas e dados apresentados, evidencia-se a importância do tema para a educação. No Brasil, as pesquisas indicam que ainda existe deficiência no que tange a permanência e êxito em cursos superiores, tendo em vista os números apresentados pelo INEP (2018). Os artigos selecionados – que abordam e apresentam as causas que levam à evasão – podem ser um norte para a análise deste problema. Diante deste cenário, tem-se a seguinte questão a ser abordada:

Como a Engenharia do Conhecimento pode auxiliar as instituições de ensino superior na compreensão do fenômeno da evasão e direcionamento das ações para evitar isto?

Para a pesquisa proposta como resposta à pergunta, estão definidos os objetivos geral e específicos.

1.4 OBJETIVOS

1.4.1 Objetivo Geral

Criar um modelo de evasão como forma de compartilhamento do conhecimento, suportado por uma ontologia e dados abertos, sendo um norteador na análise de evasão na IES.

1.4.2 Objetivos Específicos

- Identificar as principais causas da evasão no ensino superior por meio de revisão bibliográfica;
- Identificar as variáveis que representam os fatores relacionados à evasão para as causas apontadas nas publicações mais relevantes;
- Propor um modelo de Engenharia do Conhecimento baseado em ontologia;
- Desenvolver o modelo proposto;
- Realizar a prova de conceito em uma instituição de ensino superior.

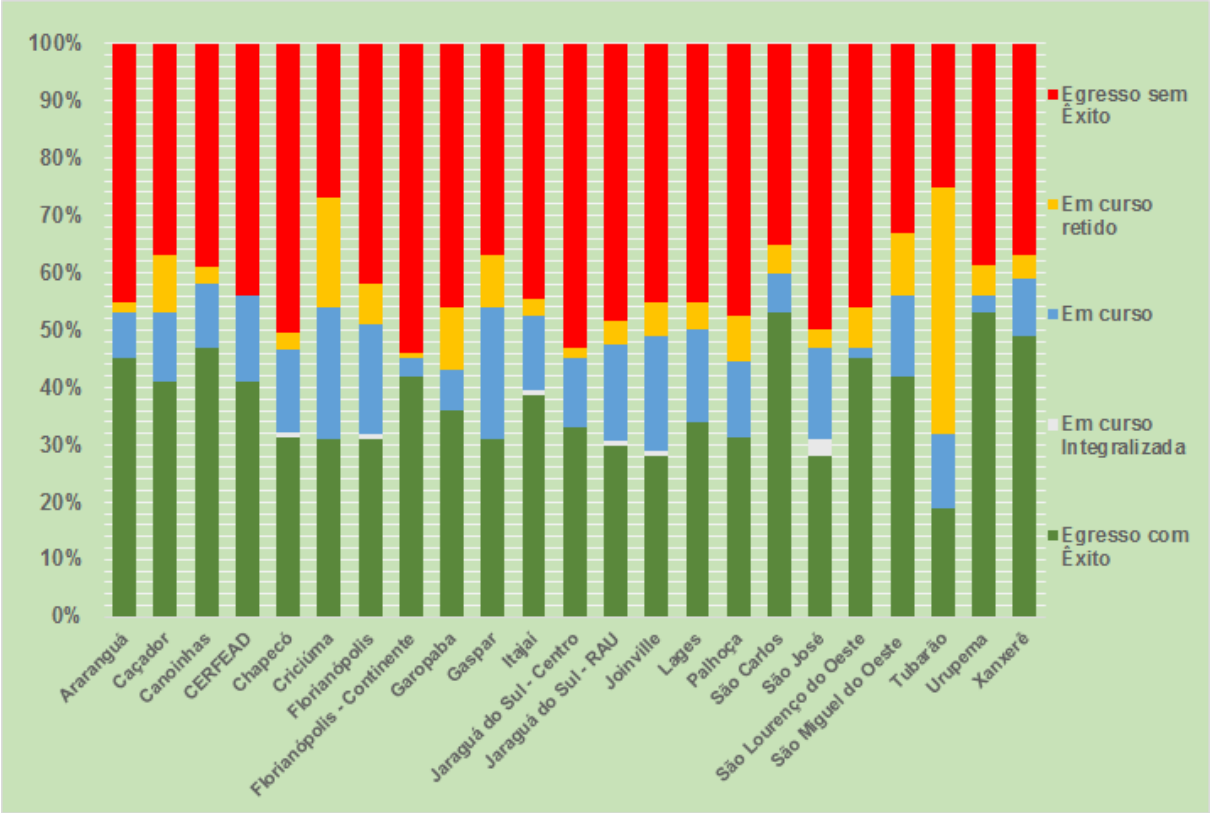
1.5 JUSTIFICATIVA

As evidências apresentadas pelo INEP (2018) em forma de dados sintéticos sobre o ensino superior no Brasil mostram que as IES têm dificuldades em manter o discente do início ao final do curso e que ele alcance o êxito.

No Brasil, no período avaliado na base de pesquisa com maior número de publicações retornadas, apenas 3 publicações brasileiras foram identificadas, conforme Figura 2. Mesmo assim, há declarações recentes de IES que estão cada vez mais preocupados com tal fenômeno, na busca da compreensão para mitigá-lo.

Especificamente na instituição que é objeto deste estudo, a Reitora do IFSC declara a necessidade de focar na permanência e êxito dos alunos da instituição, considerando isto como uma obrigação social e também uma das prioridades da instituição (IFSC, 2019). Em agosto de 2018 foi publicado o Plano Estratégico de Permanência e Êxito dos Estudantes do IFSC (PPE-IFSC) que tem como objetivo "promover a permanência e êxito dos estudantes [...] por meio de um conjunto de medidas, que visam o enfrentamento da evasão e retenção, enquanto fatores que comprometem o atendimento da missão institucional" (IFSC, 2018). O PPE-IFSC está embasado em dados relacionados com a evasão detectada em IES de mesmo tipo, sendo um dos dados em destaque a alta taxa de evasão na instituição. A Figura 3 mostra, do total (100%) de matrículas em cada campus do IFSC efetuadas de 2009 até 2017, quantos destes (também em percentual) são egressos com sucesso, em integralização, em curso, retido e egresso sem êxito (vermelho).

Figura 3 – Total de matrículas e situação das matrículas por campus do IFSC (2009-2017)



Fonte: IFSC (2018)

Depreende-se da Figura 3 que o percentual de evasão (representado pelo vermelho) sobre o total da evasão (soma de todas as cores de uma barra = 100%) para cada campus é elevado. O caso mais representativo é no campus Jaraguá do Sul Centro com aproximadamente 55% de evasão.

A instituição tem consciência e sabe da importância do tema quanto ao êxito dos alunos, e acredita que "a partir do momento em que o estudante entra no IFSC, somos responsáveis pelo seu êxito e isso significa uma mudança de paradigma e até da cultura institucional" (IFSC, 2019). Tal cenário no IFSC é corroborado por dados e estatísticas nacionais. Para o MEC (2016) este tema é considerado relevante para a educação superior e, com as restrições orçamentárias aplicadas à Educação, aumenta a dificuldade em realizar ações para redução da evasão.

Quanto à questão orçamentária, para cada aluno evadido tem-se um investimento anual da ordem de R$ 23.215,00 que não teve seu retorno pleno para a sociedade quanto à formação de nível superior (INEP, 2015). Como os cursos de instituições públicas e boa parte

dos cursos de instituições privadas (empréstimos e bolsas) são custeados pelo governo, é importante identificar estes motivos que levam à evasão (SACCARO; FRANÇA; JACINTO, 2019).

Sob a percepção de pesquisadores do assunto, Junior e Real (2017) destacam que os estudos sobre a temática no Brasil são considerados de importância para o entendimento do fenômeno da evasão, porém rudimentares. Também acreditam serem relevantes para a redução dos índices e, como reflexo, auxiliam no que tange a ampliar o acesso ao ensino superior.

Com uma abordagem suportada pela Engenharia do Conhecimento aplicada sobre as causas que levam à evasão, apresenta-se um modelo de engenharia como ferramenta para a sua compreensão. A partir do modelo, aplicando-o aos dados históricos de uma instituição de ensino, possibilita à IES aprimorar o autoconhecimento sobre o tema por meio de análise e descoberta de conhecimento.

Neste sentido, com a compreensão do fenômeno, com base no arcabouço identificado pela revisão bibliográfica do tema, a presente dissertação faz uso de métodos, técnicas e ferramentas de extração e compartilhamento do conhecimento para possibilitar um modelo de Engenharia do Conhecimento focado na compreensão do fenômeno e direcionamento das ações de melhoraria da gestão da permanência e êxito.

1.6 ADERÊNCIA AO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO CONHECIMENTO

A dissertação é desenvolvida dentro do Programa de Pós-graduação em Engenharia e Gestão do Conhecimento (PPGEGC) utilizando-se de métodos e técnicas para representação e de extração do conhecimento em dados estruturados (*Knowledge Discovery in Database* - KDD). Esta extração vem a produzir artefatos de conhecimento para apoio às áreas de gestão para a tomada de decisão referente a uma modelagem de conhecimentos, com vistas a mitigar a evasão na educação superior. Estando inserido na área Engenharia do Conhecimento, o tema abordado faz parte da linha de pesquisa de Engenharia do Conhecimento Aplicada às Organizações.

Tem seu foco nas estruturas organizacionais de governo que promovem Cursos Superiores e trata o conhecimento como um recurso importante na gestão de tais cursos. Visa proporcionar, assim, um melhor aproveitamento dos conhecimentos explícitos nas plataformas

digitais, voltados à tomada de decisão e, consequentemente, agregando valor ao serviço prestado e à sociedade. Tal conhecimento é abordado na visão cognitivista, pela qual o conhecimento é algo que pode ser armazenado, processado e facilmente compartilhado (PACHECO, 2016).

No repositório de teses e dissertações do Programa de Pós-graduação em Engenharia e Gestão do Conhecimento – PPGEGC – foram identificados quatro trabalhos que possuem alguma relação com a evasão em cursos, sendo duas teses e duas dissertações e estão listados no Quadro 1.

Quadro 1 - Publicações no banco de teses e dissertações PPGEGC

AUTOR	ANO	TIPO	TRABALHO
PACHECO, A. S. V.	2010	Tese	Evasão e permanência dos estudantes de um curso de administração do sistema Universidade Aberta do Brasil: uma teoria fundamentada em fatos e na gestão do conhecimento.
COMARELLA, R. L.	2009	Dissertação	Educação superior a distância: evasão discente. Dissertação, 2009.
CISLAGHI, R.	2008	Tese	Um modelo de sistema de gestão do conhecimento em um framework para a promoção da permanência discente no ensino de graduação.
RAMOS, B. M. S.	2007	Dissertação	Eficácia no uso de tecnologias para alavancar o aprendizado do idioma inglês no ensino médio, Dissertação, 2007.

Fonte: Banco de Teses e Dissertações do EGC[4].

Dos três trabalhos mais aderentes, Pacheco (2010) e Comarella (2009) apresentam suas pesquisas e resultados na busca da compreensão do fenômeno de evasão nos respectivos grupos delimitados. Apontam que o paradigma funcionalista é mais atuante na busca da efetividade (PACHECO, 2010) e, em uma pesquisa ampla com tutores e discentes, os fatores associados ao tempo dedicado aos estudos como relevantes (COMARELLA, 2009). Cislaghi (2008) aprofunda mais tal análise propondo um *framework* de indicadores, sensores e procedimentos com base nas causas da evasão no Brasil. Considera fatores políticos e de liderança institucional como chave para o assunto. Por tal completude, esse trabalho irá compor o conjunto selecionado para o alcance dos objetivos específicos da dissertação.

[4] Endereço BTD: http://btd.egc.ufsc.br

Já Ramos (2007) analisa o ensino de idiomas por meio de ambientes virtuais para o ensino médio, recomendando algumas ações para mitigar a evasão.

Denota-se dos trabalhos citados que a presente dissertação vem ao encontro da necessidade de compreender o fenômeno evasão. A dissertação traz como contribuição a consolidação de conhecimentos já formalizados e os compartilha em um modelo, explicitando as causas e suas variáveis que levam à evasão, por meio de artefatos suportados pela Engenharia do Conhecimento. Acrescenta-se a utilização complementar da área de Gestão em seu Ciclo de Gestão do Conhecimento e dos processos decisórios por meio da utilização dos artefatos. Assim, a partir do modelo, tais artefatos auxiliam as instituições no direcionamento das ações quanto à compreensão geral da evasão e específicas para a sua realidade, baseado na extração de conhecimento sobre suas bases estruturadas – como mostra a prova de conceito.

A partir disto, tal dissertação direciona os estudos relacionados às áreas para compreensão do fenômeno e previsão de tendência de evasão em grupos de discentes. Em uma visão mais ampla, considerando a possibilidade de um repositório de dados abertos suportado pelo modelo, várias instituições poderiam compartilhar seus dados seguindo o modelo. Assim, mais pesquisadores interessados no tema podem fazer uso dos dados na busca de novos conhecimentos e meios para evitar a evolução do fenômeno da evasão.

1.7 METODOLOGIA DA PESQUISA

Nesta seção são apresentadas as questões referentes à metodologia de pesquisa aplicada nesta dissertação.

1.7.1 Classificação metodológica

“Dito de maneira simples, Ciência é o conhecimento da natureza e a exploração desse conhecimento” (KNELLER, 1980, p. 11). É da natureza humana entender e desvendar os mistérios da natureza, sendo um dos pilares da ciência moderna o conhecimento racional. Este é recuperado por meio de procedimentos metodológicos e seus métodos, possibilitando testar, verificar e experimentar todo o novo conhecimento, denominado conhecimento científico. Conhecimento científico, portanto, é algo baseado em fatos, que pode ser organizado e

sistematizado por meio de processos bem definidos, podendo ser replicado (APOLINÁRIO, 2012).

À luz da visão sobre ciência e conhecimento científico, tal pesquisa se enquadra no paradigma funcionalista de Morgan (1980), por ter como objetivo a produção de artefatos tecnológicos com uso de métodos e técnicas da Engenharia do Conhecimento.

Sendo uma pesquisa tecnológica aplicada e exploratória promove, de forma prática e dirigida, uma solução para um problema específico. Esta é uma pesquisa que envolve solução tecnológica possuindo natureza aplicada por produzir resultados práticos (GIL, 2008). Quanto aos objetivos, trata-se de uma pesquisa descritiva e explicativa ao descrever os fatos observados e explicar suas causas (GIL, 2010).

Quanto à abordagem, é uma pesquisa quali-quanti que interpreta fenômenos, atribui significados e, ao mesmo tempo, utiliza recursos e técnicas de estatísticas (VIANNA, 2013).

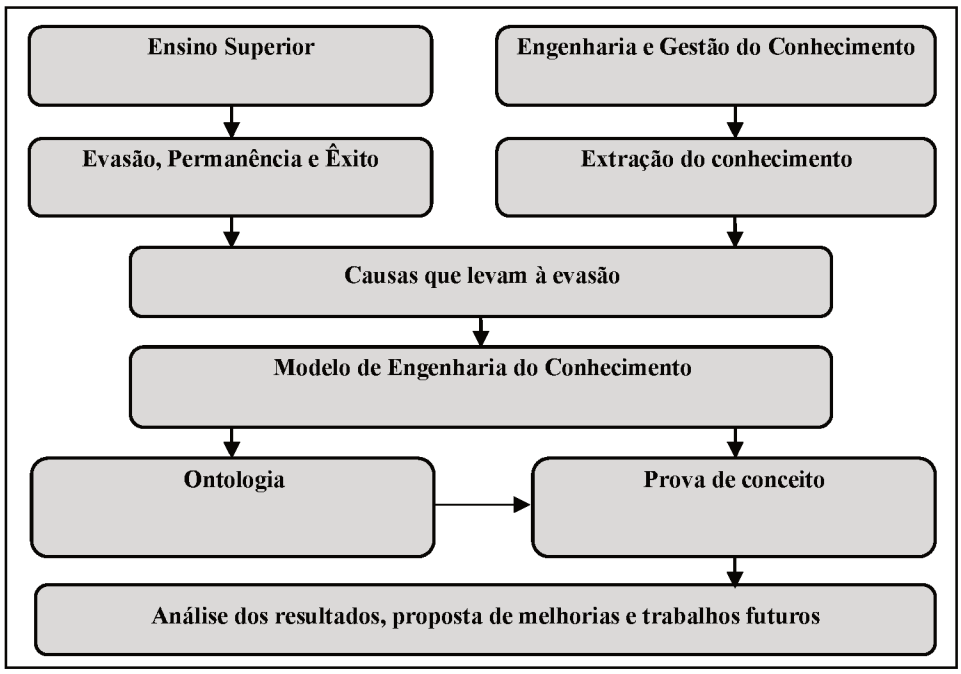
Vale-se dos procedimentos bibliográficos, baseados em publicações existentes sobre o tema e sendo um estudo de caso, coletando e analisando dados de uma instituição de ensino superior. Também experimental ao observar as causas de influência nos fenômenos associados à evasão (LAKATOS; MARCONI, 2003).

Para as fases de identificação das práticas de gestão e das causas de evasão, além das técnicas e métodos apropriados de extração do conhecimento, será aplicada uma revisão integrativa da literatura, reunindo em um arcabouço sólido de conhecimentos já explicitado sobre o tema para ser aplicado neste trabalho.

No que tange à aplicabilidade e uso dos artefatos é utilizado o método experimental que consiste, especialmente, em submeter os objetos de estudo à influência de certas variáveis, em condições controladas e conhecidas pelo investigador, para observar os resultados que a variável produz no objeto (GIL, 2008).

1.7.2 Procedimentos metodológicos

Figura 4 – Procedimentos metodológicos



Fonte: Criação do autor

A Figura 4 resume o procedimento metodológico, no qual são abordados os aspectos relacionados ao ensino superior e evasão / permanência e êxito, concomitantemente às questões de Engenharia e Gestão do Conhecimento e seus métodos, técnicas e ferramentas aplicadas à evasão.

Identificadas pelas pesquisas as principais causas e variáveis que levam à evasão, é definida uma ontologia que representa tal conhecimento e direciona à instanciação dos indivíduos para carga e disponibilização no formado de dados abertos.

Partindo do conhecimento representado, é realizada uma prova de conceito com base nos dados de uma IES, utilizando o método de extração de conhecimento em banco de dados sobre os dados e informações estruturadas disponibilizadas e autorizadas pela IES pesquisada.

Por fim, serão apresentados os resultados alcançados com a aplicação do modelo, verificando a aderência ao proposto, pontos de melhoria para o modelo e trabalhos futuros.

Tal pesquisa possui suas delimitações quanto ao escopo e abrangência aplicada sobre o tema evasão.

1.8 DELIMITAÇÃO DO TRABALHO

Neste trabalho é proposta uma ontologia com as causas de evasão e um conjunto de métricas sugeridas. A base para composição será o levantamento realizado na revisão integrativa. Quanto à prova de conceito, faz uso de métodos de extração do conhecimento sobre os dados de uma IES, instanciando o modelo ontológico e disponibilizando em formato de dados abertos. A partir desses, os dados são analisados e novos conhecimentos extraídos dos mesmos são alimentados ao modelo.

Neste âmbito, dado o objetivo e dentro da gama de recursos da Engenharia do Conhecimento, são aplicadas as ferramentas que possibilitam a extração, a representação e o compartilhamento do conhecimento norteados pelo modelo proposto.

O estudo está aplicado sobre a visão de três estados principais do discente durante a vida acadêmica:

- Cursando: são aqueles discentes ativos na instituição que estão cursando ou com trancamento em uma ou mais disciplinas;
- Evadido: o discente que interrompeu o curso, sem quaisquer perspectivas de retorno, autodeclarada ou não;
- Sucesso: são os discentes que cumpriram com todas as exigências para alcançar o grau do curso e têm seu grau atribuído por meio de diplomação.

Existe situação intermediária, na qual o aluno ainda não concluiu o curso e já ultrapassou o prazo mínimo do programa, motivado por reprovações e trancamentos, que deixa certa dúvida sobre ter ou não evadido. Esta condição será tratada como um discente considerado no estado de “cursando”, haja vista que “estudante retido ainda pode concluir o curso que ingressou, mesmo que em um prazo maior” (DE LIMA; ZAGO, 2018, p. 369). Logo, para este trabalho, tal condição se encaixa na definição.

A prova de conceito foi aplicada somente no IFSC, não sendo replicada em tempo de desenvolvimento da dissertação em outra instituição. No entanto, pretende-se que o modelo possibilite tal aplicabilidade futura para as instituições que tiverem interesse e disponibilidade de recursos de conhecimento, humanos e tecnológicos para tal.

Não está no escopo a extração de dados de sistemas legados do IFSC, devido ao fato de estarem inacessíveis ou totalmente desativados. Considera-se então os dados disponíveis nos sistemas atuais e que, porventura, podem ter em suas bases dados que foram migrados dos legados. Além disso, caso sejam identificadas dificuldades ou inconsistências em algum dado / informação no sistema, estes não serão corrigidos nos sistemas de origem por não ter autorização para tal e necessitar de análises mais aprofundadas por parte da instituição.

O autor dessa dissertação está autorizado pela Pró-reitoria de Pesquisa, Pós-graduação e Inovação - PROPPI do IFSC por meio do processo 23292.013505/2019-25 para uso dos dados da instituição. Em atendimento às normativas da PROPPI vigentes e expressas na autorização, e em cumprimento ao Art. 1º, parágrafo único da Resolução CNS 510/2016, tais dados aqui apresentados são reais, porém, são mascarados sempre que ocorrer a possibilidade de identificação de um ou mais indivíduo, sendo garantido o sigilo.

1.9 ESTRUTURA DO TRABALHO

O trabalho está organizado em 6 capítulos: introdução, evasão no ensino superior, Engenharia do Conhecimento, proposta de modelo, prova de conceito e por fim as conclusões e recomendações.

Na introdução é apresentada a temática, iniciando pelas considerações sobre o tema e identificando o problema abordado, juntamente com a pergunta de pesquisa. A partir disso, são apresentados os objetivo geral e específicos, a justificativa para a pesquisa e como ela se adere ao Programa de Pós-graduação em Engenharia e Gestão do Conhecimento. Como parte final da introdução, são apresentados a metodologia de pesquisa, a delimitação do trabalho e sua estrutura.

O capítulo 2 trata da Evasão no Ensino Superior, passando por um breve histórico do ensino superior no Brasil. A partir desse ponto, aborda-se a visão de evasão, permanência e êxito dos discentes. Por fim, os estudos associados a causas de evasão identificadas por meio da revisão integrativa.

Tendo o entendimento da temática evasão, apresenta-se uma revisão sobre a Engenharia do Conhecimento, a extração do conhecimento principalmente sobre dados estruturados, aprofundando-se em KDD. Ao final do capítulo apresenta-se conceito sobre ontologia e dados abertos.

Baseado no arcabouço de publicações, é proposto um modelo no capítulo 4, abordando a seleção dos grupos de causas de evasão e suas variáveis e a representação do modelo utilizando ontologia. Tal modelo consolida o conhecimento sobre evasão e suas causas, direcionando as ações da IES para compreender o tema dentro da instituição.

No capítulo 5 é aplicada a prova de conceito sobre o modelo, realizando a extração dos dados e a publicação em formato aberto. Para a extração do conhecimento se utiliza da análise exploratória e avaliação de algoritmos de classificação identificando a correlação entre as variáveis e a qualidade na predição de evasão.

Por fim, são apresentadas as conclusões sobre o trabalho e as recomendações, no capítulo 6.

2 EVASÃO NO ENSINO SUPERIOR

Nesse capítulo, apresenta-se a revisão da literatura que compõe a base de conhecimento necessária sobre a temática.

2.1 ENSINO SUPERIOR

O Ensino Superior iniciou no Brasil por volta de 1808 com a chegada da família real portuguesa e a criação das escolas de Cirurgia e Anatomia (Salvador), Anatomia e Cirurgia (Rio de Janeiro) e a Academia da Guarda Marinha (Rio de Janeiro). Após estas, outras instituições foram criadas, mas em um ritmo muito lento e que apresentou uma leve melhora a partir de 1850. No final do século XIX existiam apenas 24 instituições de ensino superior que atendiam em torno de 10000 estudantes. Em 1891 a iniciativa privada passa a ter suporte legal pela Constituição da República, sendo que nos 30 anos seguintes o Brasil chegou a 133 instituições (MARTINS, 2002).

Avançando rapidamente na história, a população brasileira passou de 17 milhões de habitantes no início do Século XX para uma estimativa de 210 milhões em 2019 (IBGE, 2011) (IBGE, 2019). Assim como o aumento na população, em 2019 o Brasil passou a ter 2.537 instituições de ensino superior, com um total de 37.962 cursos que oferecem 13.529.101 vagas, e que perfazem 8.450.755 alunos matriculados (INEP, 2019b).

De forma geral, as IES no Brasil estão divididas em duas Organizações Acadêmicas: pública e privada (INEP, 2019). A OECD (2018, p. 78), no estudo realizado a pedido do MEC sobre a educação superior no Brasil, define as categorias do ensino superior brasileiro como:

- Faculdades: representando 83% das instituições. São menores e, na sua maioria, dedicam-se a uma área específica;
- Centros universitários: dedicadas principalmente ao ensino, podendo ter pesquisa e programas de pós-graduação e tem maior autonomia para criar novos programas do que as faculdades;
- Universidades: instituições que abrangem ensino, com a obrigatoriedade de oferecer pesquisa e programas de pós-graduação, além de ter autonomia para criar novos programas.

A LDB (EDUCAÇÃO, 1996), em seu artigo 43, define que os fins das IES no Brasil são:

I. estimular a criação cultural e o desenvolvimento do espírito científico e do pensamento reflexivo;
II. **formar diplomados nas diferentes áreas de conhecimento**, aptos para a inserção em setores profissionais e para a participação no desenvolvimento da sociedade brasileira, e colaborar na sua formação contínua;
III. incentivar o trabalho de pesquisa e investigação científica, visando o desenvolvimento da ciência e da tecnologia e da criação e difusão da cultura, e, desse modo, desenvolver o entendimento do homem e do meio em que vive;
IV. promover a divulgação de conhecimentos culturais, científicos e técnicos que constituem patrimônio da humanidade e comunicar o saber através do ensino, de publicações ou de outras formas de comunicação;
V. **suscitar o desejo permanente de aperfeiçoamento cultural e profissional** e possibilitar a correspondente concretização, **integrando os conhecimentos que vão sendo adquiridos numa estrutura intelectual sistematizadora do conhecimento de cada geração**;
VI. estimular o conhecimento dos problemas do mundo presente, em particular os nacionais e regionais, prestar serviços especializados à comunidade e estabelecer com esta uma relação de reciprocidade;
VII. promover a extensão, aberta à participação da população, visando à difusão das conquistas e benefícios resultantes da criação cultural e da pesquisa científica e tecnológica geradas na instituição;
VIII. atuar em favor da universalização e do aprimoramento da educação básica, mediante a formação e a capacitação de profissionais, a realização de pesquisas pedagógicas e o desenvolvimento de atividades de extensão que aproximem os dois níveis escolares.

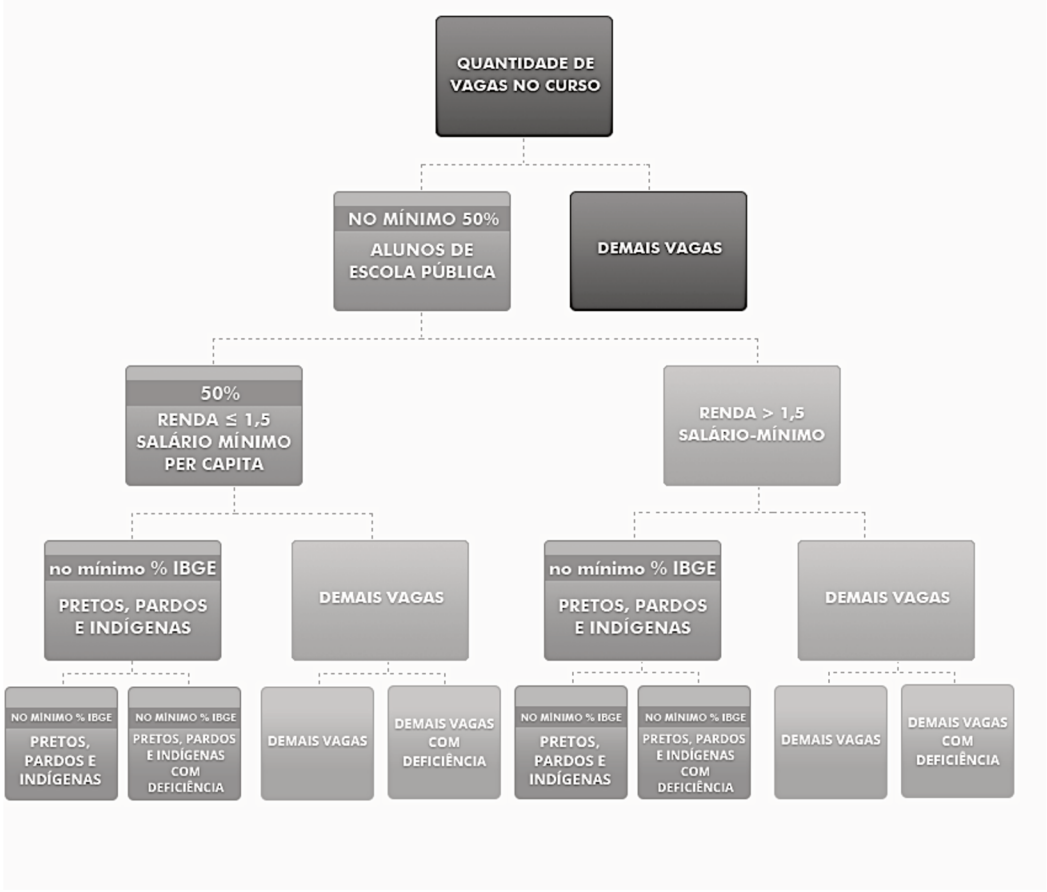
Também na LDB (EDUCAÇÃO, 1996) há a definição dos tipos de cursos superiores a serem oferecidos:

I. cursos sequenciais por campo de saber, de diferentes níveis de abrangência, abertos a candidatos que atendam aos requisitos estabelecidos pelas instituições de ensino, desde que tenham concluído o ensino médio ou equivalente;
II. de graduação, abertos a candidatos que tenham concluído o ensino médio ou equivalente e tenham sido classificados em processo seletivo;
III. de pós-graduação, compreendendo programas de mestrado e doutorado, cursos de especialização, aperfeiçoamento e outros, abertos a candidatos diplomados em cursos de graduação e que atendam às exigências das instituições de ensino;
IV. de extensão, abertos a candidatos que atendam aos requisitos estabelecidos em cada caso pelas instituições de ensino.

Para ingressar no ensino superior no Brasil e na estrutura acima apresentada, predomina o acesso via SISU (pública), FIES (privada) e ProUni (privada) com base na nota do ENEM e os vestibulares. Pensando no acesso ao ensino superior das classes menos favorecidas, alguns movimentos iniciaram seus estudos para que o governo promova políticas neste sentido. Como reflexo disto, destaca-se a Universidade Estadual do Rio de Janeiro (UERJ) que implementou o primeiro programa de cotas em 2003. A partir desta iniciativa, deu início a adoção das cotas nos processos seletivos de outras universidades, ampliando a discussão do tema, e as dúvidas quanto às consequências deste programa de cotas para o ensino superior (GUARNIERI; MELO-SILVA, 2017).

Um dos principais resultados de tais esforços e debates, que duram mais de 10 dez anos, foi a promulgação da Lei 12.711 em 2012 pelo governo brasileiro, chamada de “Lei das Cotas”. Em 2016 foi promulgada a Lei 13.409 que altera a lei anterior, ampliando as cotas para os cursos técnicos de ensino médio. (GUARNIERI; MELO-SILVA, 2017) (VAN PETTEN; DA COSTA ROCHA; BORGES, 2018). Tal lei prevê reserva de vagas para estudantes de escola pública, negros, pardos, indígenas e pessoas com deficiência (PCD), como mostra a Figura 5.

Figura 5 – Sistema de Cotas Brasileiro



Fonte: MEC (2019b)

O governo federal tem aplicado esforços na organização de um sistema único de ensino superior no Brasil. No entanto, as universidades ainda não chegaram a um modelo claro de sistema integrado universitário. Um sistema que possua inter-relação, interdependência de seus componentes, com interatividade entre o desenvolvimento de ciência, tecnologia e cultura e o setor privado e demais instituições governamentais (STALLIVIERE, 2007).

Estudos apontam que o país ainda necessita de ações mais efetivas na busca de ampliar e atender a demanda. Isto se deve ao fato que, mesmo havendo iniciativas recentes, essas levarão muitos anos para alcançar os objetivos e atender os novos cenários do mercado (STALLIVIERE, 2007) (BARROS, 2015).

Considerando superada a fase de ingresso no ensino superior, é necessário manter o discente até a conclusão do curso – conhecida como permanência e êxito –, um desafio enfrentado diariamente pelas IES.

2.2 EVASÃO, PERMANÊNCIA E ÊXITO

Evasão é o fenômeno que leva o discente a abandonar o curso ou, quando sofre uma reprovação, não efetua uma nova matrícula para cursar o ano/semestre seguinte (QEDU, 2019). Na visão da missão das IES na retenção dos alunos, uma das definições sobre a temática é "a capacidade institucional para manter e apoiar os estudantes da admissão até alcançar a graduação com sucesso" (COSTA; GOUVEIA, 2018, p. 163).

Voltando um pouco na história dos estudos sobre o tema, constata-se seu início em debates nos Estados Unidos a partir dos anos de 1950, sendo Tinto um dos pesquisadores com estudos de referência para o assunto (HOFFMAN; NUNES; MULLER, 2019) (ADACHI, 2009). Hoffman, Nunes e Muller (2019) relatam que o abandono – evasão – do curso é algo comum nos primeiros anos de curso superior.

Cislaghi (2008) sintetiza bem a evolução dos estudos relacionados à permanência e êxito entre os anos de 1950 e 2000 (Quadro 2)

Quadro 2 - Evolução dos estudos sobre evasão 1950 - 2000

Década	**Mote**	**Educação Superior e os avanços da permanência de estudantes**
1950	Expansão	Após as Grandes Guerras Mundiais, ocorre uma expansão no número de IES e no contingente de estudantes.
1960	Prevenção da evasão	Surgem situações problemáticas nas IES provocadas pelo grande contingente de estudantes, pela diversidade que os caracteriza e pela inquietação social causada por vários fatores socioculturais. São realizados os primeiros esforços para controlar a evasão com estudos que não se limitem às abordagens estatísticas descritivas.
1970	Construção de teorias	É criada uma base de conhecimento e são propostas as primeiras estruturas teórico-conceituais que vão impulsionar o avanço sistemático da compreensão dos processos relacionados ao fenômeno da evasão.

1980	Administração de matrículas	Crescem os esforços das IES para atrair e manter estudantes. O tema permanência se consolida na área do ensino superior.
1990	Abertura de horizontes	Avançam muito os estudos empíricos para validação das teorias e modelos sobre permanência e evasão. Emerge com força a tendência de considerar o processo de aprendizagem como importante para a permanência dos estudantes.
2000	Tendências	Índices de permanência passam a ser considerados como indicadores importantes a serem utilizados por órgãos oficiais para a alocação de recursos entre IES do setor público. O ensino a distância aparece como elemento novo, dentro e fora das IES. Cresce a importância da formação superior para os profissionais que disputam uma colocação em um mercado de trabalho mais exigente.

Fonte: Cislag (2008, p. 40)

É notório que, à medida que o fenômeno vem se intensificando, a preocupação e os estudos também, da mesma forma, se intensificam ao longo das décadas. Denota-se ainda que "uma das consequências da expansão das universidades brasileiras foi o aumento do número de estudantes evadidos" (JUNIOR, *et al.*, 2016, p. 488). Mesmo o ensino superior sendo considerado um diferencial para colocação no mercado de trabalho a partir dos anos 2000, o que se entende como um motivador à permanência, tal fenômeno de evasão persiste.

Mais recentemente estudos apontam que "não basta ingressar no Ensino Superior, é preciso que os estudantes tenham condições de cursá-lo e de concluí-lo" (MENEGHEL, 2018, p. 343).

Entender tais condições que levam os estudantes a decidir pela evasão passa pela análise das suas causas. Tal análise está relacionada a um domínio e seus dados específicos para os quais se deseja averiguar o fenômeno que ocorre dentro de um universo definido.

> "a evasão pode ser medida em uma instituição de ensino superior, em um curso, em uma área de conhecimento, em um período de oferta de cursos e em qualquer outro universo, desde que tenhamos acesso a dados e informações pertinentes" (SILVA FILHO, *et al.*, 2007, p. 644).

No que tange ao acesso a dados e informações pertinentes, o CENSUP 2018 apresenta – Tabela 4 – os dados referentes a ingressantes e concluintes para o Brasil.

Tabela 4 – Relação entre ingressantes e concluintes em 2018

Tipo	Ingressantes	Concluintes (% em relação aos ingressantes)
Pública	458.587	218.032 (47.5%)
Privada	1.334.124	435.322 (32,6%)

Fonte: INEP (2019).

Não obstante o levantamento realizado pelo INEP (2018), em 2017 aponta que foram ofertadas mais de 10,7 milhões de vagas, s com um total de 90% das vagas ofertadas na rede pública preenchidas. No entanto, em 2017 havia 99 mil vagas remanescentes na rede pública, e dessas 70% não foram preenchidas. Isto está relacionado com fatores que levam à evasão e, consequentemente, ao aumento das vagas remanescentes não preenchidas.

Diante dos dados no domínio do Brasil, são apresentadas as pesquisas e identificadas e analisadas as principais causas de evasão apontadas pelas publicações de 2015 até 2019.

2.3 CAUSAS QUE LEVAM À EVASÃO

O estudo de tal temática "tem sido instigador aos pesquisadores de ramos bastante distintos da ciência, uma vez que se trata de um fenômeno universal, presente nos mais diversos cursos de graduação" (JUNIOR; REAL, 2017, p. 397).

Em sua maioria foram encontradas causas no âmbito mundial e não só no Brasil. Entretanto, dentre os artigos selecionados para este trabalho, há destaque para estudos que mostram uma aderência suficiente entre os contextos nacional e internacional.

> Os resultados de estudos anteriores, em contextos diferentes do brasileiro, são semelhantes aos obtidos neste estudo, apesar de algumas diferenças em determinadas variáveis. Essa situação mostra que é possível que a evasão e a retenção tenham características semelhantes, mesmo em diferentes contextos culturais e sociais (COSTA, BISPO; PEREIRA, 2018, p. 83).

Carreira e Lopes (2019) realizaram estudos em universidade portuguesa segmentando os alunos tradicionais e não tradicionais – os que não trabalham e os que trabalham respectivamente.

Dentre os fatores positivos para os tradicionais (não trabalham), "à atribuição de bolsas, parece ser eficaz na promoção do desempenho acadêmico entre os estudantes

tradicionais, pois [...] diminui a probabilidade de evasão, embora não seja relevante para os estudantes trabalhadores" (CARREIRA; LOPES, 2019, p. 11)

Para o grupo dos não tradicionais (aqueles que trabalham), é fundamental que as instituições de ensino promovam políticas logo no início da vida acadêmica desses, como cursos preparatórios para o nível superior antes mesmo do início da trajetória acadêmica. (CARREIRA; LOPES, 2019, p. 13).

Aplicando modelo de regressão logística em uma das maiores universidades do Chile, Venegas-Muggli (2019) analisa as taxas de abandono com base nas características sociodemográficas para os estudantes não tradicionais – idade adulta. "Em relação às possíveis explicações para as taxas de abandono escolar, três dimensões das variáveis sociodemográficas foram definidas a partir da revisão da literatura: condições familiares / demográficas, situação socioeconômica e estruturas institucionais." (VENEGAS-MUGGLI, 2019, p. 13).

Pessoas com filhos, que trabalham, frequentaram ensino médio para adultos e estão matriculados em programas com duração superior a 2 anos são os fatores predominantes que levam ao abandono. No entanto, não foram encontradas relações entre questões socioeconômicas e evasão, sendo uma das possíveis causas o fato de tais indivíduos, por estarem em idade adulta, já possuírem seus projetos de vida. (VENEGAS-MUGGLI, 2019, p. 13).

Os estudos mostram a importância de realizar a análise considerando os fatores externos à instituição de ensino, como a vida acadêmica pregressa, os fatores socioeconômicos, entre outros.

Truta, Parv e Topala (2018) efetuaram estudos com diversas dimensões do comprometimento acadêmico como fator de abandono de estudantes no primeiro ano. "Os resultados da análise de regressão mostraram que a falta de dedicação é um forte preditor da intenção de abandono, em todos os modelos testados" (TRUTA; PARV; TOPALA, 2018, p. 7).

Observa-se que os alunos que têm entusiasmo nos estudos, promovido muitas vezes por desafios, se movem para atingir seus objetivos acadêmicos e, assim, reduzindo as chances de evasão. Na visão do engajamento, um ponto específico foi caracterizado no contexto dos alunos inseridos em ambientes familiares com baixa escolaridade. Esses possuem níveis de engajamento maiores que os demais (TRUTA; PARV; TOPALA, 2018, p. 8).

Estudos realizados em IES da França apontam uma nova abordagem sobre evasão – principalmente no primeiro ano – apontando como fatores a serem considerados as questões sociais dos estudantes e estruturais dos cursos oferecidos. E que o abandono no primeiro ano é, na verdade, um efeito constante e estrutural característico do contexto no qual o discente está inserido. Concluiu-se que os discentes acabam fazendo uso desse primeiro ano nos cursos que estão matriculados para "seus próprios fins, ficar tempo esperando ou sendo socializado na instituição, descobrir novas possibilidades acadêmicas ou profissionais" (BODIN; ORANGE, 2018, p. 141).

Stoessel *et al.* (2015) pesquisou sobre as questões de evasão na Alemanha com base em cinco características sociodemográficas: sexo, idade, situação dos pais, ser imigrante e exercício de atividade remunerada. "Os estudantes empregados em período integral enfrentam desafios consideráveis acima e além das demandas acadêmicas" (STOESSEL, *et al.*, 2015, p. 242). A insatisfação com o programa também é apontada em uma análise secundária quanto ao risco de abandono (STOESSEL, *et al.*, 2015).

Kamal e Ahuja (2019) analisaram as dimensões que estão relacionadas à propensão ao abandono utilizando-se de técnicas de extração do conhecimento com ferramentas estatísticas. Com isto, obtiveram uma ferramenta com 98,5% de acurácia para a massa de dados disponível.

> Para melhorar o desempenho acadêmico dos alunos, além da retenção de alunos, as instituições acadêmicas e as universidades terão que trabalhar em fatores importantes, como questões familiares, dados acadêmicos anteriores e fatores sócio comportamentais, pois desempenham um papel vital no desempenho escolar dos estudantes (KAMAL; AHUJA, 2019, p. 780).

Com base na análise de dados do semestre anterior, o estudo identifica que o aproveitamento menor que 60%, dificuldades com notas, não residir em área urbana, ter histórico de repetição de semestre, acontecimentos importantes na família e o hábito de consumo de álcool, elevam as chances de desistência. (KAMAL; AHUJA, 2019, p. 778).

Torres-Coronas e Vidal-Blasco (2019) analisaram os modelos de ensino híbridos – que contém parte do programa presencial, parte a distância – sob a perspectiva da universidade, dos alunos e do corpo docente. Sendo assim, existindo a flexibilidade do ensino a distância aliada à convivência social entre os alunos nas atividades presenciais, tal cenário contribui para a permanência. "Modelos mistos orientam melhor os alunos que não são tão organizados, o que ajuda a reduzir a taxa de evasão" (TORRES-CORONAS; VIDAL-BLASCO, 2019, p. 336).

Além disso, a convivência social propicia uma maior integração e uma relação de confiança entre os alunos que irá refletir no nível de aprofundamento na construção de novos conhecimentos na academia.

Truta, Parv e Topala (2018) deixaram de fora – em um primeiro momento – as variáveis relacionadas ao desempenho acadêmico para a retenção. No entanto, "a inclusão dessas variáveis parece ser de grande interesse não apenas para a previsão do desempenho individual, mas também para a previsão da taxa de graduação no nível universitário" (TRUTA; PARV; TOPALA, 2018, p. 8).

Como proposta para mitigar os problemas relacionados à evasão, sugerem que a instituição de ensino pode intervir na questão "tempo" dos estudantes, melhorando o planejamento de atividades, inclusive extracurriculares, levando em consideração fatores como a distância física entre ambientes de aprendizagem, flexibilidade de horários de aprendizagem e de áreas de trabalho como laboratórios. "Uma gestão responsável do tempo do aluno deve ser uma meta das universidades em seus esforços para consumir recursos com responsabilidade ao preparar os alunos capazes de enfrentar os desafios da mudança através do aprendizado, de acordo com os princípios da sustentabilidade" (TRUTA; PARV; TOPALA, 2018, p. 11).

Costa, Bispo e Pereira (2018, p. 74) analisaram um grupo de alunos de curso de Administração no Brasil entre 2004 e 2013 e apontam como os principais fatores que influenciam na evasão: a duração do curso, o desempenho do aluno, o gênero, as reprovações e os trancamentos. "Do ponto de vista estratégico, destaca-se a relevância de políticas públicas de acesso ao ensino superior, bem como o desenvolvimento de ações das universidades para atrair e manter os alunos nos cursos de graduação" (COSTA, BISPO; PEREIRA, 2018, p. 83). Quando o olhar dos pesquisadores se volta para o operacional, identifica-se que existe a necessidade de revisão das práticas pedagógicas de ensino e avaliativas por parte dos docentes como um fator importante para aumentar a retenção dos alunos (COSTA, BISPO; PEREIRA, 2018).

Saccaro, França e Jacinto (2019) avaliaram a evasão utilizando diversas técnicas estatísticas de regressão linear e apontaram as bolsas de estudo, o apoio estudantil, participar atividades extras – como estágios, projetos de pesquisa e projetos de extensão –, além da forma de ingresso baseada na nota do ENEM, como fatores relevantes para a conclusão com sucesso.

Já questões como idade mais avançada, estar trabalhando em tempo integral, dificuldades financeiras, baixo nível de organização pessoal afetam a permanência, levando o aluno a evadir.

Quanto a políticas para manter e garantir o sucesso acadêmico do aluno, "[...] os estudantes mais integrados com o ambiente acadêmico por meio da realização de atividades remuneradas e não remuneradas, e os que recebem benefícios financeiros para auxiliar com os custos do curso evadiram menos" (SACCARO; FRANÇA; JACINTO, 2019, p. 367).

Em estudo abrangente sobre modelos de análise de retenção (permanência), foram apontados vários fatores a serem considerados. Os resultados acadêmicos, as atividades acadêmicas, integração acadêmico/social, o compromisso com a instituição, os objetivos do programa, as dificuldades financeiras, o trabalho, a dedicação à família, bem como a vida pregressa são pontos de relevância (COSTA; GOUVEIA, 2018, p. 173-174).

Diogo *et. al* (2016) utiliza uma abordagem diferenciada e focada nos coordenadores de cursos superiores. Eles destacam que os coordenadores apontam como a principal causa para a evasão "as ideias equivocadas dos alunos sobre a formação; a falta de clareza acerca das características que constituiriam diferentes campos de atuação, confundindo cursos com grades curriculares semelhantes; e a incompatibilidade vocacional" (DIOGO, *et al.*, 2016, p. 136).

Além dos fatores de desconhecimento prévio do curso, a incompatibilidade de nível de exigência do curso para o aluno, condições socioeconômicas desfavoráveis que levam à necessidade de estar trabalhando, pouco tempo dedicado ao curso e a falta de interesse dos estudantes com o curso devem ser considerados (DIOGO, *et al.*, 2016, p. 146).

Campos (2017) realizou estudos focados na evasão dos alunos que são atendidos pelas cotas sociais e ações afirmativas com o intuito de identificar o impacto destas sobre o nível de evasão. A possibilidade de o candidato escolher mais de uma instituição no SISU, a insatisfação com o curso e a distância da IES – principalmente para o ingressante pelo SISU, visto que o aluno pode vir de qualquer lugar do Brasil – foram identificados como fatores que levam à evasão. No caso, os alunos ingressantes por cota e o fato de serem originários da região da IES tem um efeito contrário, reduzindo a evasão (CAMPOS, *et al.*, 2017, p. 39).

Utilizando técnicas de KDD, mais especificamente da Rede Bayesiana, Vasconcelos (2018) busca identificar os fatores que levam à evasão como forma de representação do conhecimento. Fatores com trancamentos ao longo do curso e o índice acadêmico foram identificados como relevantes.

Segundo outro estudo analisado, o tema é “complexo e pode ser explicado por uma série de fatores anteriores ao ingresso e de desempenho acadêmico” (JUNIOR, *et al.*, 2016, p. 508). Tal estudo apresenta como resultado sete causas relevantes: cotista, região de origem, meio de comunicação para se manter atualizado, participação em projetos de pesquisa, assistência estudantil, estágio e histórico de reprovação.

As várias causas encontradas estabelecem as dimensões na permanência e êxito, caracterizando os motivadores da evasão. O Quadro 3 apresenta os motivadores e seu impacto na permanência e êxito.

Quadro 3 - Motivadores e seu impacto na permanência e êxito

Motivador	**Impacto**
1. Atribuição de Bolsas	Positivo
2. Presença de Cursos preparatórios / nivelamento	Positivo
3. Características sociodemográficas	
a. Condições familiares (com filhos)	Negativo
b. Situação socioeconômica ruim	Negativo
c. Estruturas institucionais (duração até 2 anos)	Positivo
d. Vida acadêmica pregressa	Relevante
e. Gênero	Irrelevante
f. Idade	Relevante
g. Situação dos pais com baixo nível escolar	Negativo
h. Ser imigrante	Negativo
i. Trabalho	Negativo
4. Comprometimento acadêmico	Positivo
5. Entusiasmo nos estudos proporcionados por desafios	Positivo
6. Contexto familiar com baixa escolaridade	Negativo
7. Boa estrutura do curso	Positivo
8. Fatores sócio comportamentais (dependência química, violência doméstica, etc.)	Negativo
9. Residência em área urbana	Positivo

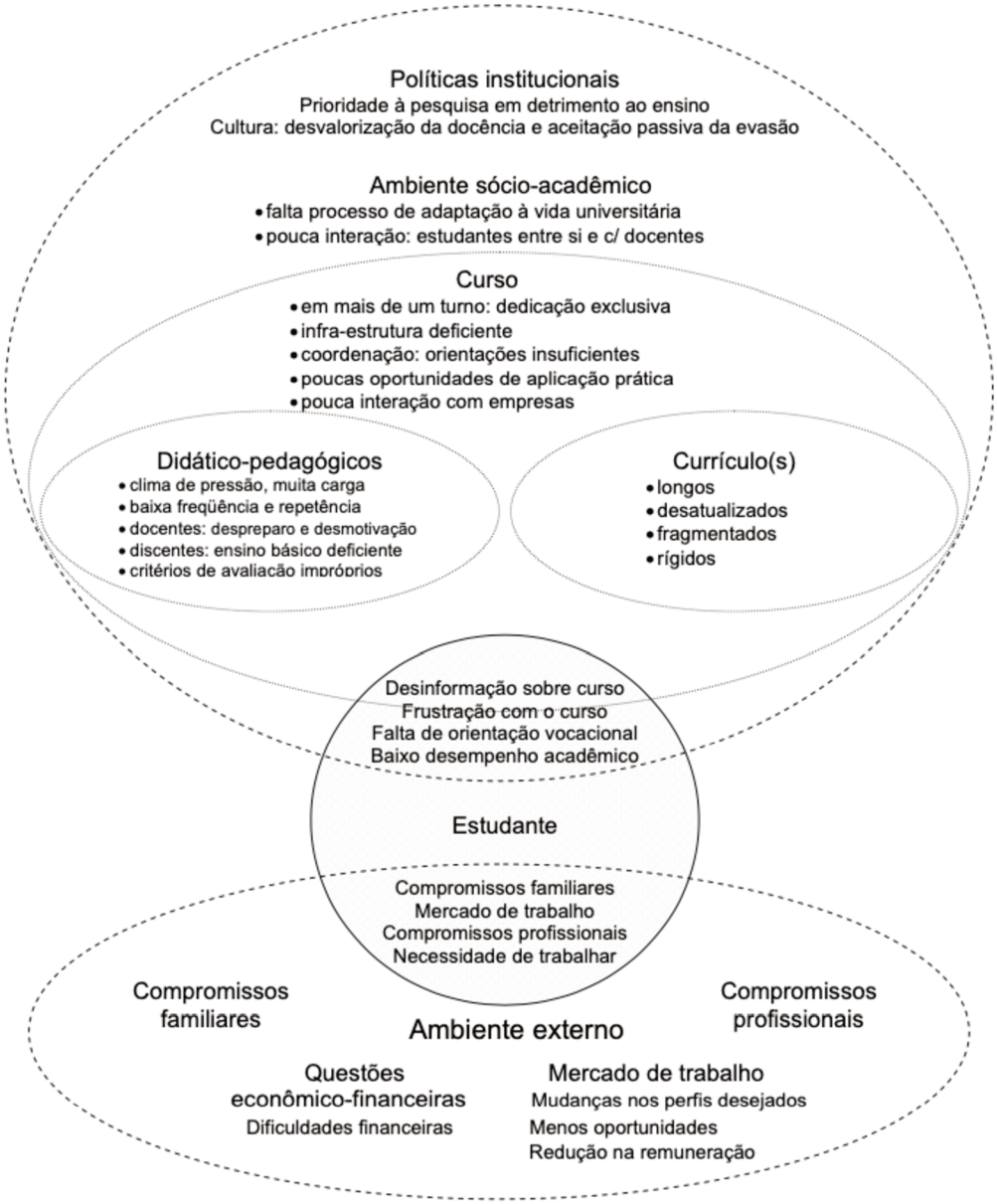
10. Histórico de reprovação/repetição	Negativo
11. Falta de organização pessoal para a vida acadêmica	Negativo
12. Convivência social e integração	Positivo
13. Relação de confiança entre os alunos	Positivo

Fonte: Do Autor

No tocante ao impacto, identifica-se que alguns têm impacto diretamente positivo à permanência e outros negativo. Alguns têm ou não relevância para o tema como a vida acadêmica pregressa que, dependendo de como ocorreu, pode ser positivo ou negativo para o contexto analisado. Contudo, tais características desses grupos motivadores devem ser consideradas nos estudos.

Cislaghi (2008), em seus estudos sobre evasão nas IES do Brasil, compilou o resultado de suas pesquisas em uma visão macro, conforme é apresentado na Figura 6.

Figura 6 – Visão macro das causas que levam à evasão nas IES brasileiras.



Fonte: Cislaghi (2008, p. 34)

Pode-se perceber que o estudante, como protagonista da questão relacionada à evasão, está inserido em um ambiente com vários estímulos que podem causar sua evasão. Desde questões macro como políticas institucionais, passando por fatores externos ao ambiente acadêmico que têm seu reflexo – positivo ou negativo – no desempenho acadêmico. Logo, tudo isso influencia diretamente não só nas aprovações e conclusões de curso, mas nas reprovações, trancamentos e possíveis evasões.

Neste cenário complexo o presente estudo aplica a Engenharia do Conhecimento para propiciar a compreensão e direcionamento das ações para mitigar os riscos.

2.4 CONSIDERAÇÕES

A evolução da oferta de vagas de ensino superior no Brasil, juntamente com o crescimento populacional, teve um forte crescimento a partir da liberação da iniciativa privada para oferecer cursos de ensino superior. Paralelo à expansão promovida pela iniciativa privada as instituições públicas de ensino também cresceram e contribuíram para o aumento da oferta. Em 2019 havia 2537 instituições, que oferecem mais de 37 mil cursos com mais de 13 milhões de vagas. Mesmo assim, não há vagas para todos, o que leva à necessidade de processos seletivos para ingresso nas instituições.

Em paralelo a isto, as dívidas históricas quanto à questões raciais, pessoas com deficiência e oriundos do ensino em escolas públicas, levaram a sociedade a mobilizar-se para propor mecanismos de correção de tais obstáculos. Como resultado, os processos seletivos passaram a contemplar cotas para cada um dos grupos historicamente menos favorecidos.

Pela escassez, pode-se perceber a existência de concorrência por vagas. Mas, mesmo havendo concorrência, algo ocorre em muitos casos que leva o discente a desistir do curso e evadir. Tito é um dos precursores nas pesquisas sobre o fenômeno da evasão (HOFFMAN; NUNES; MULLER, 2019) (ADACHI, 2009). Percebeu-se a partir da década de 1960 a iniciativa de esforços para mitigar a evasão, até que por volta dos anos 2000 os índices de permanência tornam-se indicador importante para as IES, também tem-se o EaD como um novo elemento no cenário educacional e a formação superior ganha mais importância no mercado de trabalho (CISLAGHI, 2008).

Denota-se, então, a importância de dar condições aos estudantes para que possam cursar até o fim. E, para tanto, é necessário compreender o que causa a evasão. Para tal, vários pesquisadores realizaram estudos sobre visões e grupos diferentes. De tais estudos, em comum, apontam as causas identificadas de evasão dentro de cada população pesquisada.

Existem causas que influenciam positivamente (contribui para a permanência) negativamente (leva à evasão) ou têm certo grau de relevância ou são irrelevantes. Quanto à relevância, pode-se exemplificar que a idade pode ser relevante quando influencia negativamente a permanência, por exemplo, um estudante de 50 anos pode não ter tempo para

estudar. Por outro lado, um outro estudante de 50 anos, por ser mais experiente, vai dar a devida importância à conclusão do curso, o que influencia positivamente. O gênero se mostrou irrelevante nos estudos.

É a partir destes conhecimentos sobre evasão que o modelo de EC é concebido e, para isso, se faz necessário conhecer as metodologias e ferramentas de EC. Dessa forma elas são apresentadas no capítulo seguinte.

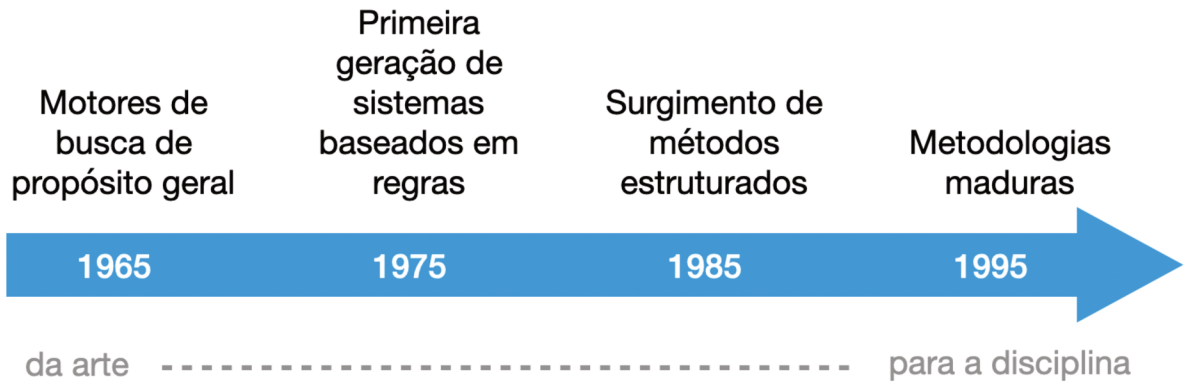
3 ENGENHARIA DO CONHECIMENTO

A Gestão do Conhecimento (GC) é uma disciplina que surgiu com a Era da Informação, como ocorreu com as engenharias elétrica e mecânica, durante a Revolução Industrial. Para a GC o conhecimento é todo o conjunto de dados e informações que as pessoas trazem para uso prático em ação, a fim de realizar tarefas e criar novas informações (SCHREIBER, *et al.*, 1999).

Dentro da Gestão do Conhecimento, existem alguns instrumentos que são aplicados sobre a base de conhecimentos organizacionais na extração desses e de novos. "Formalmente, um Instrumento de Gestão do Conhecimento consiste em um conjunto alinhado e claramente definido de medidas organizacionais, recursos humanos e Tecnologias de Informação e Comunicação, com o propósito de inferir na base de conhecimento organizacional" (RAUTENBERG; TODESCO; STEIL, 2011, p. 31).

Nesse contexto, o termo Engenharia do Conhecimento (EC) foi cunhado, em sua origem, dentro da linha de tempo e evolução da Tecnologia da Informação e Comunicação, mais especificamente da Inteligência Artificial. A EC nasceu com intuito de conceber, desenvolver e implementar sistemas especialistas (STUDER; BENJAMINS; FENSEL, 1998). A Figura 7 resume a evolução ocorrida ao longo do tempo.

Figura 7 – A evolução da EC



Fonte: Schreiber *et al.* (1999, p. 14)

Quando há necessidade por parte da Gestão do Conhecimento no uso de agentes inteligentes, suportados por tecnologia da informação e comunicação dentro dos Instrumentos

de Gestão do Conhecimento e as tarefas intensivas em conhecimento, ocorre a conexão entre a Gestão e a Engenharia do Conhecimento, que a complementa (RAUTENBERG; TODESCO; STEIL, 2011). Parte da EC está focada na Descoberta de Conhecimento. Conhecimento é para a EC, assim como a elétrica e mecânica são para as suas respectivas.

O objetivo da disciplina Engenharia do Conhecimento é semelhante ao da Engenharia de Software: transformar o processo de construção de sistemas de uma arte em uma disciplina de engenharia. Isso requer a análise do próprio processo de construção e manutenção, além do desenvolvimento de métodos apropriados, linguagens e ferramentas especializadas para o desenvolvimento de Sistemas Baseados em Conhecimento - SBC.

Para atingir o seu propósito, a EC define metodologias e ferramentas que permitem modelar e adquirir o conhecimento, formalizando-o e tornando-o reutilizável independentemente de pessoas (HASSLER, *et al.*, 2016).

Tal arcabouço composto por metodologias e ferramentas dá suporte à EC no processo de descoberta de conhecimento.

3.1 DESCOBERTA DO CONHECIMENTO

O processo aplicado sobre dados não estruturados, semiestruturados e estruturados com o objetivo de verificar a hipótese de usuários, bem como novos padrões é denominado Descoberta do Conhecimento. A descoberta de padrões pode ser suportada por análise histórica de dados para prever um comportamento futuro ou a representação de padrões identificados na análise de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p. 7).

Para realizar tal processo de extração do conhecimento a EC tem o suporte de diversas metodologias e ferramentas para cada tipo de dado. Neste trabalho, é abordado o processo de Descoberta de Conhecimento em Banco de Dados, do inglês *Knowledge Discovery in Databases* (KDD) que é utilizado para a prova de conceito do modelo proposto, extraindo novos conhecimentos sobre evasão que realimentam o modelo.

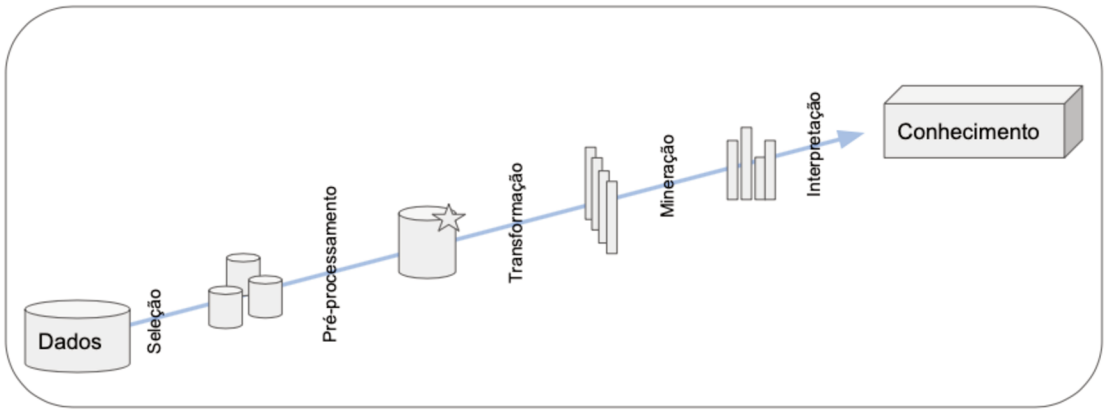
KDD é a área relacionada ao processo de descoberta de conhecimento em bases de dados estruturadas, possui técnicas e mecanismos adequados para tal. (DA SILVA, *et al.*, 2018, p. 614) . “Descoberta de conhecimento em bancos de dados é o processo não trivial de

identificar padrões válidos, novos, potencialmente úteis e, em última análise, compreensíveis nos dados" (FAYYAD, 1996, p. 21).

Os termos KDD e mineração têm estreita relação. No entanto, Fayyad, Piatetsky-Shapiro e Smyth (1996) caracterizam que KDD é o processo geral de descoberta de conhecimento e mineração é uma parte desse processo, no qual as etapas adicionais de preparação, seleção, limpeza, incorporação de conhecimento prévio e interpretação dos resultados são essenciais para garantir um conhecimento útil derivado dos dados.

A Figura 8 apresenta o processo que passa pela seleção das fontes de dados de origem, o pré-processamento, a transformação, o armazenamento, a mineração e a interpretação, tendo como produto resultante o conhecimento descoberto.

Figura 8 – Processo KDD



Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 41), adaptação nossa.

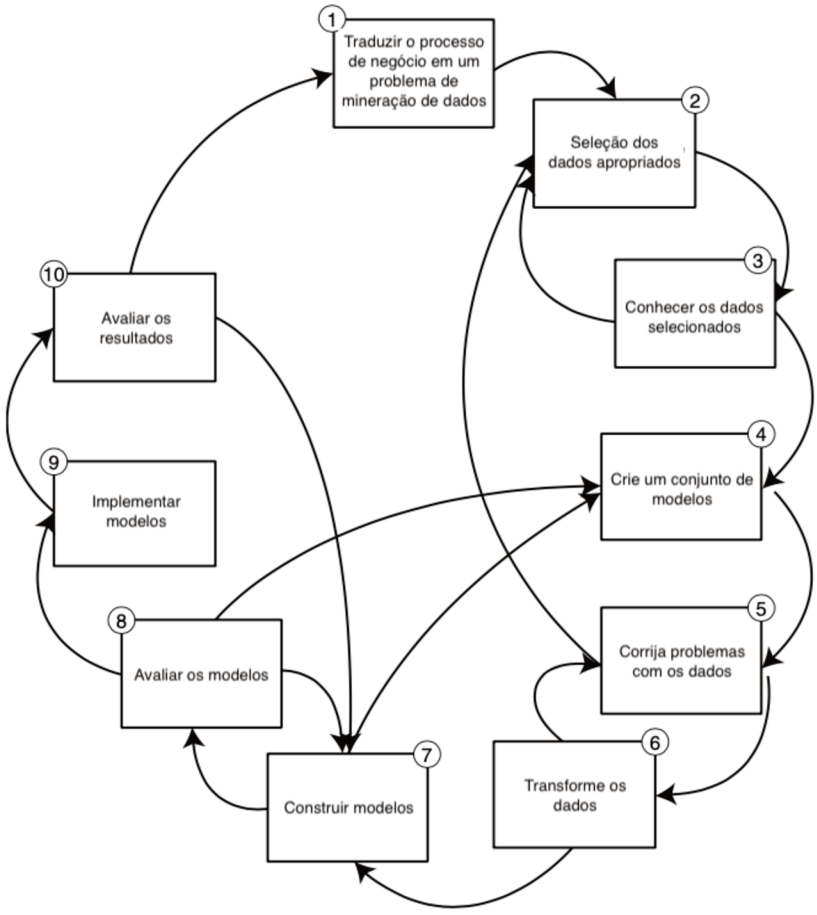
Pelo fato de abranger técnicas que estão além de uma disciplina de aprendizagem de máquina e por convergir vários paradigmas da computação, ela é considerada uma área multidisciplinar (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) (ROMERO; VENTURA; GARCÍA, 2008).

Desta forma, KDD é aplicado sobre bases de dados estruturadas, como bancos de dados relacionais, planilhas e arquivos no formato CSV, a fim de identificar e extrair conhecimentos de um determinado domínio de interesse.

Berry e Linoff (2004) propõem uma visão mais detalhada do processo KDD dividido em onze passos – Figura 9 – e reforçando que não se trata necessariamente de um processo totalmente linear, no qual um passo se inicia ao final da execução total do anterior.

Pelo contrário, tal comportamento linear é indesejado, devendo-se considerar a interação entre passos e um ciclo global recorrente.

Figura 9 – Passos do KDD



Fonte: Berry e Linoff (2004, p. 55)

3.1.1 Traduzir em um problema de mineração

Berry e Linoff (2004) definem que o processo passa inicialmente por traduzir problema de negócio em um problema de mineração de dados. Nele se define qual método será utilizado para a descoberta do conhecimento buscando a previsão ou descrição. A previsão trata da utilização de variáveis e dados conhecidos para prever valores desconhecidos ou valores futuros de interesse. Já a descrição trata da busca por padrões interpretáveis por humanos e que descrevem os dados em questão.

Para alcançar tais objetivos, os métodos disponíveis são: classificação, regressão, agrupamento, sumarização, modelagem de dependência e a detecção de mudança e desvio (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

- Classificação trata da busca por entendimento que permitirá classificar algo. Como exemplo, classificar os carros por tipo de veículo como SUV, passeio, de carga, entre outros, depende de entendimento das características que possibilitam atribuir tais classes;
- Regressão é o aprendizado que possibilita a predição de um valor real com base em um conjunto de dados. Avaliar os exames de um paciente e predizer as chances dele se curar, avaliar a previsão do tempo com base em dados de temperatura, pressão, humidade, entre outros, são alguns exemplos da regressão;
- Agrupamento é a busca por um conjunto finito de categorias que possibilita descrever os dados em análise. Tomando novamente uma sala de aula, dado o conjunto de alunos que compõem a turma, pode-se encontrar uma função e agrupá-los por estatura entre alta, mediana e baixa, sendo que, para a população analisada, o mais alto da turma tem 1,50 metros, por exemplo;
- Sumarização busca encontrar uma descrição para o subconjunto de dados que são frequentemente utilizados durante a análise exploratória de dados. Tabular os dados por média e desvio padrão são exemplos simples de sumarização;
- Modelagem de dependência busca identificar quais são as dependências significativas existentes entre as variáveis analisadas. Analisar a correlação e covariância entre as variáveis são exemplos de modelagem de dependência;
- Detecção de mudanças e desvios tenta descobrir, com base no comportamento das variáveis as mudanças e desvios ocorridos nelas referentes ao que foi normatizado anteriormente.

3.1.2 Selecionar dados apropriados

O segundo passo trata da seleção dos dados mais apropriados para o problema.

> No melhor dos mundos possíveis, os dados necessários já estariam residentes em um *data warehouse* corporativo, limpos, disponíveis, historicamente precisos e atualizados com frequência. Na verdade, ele está mais frequentemente espalhado em uma variedade de sistemas operacionais em formatos incompatíveis em computadores que executam sistemas operacionais diferentes, acessados por meio de ferramentas de desktop incompatíveis (BERRY; LINOFF, 2004, p. 60).

Portanto, caso os dados não estejam disponíveis em um *dataware house*, parte-se para as bases de dados disponíveis pelos vários sistemas da organização que podem contribuir com dados e informações associados ao problema modelado. Para extrair os dados, pode-se utilizar o ETL (*Extract, Transformation, and Load*). ETL são ferramentas que permitem extrair (*extract*), transformar (*transformation*) e carregar (*load*) dados de diversas fontes para um repositório central, como representado na Figura 10. Tais ações compões as três primeiras fases do KDD: seleção dos dados, pré-processamento e transformação, disponibilizando as informações para as próximas fases (vide fases do KDD na Figura 8).

Figura 10 – Fases de extração, transformação e carga

Fonte: Fayyad (1996), adaptação nossa.

A tarefa de ETL permite realizar a extração dos dados de várias origens. “As ferramentas de extração, transformação e carga resolvem o problema de coletar dados de sistemas diferentes, fornecendo a capacidade de mapear e mover dados dos sistemas de origem para outros ambientes” (BERRY; LINOFF, 2004, p. 487). Sendo assim, aplica padrões aos

dados visando garantir a qualidade e a consistência dos dados de fontes diferentes, para que possam ser utilizadas juntas posteriormente, permitindo, então, entregar os dados prontos para os desenvolvedores de aplicações que criam as soluções para o usuário final (KIMBALL; CASERTA, 2004, p. xxi).

Houve um tempo em que programadores eram responsáveis pelas tarefas como movimentação e limpeza de dados para posterior uso, escrevendo linhas de códigos para tal. No entanto, tais códigos eram específicos para aquele cenário tornando-os críticos à medida que acrescem o número de sistemas envolvidos e ocorrem mudanças nos sistemas que originam os dados. Com o surgimento de ferramentas ETL, não há preocupação com a escrita de códigos para estas tarefas, mas sim em descrever de onde vem os dados, o que ocorre com eles durante a transformação, isto sendo realizado com códigos baseados em metadados – ao invés de código fonte – que facilitam o entendimento dos usuários. (BERRY; LINOFF, 2004).

3.1.3 Conhecer os dados

Deve-se mergulhar nos dados e entender o seu comportamento. Naturalmente, à medida que se mergulha nos dados, descobre-se problemas com a sua qualidade, tais como inconsistência ou falta de valores, que podem interferir no modelo (BERRY; LINOFF, 2004).

Pode-se, por exemplo, fazer uso de histograma para visualizar a distribuição da variável e avaliar criticamente se aquelas informações de distribuição fazem sentido. Uma instância de pessoa com idade de 200 anos, uma altura e uma pessoa com poucos 20 centímetros, a área de um terreno de casa do tamanho de uma cidade, entre outros, nos leva a crer que há algo de errado e necessita de análise detalhada daquela área do histograma.

3.1.4 Criar modelos intermediários

Aqui todos os dados são considerados, sendo que algum conjunto desses pode ser utilizado para encontrar padrões, outros para verificar a estabilidade, ou o desempenho, por exemplo (BERRY; LINOFF, 2004). Para modelos de predição, faz parte do processo a definição dos conjuntos de dados de treino e teste. Deve-se cuidar com tal definição de conjuntos pois, para avaliar o desempenho de um classificador, precisamos avaliar sua taxa de erro sobre instâncias que não foram aplicadas para o treino (WITTEN; FRANK, 2002).

3.1.5 Resolver os problemas encontrados nos dados

Tendo os modelos intermediários e conhecendo a qualidade dos dados, é necessário (passo 5) resolver os problemas encontrados nos dados durante o passo 3 – e não somente agora. Todos os dados possuem alguma "sujeira", portanto, todos têm algum problema. Sendo assim, cada variável deve ser analisada de forma a não trazer problemas para as técnicas de mineração escolhidas. Por exemplo, a falta de valores pode não ser um problema para classificadores do tipo árvore de decisão, mas para redes neurais pode causar todo o tipo de problema (BERRY; LINOFF, 2004). Basicamente, é a remoção de ruído quando necessário, a coleta das informações que possibilite modelar ou contabilizar o ruído, definir estratégias para tratar dados ausentes e contabilizar informações de sequência de tempo e mudanças conhecidas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

3.1.6 Transformar os dados

Com os dados "limpos", passa-se a transformar estes dados para trazer algum conhecimento, agregando outras informações, removendo dados com valores discrepantes, transformando classes categóricas, entre outros. Tais tarefas preparam os dados para análise (BERRY; LINOFF, 2004). Recuperar latitude e longitude de dois endereços para adicionar posteriormente a distância entre eles é um exemplo de adição de informação baseado nos dados disponíveis inicialmente. Uma das formas de trazer algum conhecimento é a identificação da existência de algum grau de correlação entre as variáveis.

Correlação é uma medida que permite identificar como a mudança em uma variável afeta uma segunda, sendo medida em valores entre -1 a 1. Quando a variável "A" tem correlação positiva com a variável "B" (exemplo 0,45), significa que existe uma correlação positiva entre elas. Ou seja, à medida que o valor de A cresce, B também cresce. Do contrário, um valor negativo (exemplo -0,60) informa uma correlação negativa (uma cresce e a outra diminui ou vice-versa). Correlações próximas de zero significa que não foi identificada qualquer correlação entre elas (BERRY; LINOFF, 2004).

Figueiredo Filho e Silva Júnior *apud.* Bohen (2009), apresentam uma classificação para o nível de correlação entre duas variáveis de acordo com a faixa de coeficiente de

correlação. O Quadro 4 mostra tais faixas, sendo que se aplicam os mesmos para os valores negativos. Ou seja, entre 0,10 e 0,29 é considerado pequeno, o mesmo vale para entre -0,10 e -0,29.

Quadro 4 – Nível de correlação por faixa de valor

Faixa	**Nível de correlação**
0,10 a 0,29	Pequeno
0,30 a 0,49	Médio
0,50 a 1,00	Grande

Fonte: adaptado de Figueiredo Filho e Silva Júnior (2009, p. 119).

Uma forma de representar visualmente a correlação entre variáveis é no formato de matriz, na qual cada variável é apresentada tanto na linha quanto na coluna. Em cada intercessão entre as duas variáveis é apresentada a medida de correlação entre elas. A diagonal da matriz – que representa o cruzamento entre linha e coluna de uma mesma variável – normalmente é preenchida pelo valor 1. Isto ocorre porque a correlação da variável com ela mesmo é sempre uma correlação positiva máxima.

3.1.7 Construção do modelo

A construção dos modelos varia de acordo com a mineração, se é direcionada ou não. Na mineração não direcionada, a busca por relacionamentos entre registros com utilização de técnicas de clusterização é um dos exemplos. "A construção de modelos é a única etapa do processo de mineração de dados que foi verdadeiramente automatizada por um software moderno de mineração de dados" (BERRY; LINOFF, 2004, p. 77). Softwares como *Weka*[5] e *sklearn*[6] são exemplos de automatização disponíveis que já possuem todo o arcabouço associado ao conhecimento necessário para a aplicação de alguns modelos, bastando o engenheiro do conhecimento ou outra pessoa conhecedora do tema saber aplicar os recursos de forma correta.

[5] https://scikit-learn.org/stable/
[6] https://www.cs.waikato.ac.nz/ml/weka/

3.1.8 Avaliação do modelo

A avaliação do modelo busca responder se o modelo "funciona ou não funciona". Para saber isto, utilizam-se "medidas de confiança" sobre o modelo que podem responder a perguntas com a precisão, acurácia e representatividade dos dados observados, entre outras. A seleção das "medidas de confiança" depende do modelo escolhido (BERRY; LINOFF, 2004).

No caso de modelos de predição, sobre os resultados apresentados pelo modelo, são avaliadas as métricas sobre a capacidade de predição: matriz de confusão, acurácia, *recall,* precisão, *Fscore* e curva de ROC (*Receiver Operation Characteristic*) com a respectiva AUC (*Area Under Curve*).

Todas as medidas acima citadas partem da matriz de confusão do modelo, que é uma representação na qual são identificadas, dentre o resultado do conjunto de teste, a quantidade de verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN). Os VPs e VNs são classificações efetuadas de forma correta. Ou seja, um caso conhecido (VP, por exemplo) é submetido à predição e o modelo responde conforme esperado. Um FP tem tal classificação quando o resultado é incorretamente previsto como sim (ou positivo) quando na verdade era não (negativo). Um FN é quando o resultado é incorretamente previsto como negativo, quando na verdade é positivo. O Quadro 5 apresenta um modelo de matriz de confusão.

Quadro 5 – Matriz de confusão

		PREDITO	
		Sim	**Não**
REAL	**Sim**	Verdadeiros Positivos (VP)	Falsos Positivos (FP)
	Não	Falsos Negativos (FN)	Verdadeiros Negativos (VN)

Fonte: Adaptado Berry e Linoff (2004, p. 80).

Com a matriz de confusão conhecida, parte-se para as medidas de confiança.

A acurácia de um algoritmo de classificação é uma forma de medir a frequência com que o algoritmo classifica instâncias corretamente, calculando a proporção dos itens

classificados corretamente sobre o total de itens (GOOGLE DEVELOPERS, 2019). A acurácia é calculada pela fórmula:

$$Acurácia = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN}}$$

Precisão é a proporção dos casos VPs em relação aos VPs e FPs. Corresponde à a taxa verdadeiros positivos dividida pelo número total de positivos (WITTEN, FRANK; HALL, 2011, p. 164). A fórmula de cálculo da acurácia é (POWERS, 2011):

$$Precisão = \frac{VP}{VP + FP}$$

Recall é a proporção de casos reais positivos que são corretamente preditos positivos (POWERS, 2011) (WITTEN, FRANK; HALL, 2011, p. 174).

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

Fscore trata da média ponderada entre precisão e *recall*, também conhecida como proporção de concordância específica (POWERS, 2011). É calculada pela fórmula:

$$Fscore = 2 * \frac{Precisão * Recall}{Precisão + recall}$$

A curva ROC (*Receive Operation Characteristic*) representa o desempenho de um classificador sem levar em conta a distribuição de classe ou custos de erro, traçando um gráfico no qual as taxas de verdadeiros positivos são representadas pelo eixo Y e as de falsos positivos no eixo X. Os verdadeiros positivos são representados pelo percentual de total de positivos e o X pelos percentuais de falsos positivos (BERRY; LINOFF, 2004).

Um valor excelente (e máximo) é uma curva com um ângulo reto no canto superior esquerdo do gráfico. Já um gráfico que traça uma linha próxima à diagonal de 45º, começando do ponto (0,0) e finalizando no ponto (1,1) é considerado inútil (BERRY; LINOFF, 2004).

AUC (*Area Under Curve*) nada mais é que a área sobre a curva ROC como uma sumarização do gráfico em um valor, já que quanto maior o valor da área abaixo da curva, melhor o modelo (mais perto do canto superior esquerdo do gráfico). Além disso, é considerada uma boa métrica para verificar a probabilidade de que o classificador classifique uma instância aleatória como positiva sobre outra negativa (WITTEN; FRANK, 2002).

3.1.9 Implementar em produção

É, basicamente, levar o modelo para o ambiente de produção. A implantação em produção possui um grau de dificuldade que dependerá da aderência entre o ambiente inicial de mineração e o de produção. Um dos pontos é ter no modelo variáveis de entrada que não estão disponíveis em produção, como variáveis derivadas de análise (BERRY; LINOFF, 2004). A distância entre dois endereços baseada na latitude e longitude é um exemplo de variável derivada. Tais necessidades devem ser tratadas garantindo a operação correta do modelo em produção.

3.1.10 Avaliação dos resultados

A avaliação dos resultados ocorre acompanhando as saídas em produção e verificando se o objetivo está sendo alcançado (BERRY; LINOFF, 2004). Se o objetivo é o aumento na taxa de vendas em uma loja on-line, tal comportamento deve ser acompanhado. Portanto, avaliar o comportamento do "objetivo" antes e após a implantação é extremamente importante para confirmação dos resultados esperados.

3.1.11 Começar de novo

"Cada projeto de mineração de dados levanta mais perguntas do que respostas. Isto é uma coisa boa" (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p. 82). Este é considerado o "passo 11" que não está representado na Figura 9, mas que se refere ao conhecimento exposto pelo processo trazer novas hipóteses que, consequentemente, levarão ao refinamento do modelo ou criação de novos (BERRY; LINOFF, 2004).

Considerando a extração de conhecimento por meio do KDD, pode-se passar para a representação deste de forma a compartilhá-lo. Uma das ferramentas da EC é a Ontologia.

3.2 ONTOLOGIA

Sendo um termo originário da Filosofia, ontologia vem da junção das palavras gregas *onto* (ser) e *logia* (escrito ou falado), tendo sido aplicada a semântica da natureza dos seres ou ainda uma elucidação sistemática da existência (GRUBER, 1993).

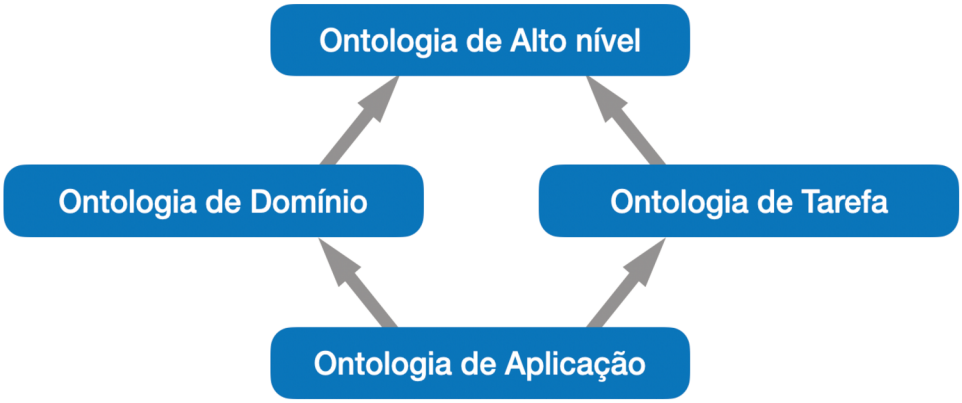
Sob a ótica da inteligência artificial, a definição de ontologia é "uma especificação explícita de uma conceitualização" (GRUBER, 1995, p. 1). Relacionando o conceito da Filosofia sobre a existência com a Inteligência Artificial, para esta o que existe é o que pode ser representado (GRUBER, 1995). Mais tarde, Borst (1999) apresenta uma definição de ontologia como sendo "uma especificação formal e explícita de uma conceitualização compartilhada".

A conceitualização trata de um modelo abstrato, seus fenômenos e os conceitos relevantes a ele. Tais conceitos e fenômenos estão necessariamente explícitos, tendo o formalismo necessário para possibilitar a leitura por máquina. Além disso, a ontologia promove o compartilhamento de tal conhecimento consensual por ela representado (SHUE; CHEN; SHIUE, 2009, p. 2132). Portanto, tal definição considera o fato de que deve existir um senso comum relacionado ao conceito abordado para que tal conhecimento possa ser reutilizado efetivamente.

Por certo, a ontologia apresenta descrição de conceitos, propriedades, atributos, suas relações, restrições entre as relações, além das instâncias dos indivíduos, dentro de um domínio específico (TODESCO, *et al.*, 2009).

Autores classificam as ontologias sob várias perspectivas. Guarino (1998, p. 8) classifica as ontologias por nível de dependência de determinada tarefa, conforme Figura 11.

Figura 11 - Classificação Ontológica de Guarino



Fonte: Guarino (1998, p. 8)

Sendo que (GUARINO, 1998, p. 8):

- Alto nível: são ontologias que tratam de conceitos muito gerais, independentes de um problema ou domínio específico;
- Domínio e tarefa: categorizadas em um mesmo nível de granularidade, tratam de um domínio ou tarefa específica. Automóvel e análise geral de falhas em automóveis são exemplos de ontologias de domínio e tarefa, respectivamente;
- Aplicação: é a especialização das ontologias de domínio e tarefa, muitas vezes envolvendo os papéis relacionados ao domínio e tarefas a serem representadas.

Gruber (1993) define preceitos para nortear o desenvolvimento de ontologias a fim de alcançar os níveis de compartilhamento de conhecimento e interoperabilidade sistêmica compartilhada: clareza, coerência, extensibilidade, compromisso mínimo de codificação e compromisso mínimo ontológico.

- Clareza: refere-se a uma ontologia possuir definições objetivas repassando o significado dos termos compreendidos;
- Coerência: possuir uma consistência dos seus axiomas provê a coerência que uma ontologia necessita;

- Extensibilidade: ter os fundamentos conceituais bem definidos que possibilitem sua extensibilidade além de especialização, na qual novos termos podem ser adicionados dentro das definições previstas na ontologia;
- Compromisso mínimo de codificação: representar o conhecimento com a menor dependência possível de codificação permitirá que vários agentes diferentes façam uso da ontologia, sem se importar com eventuais dependências sistêmicas;
- Compromisso mínimo ontológico: a facilidade e liberdade de especialização de uma ontologia tem relação direta com o seu mínimo compromisso ontológico, sem pré-condições impostas sobre o domínio abordado.

Assim como em várias áreas de conhecimento, o desenvolvimento de ontologias é suportado por metodologias que provêm as diretrizes e processos para a sua construção. Dentre elas, são abordadas brevemente as metodologias On-to-Knowledge, METHONOLOGY, 101 Ontology e ontoKEM que foi concebida com base nas anteriores e disponibiliza uma ferramenta.

On-to-Knowledge Methodology é uma metodologia destinada à manutenção de aplicações de gestão de conhecimento. Tendo o foco principal nos processos e meta-processos de conhecimento (SURE; STUDER, 2002) (SURE; STAAB; STUDER, 2004). Como principais benefícios, os autores destacam ser uma ontologia orientada ao processo, disponibiliza um conjunto de ferramentas para cada etapa deste e possui exemplos de aplicação das etapas de processo baseadas em estudos de caso.

METHONTOLOGY foi concebida com base em experiências em construção de ontologias de domínio em produtos químicos. Ela define seis fases para a sua aplicação (FERNÁNDEZ; GÓMEZ-PÉREZ; JURISTO, 1997):

- Especificação: utiliza linguagem natural para gerar um documento de especificação formal, semiformal ou informal utilizando representações intermediárias ou questões de competência;
- Aquisição do conhecimento: ocorre por todo o processo de desenvolvimento da ontologia, sendo mais intensa sua concomitância na fase de especificação e vai reduzindo sua intensidade à medida que o processo de desenvolvimento da ontologia avança. Faz uso de técnicas com análise formal e *brainstorm*,

aplicadas sobre conhecimentos prévios disponíveis em documentos, com especialistas ou até mesmo em outras ontologias – como exemplos - para elucidar o conhecimento necessário;

- Conceitualização: busca estruturar o conhecimento do domínio por meio de um modelo conceitual. Tal modelo descreve o problema e sua solução utilizando o vocabulário de domínio identificado durante a especificação da ontologia. Os termos que compõem o vocabulário irão formar um glossário, incluindo conceitos, instâncias, verbos e propriedades;
- Integração: relacionado ao reuso de definições disponíveis em outras ontologias ao invés de começar do zero;
- Implementação: trata da codificação da ontologia em linguagem formal utilizando ambiente que suporte tanto a meta-ontologia quanto as ontologias selecionadas para reuso;
- Avaliação: atua na verificação técnica da ontologia, avaliando se a ontologia, seu ambiente de software e documentação estão representando corretamente o sistema proposto;
- Documentação: não é, necessariamente, uma fase isolada. A cada final de fase, há um artefato de documentação. De fato, a definição da documentação como sendo uma fase nessa metodologia procura enfatizar a necessidade da documentação em cada uma das fases.

OntoKEM[7] é identificada pelos autores como uma ferramenta concebida para uso acadêmico nos âmbitos de ensino e pesquisa. Para pesquisa, é aplicada ao cenário onde existe um ou mais especialistas em ontologia que a utilizam para execução de projetos reais, com a necessidade de reportar o andamento dos projetos de forma rápida e eficiente. No ensino, dá suporte aos alunos no desenvolvimento de ontologias. Como este último público possui pouca experiência com ontologias, a ferramenta possibilita um processo de aprendizagem eficiente por meio de *feedbacks* rápidos, além da facilidade de reformular e renomear elementos.

[7] Disponível em http://ontokem.egc.ufsc.br

OntoKEM se baseia nos processos das metodologias 101 e nos artefatos de outras duas metodologias: On-to-Knownledge e METHONTOLOGY. (TODESCO, *et al.*, 2009).

Possibilitando uma abordagem rápida e metodológica sobre a ontologia concebida na corrente dissertação, o OntoKEM permite exportar a ontologia em forma de relatórios e no formato OWL (*Ontology Web Language*), sendo possível importar em outras ferramentas de construção de ontologias, como o Protégé[8], e evoluí-la.

Existem basicamente duas formas de representar uma ontologia: por meio de representação gráfica ou por representação formal. A representação gráfica é mais utilizada para interpretação por seres humanos, enquanto a formal é direcionada ao processamento pelos computadores. Devemos considerar sempre as duas, haja visto que a ausência de uma irá afetar a qualidade da ontologia (ISOTANI; BITTENCOURT, 2015).

Para o presente trabalho, a ontologia em ambas as formas de representação auxilia no entendimento do conhecimento explicitado pelas partes humanas interessadas e a formal permitirá o compartilhamento e uso por outros recursos tecnológicos para processamento como a instanciação dos indivíduos e aplicação de processamento computacional sobre estes.

Entre as linguagens formais mais utilizadas estão a XML, RDF e OWL que serão abordadas nas questões relacionadas a dados abertos.

3.3 DADOS ABERTOS

O conceito de dados abertos parte da teoria e prática sobre Web Semântica. “A Web Semântica dá às pessoas a capacidade de criarem repositórios de dados na Web, construírem vocabulários e escreverem regras para interoperarem com esses dados” (W3C, 2011). Trata-se de uma extensão da web tradicional que possibilita que pessoas e computadores façam uso da mesma informação disponibilizada (BERNERS-LEE; HENDLER; LARISSA, 2002). Com isso, a Web Semântica faz uso de recursos da web tradicional agregando semântica e disponibilizando como um repositório de dados que é acessível tanto para seres humanos quanto para máquinas.

> De forma mais detalhada, a Web Semântica busca utilizar recursos provenientes da Inteligência Artificial (como agentes inteligentes e representação de conhecimento), Engenharia de Software (como frameworks e plataformas), Computação Distribuída

[8] Protégé – Editor de Ontologia de código aberto disponibilizado pela da Universidade de Stanford

> (como *web services*), entre outras, para executar atividades na Web que antes só eram possíveis por agentes humanos (ISOTANI; BITTENCOURT, 2015)

Dentre as tecnologias mais importantes e aliadas ao desenvolvimento da Web Semântica, Berners-Lee, Hendler e Larissa (2002) apontam o XML (*eXtensible Markup Language*) e o RDF *(Resource Description Framework*).

A linguagem XML possibilita representar dados e informações por meio de marcações como <pessoa></pessoa> e <nome></nome>. Ela consiste em um conjunto de regras que permite às pessoas criarem suas próprias marcações, sendo que as regras permitirão a um sistema processar tais marcações definidas pela pessoa (BOSAK; BRAY, 1999). Sendo assim, analisando o trecho de XML apresentado na Figura 12:

Figura 12 - Trecho de código XML

```
<produto>
    <nome>Resma Papel A4 </nome>
    <descricao>Resma de papel tamanho A4 de gramatura 70 na cor branca.</descricao>
    <preco>30,00</preco>
</produto>
```

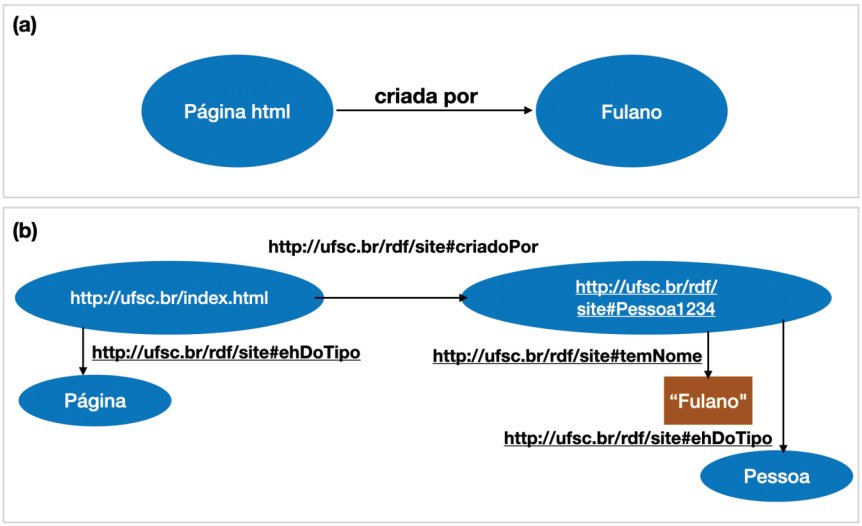
Fonte: do autor.

Uma pessoa ao analisar tal trecho tende a ter certa facilidade para identificar o que significa. Um sistema ao conhecer as regras aplicadas a cada marcação, também poderá facilmente interpretar os dados. Analisando as regras ele saberia, por exemplo, que é um produto composto por nome, descrição e valor, sendo que o valor é monetário.

Já o RDF é composto por triplas que possuem um sujeito, um predicado e um objeto. Logo, uma tripla é uma frase elementar. Fazendo afirmações sobre algo em particular com suas propriedades e valores, tal estrutura é capaz de ser uma forma simples de descrever os dados para serem processados por máquinas. Em termos de representação de uma tripla, tanto o sujeito quanto o objeto são representados por URI (como os links utilizados na web tradicional) ou valores. O predicado normalmente é representado por uma URI que deixa claro qual o conceito por trás do mesmo (BERNERS-LEE; HENDLER; LARISSA, 2002) (LASSILA; SWICK, 1998).

A Figura 13 apresenta uma representação gráfica exemplificando o RDF.

Figura 13 - Exemplo de grafo RDF



Fonte: adaptado de Larissa *et al.* 1998)

Basicamente busca-se representar que uma determinada página de um site foi criada por uma pessoa (a). Para isso utilizando-se de URIs para sujeito, predicado e objetos e relação (b). Transcrevendo tal diagrama para triplas RDF se obtém (Figura 14):

Figura 14 - Exemplos de triplas RDF

```
<http://ufsc.br/index.html>
<http://ufsc.br/rdf/site#ehDoTipo>
<http://ufsc.br/rdf/site#Página> .

<http://ufsc.br/index.html>
<http://ufsc.br/rdf/site#criadaPor>
<http://ufsc.br/rdf/site#Pessoa1234> .

<http://ufsc.br/rdf/site#Pessoa1234>
<http://ufsc.br/rdf/site#ehDoTipo>
<http://ufsc.br/rdf/site#Pessoa> .

<http://ufsc.br/rdf/site#Pessoa1234>
<http://ufsc.br/rdf/site#temNome>
"Fulano" .
```

Fonte: do autor, adaptado de Larissa *et al.* (1998)

Além das linguagens XML e RDF, há também a OWL (*Ontology Web Language*) que foi projetada para "representar conhecimentos ricos e complexos sobre as coisas" (OWL WORKING GROUP, 2012).

A diferença da OWL em comparação com as recomendações XML e RDF da W3C para Web Semântica é que OWL adiciona mais vocabulário para descrever classes e

propriedades, dentre elas as relações entre classes e outros, a cardinalidade, a igualdade, mais possibilidades de tipificação de propriedades, características de propriedades e classes enumeradas (W3C, 2004).

Com o uso de dados abertos para a publicação, possibilita um entendimento comum do domínio modelado na ontologia e, consequentemente, que outros pesquisadores façam uso deles contribuindo para a evolução dos estudos sobre o fenômeno evasão. Da mesma forma, ao utilizar o modelo, cada IES poderá publicar seus dados seguindo o padrão definido e utilizar dados de outras IES realizando:

- Meta-análises em uma população amostral mais abrangente – várias IES de uma região, por exemplo;
- Comparação entre realidades diferentes identificando similaridades.

3.4 CONSIDERAÇÕES

Tendo origem na evolução da TIC, especialmente da inteligência artificial, a EC faz uso de técnicas, métodos e ferramentas para modelar e adquirir conhecimento, formalizando-o e tornando-o reutilizável independentemente de pessoas (HASSLER, *et al.*, 2016).

Neste sentido, o processo de descoberta de conhecimento deve possibilitar, aplicado ao conhecimento de evasão adquirido no capítulo 2, entender os padrões comportamentais dos dados e extrair novos conhecimentos sobre evasão.

Para tal e diante do cenário apresentado, especificamente o processo KDD permite a realização de tais tarefas sobre dados estruturados, comumente utilizados pelas instituições de ensino para a operacionalização de suas atividades administrativas e acadêmicas. Tal processo – resumidamente apresentado na Figura 8 – possibilitará selecionar os dados adequados nas bases de dados, processá-los e chegar à extração de novos conhecimentos.

Dentre tais conhecimentos podemos citar a análise exploratória de dados para identificar padrões e algoritmos de *machine learning* para realizar predição, os quais tornam possível validar a qualidade desses por meio de métricas documentadas.

A ontologia possibilita a representação formal do conhecimento relacionado a um determinado domínio, além do compartilhamento deste para humanos e o processamento computacional.

Essa dissertação faz uso da ontologia para modelar o conhecimento adquirido por meio do arcabouço teórico provido pelas publicações sobre evasão e para nortear o processo de KDD para extração de conhecimento das bases estruturadas de uma instituição de ensino superior.

Acrescenta-se a isto a publicação dos dados no formato aberto. Além de facilitar o seu compartilhamento , possibilita que toda a instituição que fizer uso do modelo compartilhe suas informações – se assim desejar – em um formato padronizado. Isto facilita o intercâmbio de conhecimento (ontologia) e dados (dados abertos) entre instituições e amplia a perspectiva de pesquisas sobre o tema.

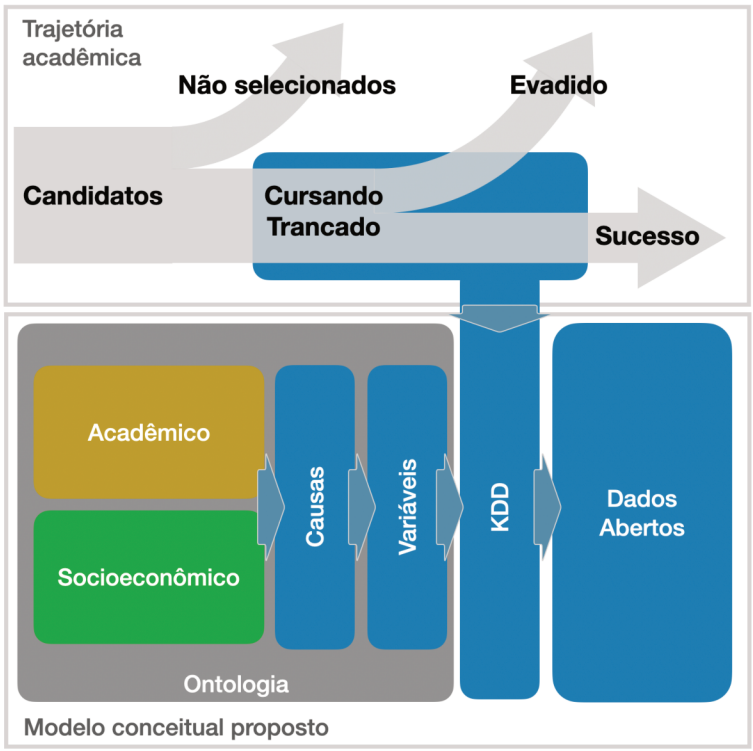
Compreendendo a evasão, bem como as causas e variáveis associadas a elas e a seleção de métodos, técnicas e ferramentas de EC aplicáveis é proposto o modelo de EC.

4 PROPOSTA DO MODELO

Com base nas análises dos estudos e seus resultados apresentados, este trabalho tem por objetivo criar um modelo de evasão como forma de compartilhamento do conhecimento, com suporte a dados abertos que possibilitam ser apoio à análise de evasão em IES.

A revisão integrativa dos artigos apresentou uma ampla gama de temáticas que, para uma análise mais aprimorada, deve identificar trabalhos com certa correlação. Tal constatação corrobora com o relatado por Júnior e Real (2017) nos resultados alcançados em estudos sobre evasão.

Figura 15 - Modelo conceitual proposto



Fonte: do Autor

O modelo proposto, apresentado na Figura 15, está suportado pelas condições típicas ocorridas com o discente ao longo da sua trajetória acadêmica, mais especificamente a partir do momento do seu ingresso na instituição. Desse ponto em diante o modelo sugere a coleta dos dados que comporão as medições até o momento em que o discente evade ou conclui

com sucesso a sua trajetória. Tal modelo conceitual, juntamente com a pergunta de pesquisa deste trabalho, traduzem o problema de negócio em um problema de mineração para a extração de conhecimento.

Este é um modelo aceitável para o proposto, porém mutável. Deve-se considerar que o processo de modelagem é cíclico e cada modelo pode auxiliar na descoberta de novos conhecimentos a partir deste por meio de refinamentos, modificações ou conclusões. Outro ponto importante é o fato de ser dependente de interpretações subjetivas do Engenheiro do Conhecimento. Isso obriga uma avaliação do modelo perante a realidade na busca de um modelo adequado durante todo o seu processo de criação (STUDER; BENJAMINS; FENSEL, 1998). Portanto, há necessidade de refinamentos ao longo da vida do modelo, agregando visões de especialistas de outras áreas ligadas à evasão.

A partir do modelo conceitual proposto e apresentado na Figura 15, são realizados os levantamentos dos grupos, causas e variáveis, compondo o modelo que está suportado pelos estudos identificados e analisados.

4.1 GRUPOS, CAUSAS E MÉTRICAS QUE INFLUENCIAM NA EVASÃO

O modelo proposto está baseado nos estudos visitados neste trabalho durante a revisão da literatura sobre as principais causas de evasão. Entre eles, algumas causas comuns nos estudos estão relatadas e a forma como afetam a evasão. Neste sentido, busca-se relacionar a percepção dos autores diante de uma mesma causa, identificando a sua relevância nos estudos. O Quadro 6 consolida tais percepções comuns, apontando as causas, agrupando-as em questões socioeconômicas e acadêmicas.

Quadro 6 - Causas que levam à evasão com base na revisão da literatura

GRUPO	
CAUSA	**FONTES**
SOCIOECONÔMICA	
Saúde Problemas de saúde consigo ou alguém na família	(KAMAL; AHUJA, 2019)
Distância	(CAMPOS, *et al.*, 2017)

Relacionado a distância entre endereços, como residência – IES ou por ser imigrante	(KAMAL; AHUJA, 2019) (STOESSEL, *et al.*, 2015) (TRUTA; PARV; TOPALA, 2018)
Característica pessoais Temas como idade, gênero	(COSTA; BISPO; PEREIRA, 2018) (SACCARO; FRANÇA; JACINTO, 2019) (STOESSEL, *et al.*, 2015) (VENEGAS-MUGGLI, 2019)
Atualidades Como a pessoa lida com assuntos da atualidade e se mantém atualizada sobre eles	(JUNIOR, *et al.*, 2016)
Família, Trabalho e Renda Chefe de família, ter filhos, estar trabalhando e ser o principal provedor de renda ou auxiliar na renda familiar. Algum fato importante na família que pode desestabilizar	(COSTA; GOUVEIA, 2018) (DIOGO, *et al.*, 2016) (KAMAL; AHUJA, 2019) (SACCARO; FRANÇA; JACINTO, 2019) (STOESSEL, *et al.*, 2015) (VENEGAS-MUGGLI, 2019)
Auxílio financeiro Recebe algum auxílio financeiro da instituição como assistência estudantil e/ou bolsa de estudos	(CARREIRA; LOPES, 2019) (COSTA; GOUVEIA, 2018) (JUNIOR, *et al.*, 2016) (SACCARO; FRANÇA; JACINTO, 2019)
Cotas / Ações afirmativas Seu ingresso se deu por meio de leis de cotas ou ações afirmativas	(CAMPOS, *et al.*, 2017) (JUNIOR, *et al.*, 2016)
ACADÊMICO	
Curso Questões relacionadas ao currículo, duração, sociabilidade entre discentes e com os docentes, forma de ingresso e o nível de satisfação com o curso, são alguns exemplos	(CAMPOS, *et al.*, 2017) (COSTA; BISPO; PEREIRA, 2018) (COSTA; GOUVEIA, 2018) (VENEGAS-MUGGLI, 2019) (SACCARO; FRANÇA; JACINTO, 2019) (TORRES-CORONAS; VIDAL-BLASCO, 2019)

	(TRUTA; PARV; TOPALA, 2018)
Desempenho Suas notas, aproveitamento, trancamentos e reprovações	(COSTA; BISPO; PEREIRA, 2018) (COSTA; GOUVEIA, 2018) (JUNIOR, *et al.*, 2016) (KAMAL; AHUJA, 2019) (TRUTA; PARV; TOPALA, 2018) (VASCONCELOS, *et al.*, 2018) (BODIN; ORANGE, 2018)
Vida acadêmica pregressa Informações relacionadas ao nível de ensino anterior, como o tipo de ensino cursado, tipo de instituição de ensino.	(COSTA; GOUVEIA, 2018) (DIOGO, *et al.*, 2016) (VENEGAS-MUGGLI, 2019)
Interesse acadêmico Capacidade de absorver conteúdo, entusiasmo, dedicação e organização para tal.	(COSTA; GOUVEIA, 2018) (DIOGO, *et al.*, 2016) (SACCARO; FRANÇA; JACINTO, 2019) (TRUTA; PARV; TOPALA, 2018)
Atividades complementares Atividades executadas como estágio, projetos de pesquisa e de extensão	(SACCARO; FRANÇA; JACINTO, 2019) (JUNIOR, *et al.*, 2016)

Fonte: Do autor

Comparando a macro visão apresentada na Figura 6 com as principais causas apontadas nas publicações analisadas (Quadro 6), verifica-se que há um grau elevado de aderência entre os estudos. Comparando as publicações entre 2015 e 2019 com as pesquisas de Cislag (2008), várias das causas continuam presentes e podem levar à evasão nos dias atuais, como as questões familiares, trabalho e renda, questões relacionadas à saúde, interesses, entre outras.

Denota-se do Quadro 6 que ações relacionadas ao incentivo à participação na vida socioacadêmica, além das aulas regulares como uma causa de evasão, pela qual a falta do envolvimento do discente em atividades extra classe e a boa convivência acadêmica influenciam na decisão de evasão. Portanto, tal causa – a participação – pode ser considerada

de permanência quando se analisa a participação e envolvimento. Ou seja, é uma causa de retenção e que é necessário compor o modelo.

Para a definição de variáveis para cada uma das causas, foram analisados os estudos e identificados as variáveis que são apresentadas associadas às respectivas causas. Logo, cada variável listada no Quadro 7 foi identificada no estudo original como compondo uma determinada causa.

Quadro 7 - Lista de variáveis para cada causa

CAUSA	
VARIÁVEL	**IDENTIFICADA NA PUBLICAÇÃO**
SAÚDE	
Ser dependente químico	(KAMAL; AHUJA, 2019)
Ter doença pré-existente	(KAMAL; AHUJA, 2019)
DISTÂNCIA	
Distância campus-residência	(TRUTA; PARV; TOPALA, 2018) (CAMPOS, *et al.*, 2017)
Imigrante	(STOESSEL, *et al.*, 2015)
Reside em área rural	(KAMAL; AHUJA, 2019)
CARACTERISTICAS PESSOAIS	
Idade	(VENEGAS-MUGGLI, 2019) (STOESSEL, *et al.*, 2015) (SACCARO; FRANÇA; JACINTO, 2019)
Gênero	(STOESSEL, *et al.*, 2015) (COSTA; BISPO; PEREIRA, 2018)
Nível de organização pessoal	(SACCARO; FRANÇA; JACINTO, 2019)
ATUALIDADES	
Principal meio de comunicação para se manter atualizado	(JUNIOR, *et al.*, 2016)
FAMÍLIA, TRABALHO E RENDA	
Ajuda no sustento da família	(SACCARO; FRANÇA; JACINTO, 2019) (COSTA; GOUVEIA, 2018)

Algum acontecimento importante e recente na família	(KAMAL; AHUJA, 2019)
Quantidade de filhos	(VENEGAS-MUGGLI, 2019) (COSTA; GOUVEIA, 2018)
Renda familiar	(SACCARO; FRANÇA; JACINTO, 2019) (COSTA; GOUVEIA, 2018)
Trabalha	(STOESSEL, *et al.*, 2015) (SACCARO; FRANÇA; JACINTO, 2019) (COSTA; GOUVEIA, 2018) (DIOGO, *et al.*, 2016)
AUXÍLIO FINANCEIRO	
Assistência estudantil	(SACCARO; FRANÇA; JACINTO, 2019) (COSTA; GOUVEIA, 2018) (JUNIOR, *et al.*, 2016)
Bolsa de estudo	(CARREIRA; LOPES, 2019) (SACCARO; FRANÇA; JACINTO, 2019)
COTAS / AÇÕES AFIRMATIVAS	
Ingresso por cota	(CAMPOS, *et al.*, 2017) (JUNIOR, *et al.*, 2016)
CURSO	
Duração superior a 2 anos	(VENEGAS-MUGGLI, 2019) (COSTA; BISPO; PEREIRA, 2018)
Modalidade do curso	(TORRES-CORONAS; VIDAL-BLASCO, 2019)
Nível de satisfação com o curso	(TRUTA; PARV; TOPALA, 2018) (COSTA; BISPO; PEREIRA, 2018) (CAMPOS, *et al.*, 2017)
Nível de satisfação com a convivência acadêmico-social	(TORRES-CORONAS; VIDAL-BLASCO, 2019) (COSTA; GOUVEIA, 2018)
Ingresso pelo ENEM	(SACCARO; FRANÇA; JACINTO, 2019) (CAMPOS, *et al.*, 2017)
DESEMPENHO	

Reprovações no primeiro semestre/ano	(KAMAL; AHUJA, 2019) (COSTA; BISPO; PEREIRA, 2018) (JUNIOR, *et al.*, 2016) (BODIN; ORANGE, 2018)
Reprovações totais no curso	(KAMAL; AHUJA, 2019) (TRUTA; PARV; TOPALA, 2018) (JUNIOR, *et al.*, 2016)
Trancamentos durante o curso	(TRUTA; PARV; TOPALA, 2018) (COSTA; BISPO; PEREIRA, 2018) (VASCONCELOS, *et al.*, 2018)
Índice de Aproveitamento Acadêmico	(KAMAL; AHUJA, 2019) (TRUTA; PARV; TOPALA, 2018) (COSTA; BISPO; PEREIRA, 2018)
VIDA ACADÊMICA PREGRESSA	
Ensino anterior na modalidade Educação de Jovens e Adultos	(VENEGAS-MUGGLI, 2019) (COSTA; GOUVEIA, 2018)
IDEB como medida de qualidade acadêmica no nível anterior de ensino	(COSTA; GOUVEIA, 2018)
Teve orientação vocacional	(DIOGO, *et al.*, 2016)
INTERESSE ACADÊMICO	
Capacidade de absorção do conteúdo	(TRUTA; PARV; TOPALA, 2018)
Dedicação ao estudo	(TRUTA; PARV; TOPALA, 2018) (DIOGO, *et al.*, 2016)
Entusiasmo com o estudo e/ou curso	(TRUTA; PARV; TOPALA, 2018) (DIOGO, *et al.*, 2016)
Tempo dedicado aos estudos	(TRUTA; PARV; TOPALA, 2018) (SACCARO; FRANÇA; JACINTO, 2019)
Estar comprometido com o curso	(COSTA; GOUVEIA, 2018)
ESTÁGIO, PESQUISA E EXTENSÃO	

Realiza ou realizou estágio durante o curso	(SACCARO; FRANÇA; JACINTO, 2019) (JUNIOR, *et al.*, 2016)
Participa ou participou de projeto de pesquisa durante o curso	(SACCARO; FRANÇA; JACINTO, 2019) (JUNIOR, *et al.*, 2016)
Participa ou participou de projeto de extensão durante o curso	(SACCARO; FRANÇA; JACINTO, 2019) (JUNIOR, *et al.*, 2016)

Fonte: do autor.

No total, 37 variáveis foram identificadas como pontos de partida para a avaliação da evasão nas instituições de ensino. Ressalta-se que foram elencadas apenas as variáveis que fazem parte das causas conclusivas de evasão nos estudos avaliados. Algumas variáveis foram identificadas de forma indireta. Como por exemplo o IDEB – Índice de Desenvolvimento da Educação Básica – que não é diretamente citada no referido estudo, mas os autores indicam a necessidade de medir a qualidade do estudo no nível médio (vida acadêmica pregressa). No entanto o IDEB "reúne, em um só indicador, os resultados de dois conceitos igualmente importantes para a qualidade da educação: o fluxo escolar e as médias de desempenho nas avaliações" (INEP, 2020).

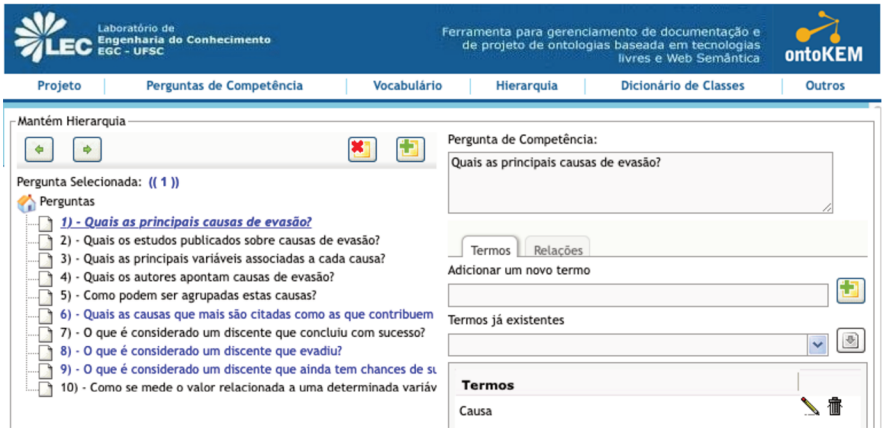
Tais variáveis comporão a seleção de dados apropriados para a continuidade do modelo, sendo o próximo passo a definição formal do modelo por meio de uma ontologia.

4.2 DEFINIÇÃO DA ONTOLOGIA

De posse do modelo geral, da identificação dos grupos, causas e métricas, é definida a ontologia de domínio para o contexto para o SBC. Para a construção da ontologia, utiliza-se o OntoKEM. Consequentemente, todo o arcabouço relacionado com as metodologias sobre as quais ele foi concebido estão presentes nesta fase inicial. Na ontologia representa-se os grupos e suas causas identificadas de evasão, as variáveis associadas a cada causa e demais propriedades mapeadas com base no teórico avaliado.

Deste ponto, foram inicialmente definidas quais as Perguntas de Competência que se deseja responder por meio da ontologia sobre a evasão em instituição de ensino superior. No total foram elencadas dez perguntas.

Figura 16 - Tela com as perguntas de competência

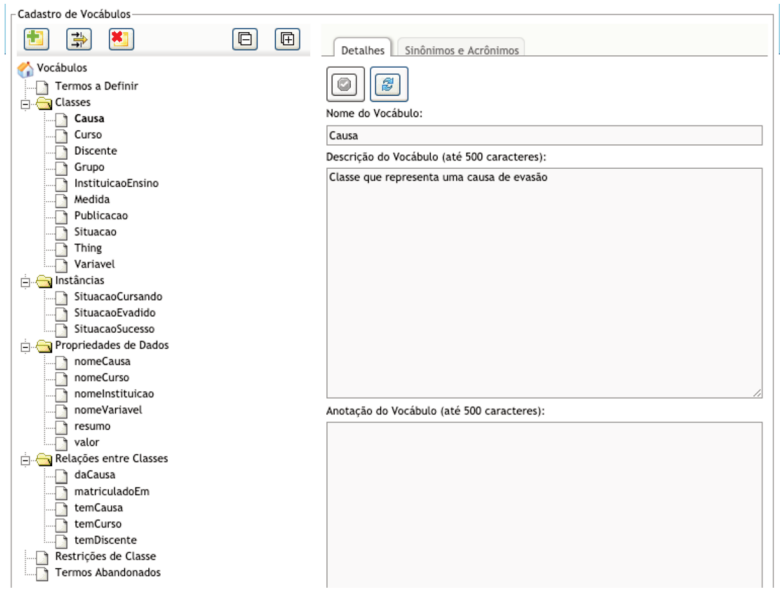


Fonte: Do autor.

Durante a criação das Perguntas de Competência, são identificados os Termos e Relações iniciais a partir das perguntas. Na Figura 16, selecionando em "Perguntas" a pergunta de número 4, o sistema apresenta – na direita – a pergunta e os termos e relações que foram cadastrados para a pergunta.

À medida que vamos avançando no cadastro no OntoKEM, mais relações e classes irão surgindo. No Cadastro de Vocabulário – Figura 17 – são definidas as classes, instâncias, propriedades de dados, propriedades de objetos, restrições de classes e termos de abandono. Tais itens são oriundos da análise realizada das perguntas de competência. Além de definir qual desses itens está sendo acrescido à ontologia, a descrição do vocábulo é parte importante para, posteriormente, facilitar o entendimento da ontologia e identificar possíveis mudanças quando a descrição tende a ser extensa ou pouco coesa.

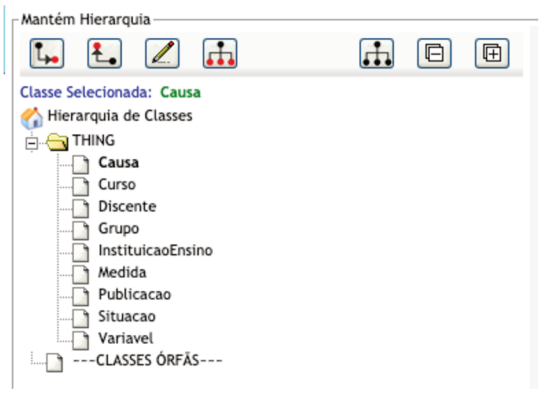
Figura 17 - Cadastro de vocábulos



Fonte: do autor.

Do cadastro de vocábulos é realizado a organização da hierarquia de classes da ontologia, como mostra a Figura 18.

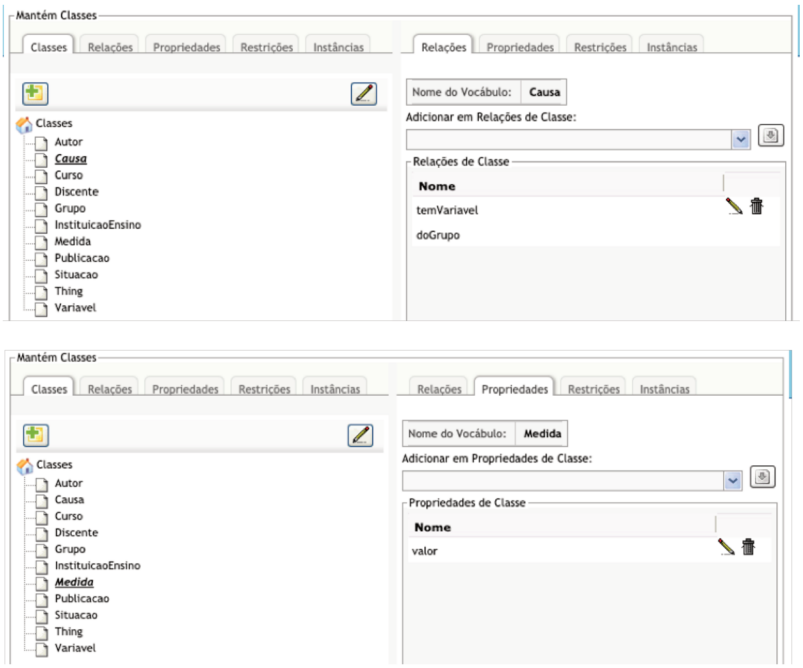
Figura 18 - Hierarquia de classes



Fonte: do autor.

O próximo passo é o refinamento do Dicionário de Classes, no qual são definidas a inter-relação entre as classes, propriedades, restrições e instâncias. Exemplificando (vide Figura 19), para classe Discente, foram definidas as relações que pertencem a tal classe. Para a relação “matriculado em”, foi definido a que domínio de classe tal relação pertence e qual o range.

Figura 19 - Manter o Dicionário de Classes



Fonte: do autor.

Ao final do processo inicial, exporta-se a ontologia-base no formato OWL (*Ontology Web Language*) para refinamento e aprimoramento no Protégé. Nessa, foram revistos alguns relacionamentos, propriedades, bem como adicionadas as classes específicas para cada uma das causas, variáveis e grupos de causas. Após isso, foram instanciados na própria ontologia os indivíduos que fazem parte do modelo, como as causas, grupos e variáveis, além das publicações e seus autores. Como resultado, temos as hierarquias de classes, de propriedades de objetos e de propriedades de dados, apresentada na Figura 20.

Figura 20 - Classes, propriedades de objetos e propriedades de dados

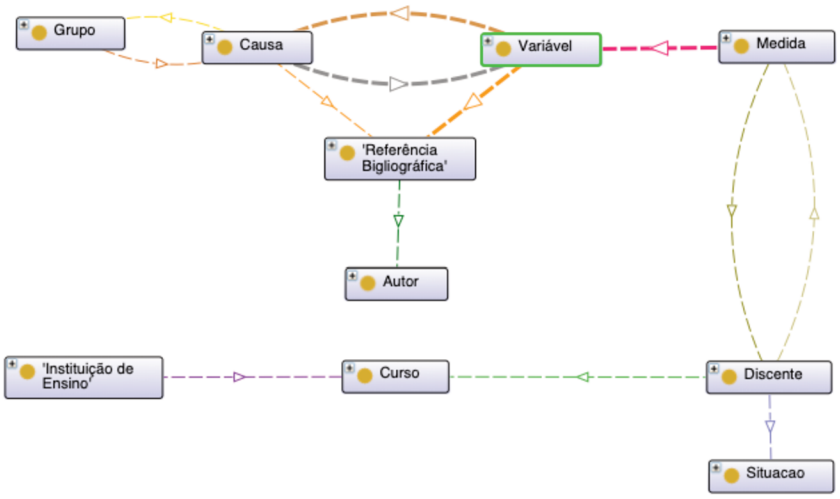
Fonte: do autor.

Com tal representação, o modelo contempla o arcabouço referencial para cada causa e a variável identificada, o que possibilitará a qualquer instituição de ensino fazer uso deste como ponto de partida para análise da evasão, visto que ele já possui o direcionamento para tal. De fato, nem todas as instituições terão os dados para medir todas as variáveis definidas no

modelo, o que não impede o uso do modelo de forma parcial – medir apenas as variáveis possíveis – sob as informações disponíveis em suas bases de dados estruturadas.

A representação gráfica resumida da ontologia – Figura 21 – contendo as classes, na qual as linhas tracejadas são os relacionamentos entre classes e hierarquias de classes. A representação gráfica total, com as causas e variáveis disponíveis, está no ANEXO B – Ontologia completa.

Figura 21 - Representação gráfica da ontologia



Fonte: do autor.

É importante ressaltar que "novas observações podem levar a um refinamento, modificação ou conclusão do modelo já construído. Por outro lado, o modelo pode orientar a futura aquisição de conhecimentos." (STUDER; BENJAMINS; FENSEL, 1998, p. 3). Sendo assim, a ontologia poderá ser atualizada e evoluída com revisão de novos estudos, adicionando tais referenciais, novas causas e variáveis. Tal característica dá ao modelo flexibilidade para evolução.

O modelo ontológico proposto propicia então descobrir o conhecimento consolidado sobre as principais causas de evasão e direciona o processo de análise dos dados de uma instituição de ensino que, ao final, tende a identificar novos conhecimentos a partir dos dados e retroalimentar o modelo com tais informações importantes.

A partir do modelo como norteador, passa-se para a extração de conhecimento por meio do KDD.

4.3 KDD

Além da Ontologia do modelo – que representa o conhecimento-base do domínio em questão com suas classes, atributos, propriedades, relacionamentos e indivíduos relacionados ao conhecimento extraído das publicações – o modelo prevê o uso de KDD para popular os indivíduos e suas relações das classes “Instituição de Ensino”, “Curso”, “Discente” e “Medida”.

Boa parte das IES, se faz uso de sistemas estruturantes diferentes para operacionalizar os processos de ingresso, acadêmico e assistencial. Por este motivo o KDD é sugerido para extração dos dados dos sistemas estruturantes e popular as instâncias das classes.

Como apresentado no capítulo 2.3 Descoberta do Conhecimento, o objetivo do KDD é a extração de conhecimento. Neste sentido, o modelo direciona para tal, sendo que os novos conhecimentos extraídos para o cenário específico – realidade de uma IES – possam ser, agregando a ontologia – extensão e especialização do modelo para cenário particular.

Diante do cenário no qual a extração de conhecimento é particular para o cenário de cada IES, o modelo faz apenas o direcionamento para tal processo.

4.4 PUBLICAÇÃO DOS DADOS (FORMATO ABERTO)

Tendo o modelo representado, com seus indivíduos instanciados – gerais e específicos para o cenário da IES –, a publicação de dados no formato aberto possibilita um entendimento comum sobre o que está representado e, para análises por meio da tecnologia, o acesso e processamento por sistemas computacionais.

Além dos benefícios para a instituição, o horizonte é ampliado quando os dados abertos são publicados para outros de interesse na área. Dependendo do nível de publicidade a ser atribuído aos dados, a IES pode optar por utilizar dados que não identifiquem a pessoa. Ou seja, ao publicar dados no formato aberto e de acesso público deve-se levar em consideração publicar dados não pessoais (OPEN DATA HANDBOOK, 2020).

Permitir que pesquisadores independentes ou associados da instituição que têm a custódia dos dados tenham acesso a eles possibilitará que possam formular e executar políticas públicas, avaliar como os recursos da sociedade são aplicados e verificar os impactos na melhoria da qualidade de vida. Pesquisas bem fundamentadas promovem um aumento no nível

de autoconhecimento de determinada sociedade, do conhecimento de outras sociedades, além de possibilitar que os interessados tomem decisões justificadas com base na ciência (PIRES, 2015).

Neste sentido, a publicação em dados abertos fomenta não só pesquisas internas na instituição específica, mas da comunidade de pesquisa com interesse no tema.

4.5 CONSIDERAÇÕES

A partir do conhecimento adquirido sobre evasão e dos métodos, técnicas e ferramentas da EC, o modelo de EC proposto leva em consideração a vida acadêmica do discente (do ingresso à conclusão ou evasão).

Ainda, as causas e variáveis que foram identificadas como fatores de evasão nos estudos que compõem o arcabouço de publicações. Esse arcabouço também faz parte do modelo que norteia as análises, sendo representado pelas referências bibliográficas instanciadas na ontologia. Assim, tendo conhecimento das variáveis, uma instituição pode fazer uso de KDD para extrair os dados dessas variáveis das bases estruturadas e armazenar esses dados em formato aberto e inferir novos conhecimentos.

Isto é possível aplicando os métodos contidos no KDD para identificação de padrões comportamentais dos dados e possibilidades de predição de tendências à evasão.

Para observar como o modelo se comporta em um caso prático é realizada a prova de conceito aplicando o modelo sobre os dados do Instituto Federal de Santa Catarina - IFSC.

5 PROVA DE CONCEITO

Com o modelo proposto disponível, é realizada a prova de conceito sobre os dados dos cursos de graduação do IFSC. “Prova de Conceito, do inglês *Proof of Concept* (PoC), é um termo utilizado para denominar um modelo prático que possa provar o conceito (teórico) estabelecido por uma pesquisa ou artigo técnico” (SILVA, 2014)

Durante este capítulo serão descritos os procedimentos, métodos e técnicas utilizadas para a identificação das fontes de dados, o mapeamento, a extração e a carga deles em formato de dados abertos, conforme modelo.

A pesquisa está registrada na instituição e tem autorização expressa para este fim, desde que se cumpram as normas relacionadas ao sigilo dos indivíduos. Por este motivo, os dados sensíveis que possam identificar um ou mais indivíduos serão mascarados.

Tendo a pergunta de pesquisa como problema de negócio a ser resolvido a as variáveis como direcionadores para a fase de seleção de dados no KDD, são realizados os procedimentos para se chegar na extração do conhecimento, iniciando-se pela extração dos dados das bases de dados dos sistemas do IFSC.

5.1 EXTRAÇÃO DOS DADOS

Norteados pelas variáveis sugeridas pelo modelo (Quadro 7 - Lista de variáveis para cada causa), foram avaliados cada um dos sistemas cujos dados e informações estão disponíveis, sendo eles:

- SIGAA[9] - Sistema Integrado de Gestão de Atividades Acadêmicas. Implantado em 2017 no IFSC – substituindo o sistema ISAAC –, mantém toda a história do discente dentro do IFSC, os cursos, docentes, estruturas curriculares e todo o necessário que um sistema acadêmico exige. Nele é possível extrair informações sobre docentes, disciplinas, o desempenho do aluno, dados do campus, curso e informações sobre endereço do aluno e do campus. Ressalta-se que os dados disponíveis no sistema ISAAC e que podiam ser migrados,

[9] Faz parte deu um pacote de sistemas integrados desenvolvidos pela UFRN. https://info.ufrn.br

foram levados para o SIGAA, tornando-o o mais completo em informações acadêmicas. No entanto, pode haver inconsistências;

- Sistema de Ingresso: operacionaliza todo o processo de ingresso do discente no IFSC. Pelo fato de fazer parte do processo de ingresso as questões relacionadas aos dados socioeconômicos e cota, ele possibilita extração de informações de nível socioeconômico, inclusive;
- PAEVS: sistema que operacionaliza o processo do Programa de Assistência Estudantil para discentes em Vulnerabilidade Social. É por meio desse sistema que a área responsável, composta por uma equipe multidisciplinar de psicólogos, pedagogos, nutricionistas, entre outros, faz a análise e identifica discentes em vulnerabilidade social, definindo um Índice de Vulnerabilidade Social (IVS) e o auxílio financeiro às assistências mais adequadas em cada caso.

Analisando então o "Quadro 7 - Lista de variáveis para cada causa" e os sistemas da instituição, o Quadro 8 identifica quais variáveis estão disponíveis e em qual sistema.

Quadro 8 - Sistemas *versus* variáveis

SISTEMA	VARIÁVEIS
SIGAA	Distância campus-residência
	Idade
	Gênero
	Duração superior a 2 anos
	Modalidade do curso
	Ingresso pelo ENEM
	Ingresso por cota
	Reprovações no primeiro semestre/ano
	Reprovações totais no curso
	Trancamentos durante o curso
	Índice de Aproveitamento Acadêmico
	Realiza ou realizou estágio durante o curso

	Participa ou participou de projeto de pesquisa durante o curso
	Participa ou participou de projeto de extensão durante o curso
Sistema de Ingresso	Imigrante
	Reside em área rural
	Ajuda no sustento da família
	Quantidade de filhos
	Renda familiar
	Trabalha
PAEVS	Assistência estudantil
	Índice de Vulnerabilidade Social
Não existe informação	Ser dependente químico
	Ter doença pré-existente
	Principal meio de comunicação para se manter atualizado
	Algum acontecimento importante e recente na família
	Bolsa de estudo
	Nível de satisfação com o curso
	Ensino anterior na modalidade Educação de Jovens e Adultos
	IDEB como medida de qualidade acadêmica no nível anterior de ensino
	Teve orientação vocacional
	Capacidade de absorção do conteúdo
	Dedicação ao estudo
	Entusiasmo com o estudo e/ou curso
	Tempo dedicado aos estudos
	Nível de organização pessoal
	Estar comprometido com o curso

Fonte: do autor.

Como proposto pelo modelo, ele leva em consideração um conjunto de estudos que foram realizados em instituições diferentes. Naturalmente, e mesmo entre os estudos, dificilmente uma instituição hoje possui as informações para compor todas as variáveis. Considerando os estudos avaliados, o grupo de variáveis possui diferenças. Logo, o modelo será aplicado utilizando parcialmente as variáveis, o , que viabiliza o uso dos dados existentes.

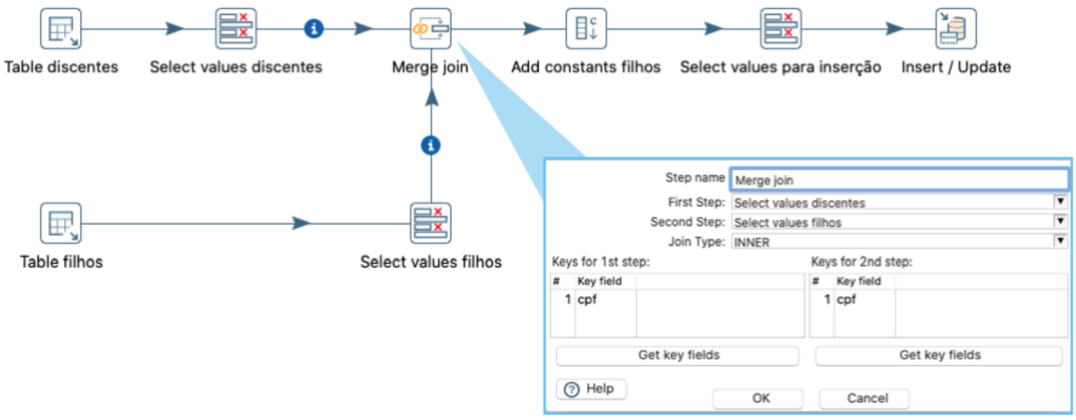
A maior parte das variáveis que estão na linha “Não existe informação” do Quadro 8 são as que necessitam de um estudo adicional por meio de análise detalhada do tema. Outras, como o IDEB da instituição de ensino médio, uma vez que a instituição analisada não possui a informação sobre qual instituição de ensino médio o discente concluiu esta etapa acadêmica. É uma melhoria no processo de ingresso da instituição que viabilizará aprimorar análises futuras. Com a identificação dos sistemas e as respectivas variáveis é realizada a extração dos dados e a correção de problemas nesses para cada variável.

A fase de extração dos dados por meio do KDD é uma das mais delicadas, pois necessita de conhecimento no que se refere às questões educacionais e tecnológicas, para garantir, principalmente, que a extração e transformação estejam aderentes ao domínio da ontologia, ao modelo de ensino e ao propósito o que a transforma em uma atividade multidisciplinar pelo fato de abranger técnicas que estão além de uma disciplina de aprendizagem de máquina e por convergir vários paradigmas da computação (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) (ROMERO; VENTURA; GARCÍA, 2008).

Nesse sentido, com base em orientações da área de ensino da instituição, foram realizadas as fases do ETL utilizando a ferramenta Pentaho Data-integration, que usa dois conceitos fundamentais: *jobs* e *transformations.*

Uma *transformation* pode ser considerada como uma pequena parte do KDD de um todo. Por exemplo, a extração, transformação e carga das reprovações dos discentes é uma parte da extração, considerando todo o resto para as demais variáveis. A ferramenta em questão é composta por uma representação visual dos componentes, suas ligações com o fluxo dos dados e, para cada componente, são definidos os comportamentos desejados. Há componentes cujo comportamento é definido por simples parâmetros, outros por códigos SQL ou até mesmo de programação como JavaScript. A Figura 22 mostra uma transformação, na qual cada caixa é um componente que provê funcionalidades específicas, as setas são as conexões que indicam o fluxo do dado. Em detalhe pode-se ver como são definidos os parâmetros do componente *merge join*.

Figura 22 - Exemplo de *transformation* com detalhe do *merge join*



Fonte: do autor

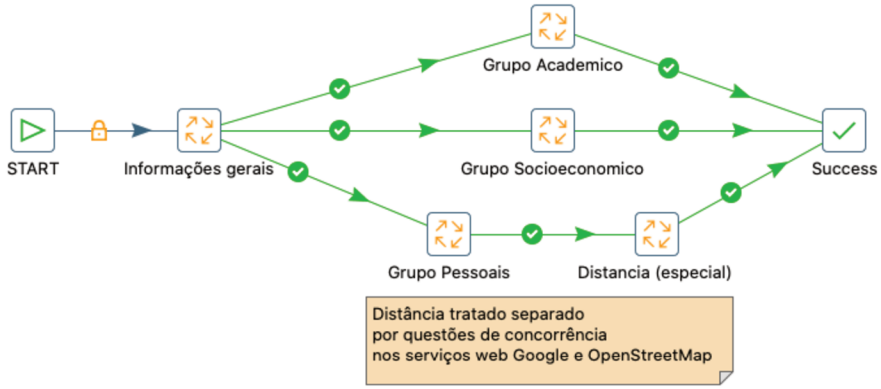
Na transformação acima, são compatibilizados dados de duas origens diferentes – representados pelos componentes *table* que são dois bancos de tecnologias diferentes: PostgreSQL e MySQL. Com os dados de origem disponíveis, passa-se, então, para a fase de transformação, na qual dados similares em bases diferentes são tornados equivalentes e padronizados. Exemplificando: se em um sistema a informação CPF contém somente números, no outro números, pontos e hífen, é possível transformar uma das origens para compatibilizar com a outra. Assim, passam a ser compatíveis ao longo de todo o processo de transformação subsequente, visto que, conceitualmente, CPF representa a mesma informação nas duas origens, apenas estão em formatos diferentes.

Por fim, entra-se na fase de carga – representada pelo componente *insert* na imagem anterior –, na qual todos os dados coletados e transformados são disponibilizados em uma local, que pode ser também um banco de dados, um arquivo ou outro destino suportado.

Após criar todas as transformações, pode-se criar um ou mais *jobs*. *Job* é um conjunto de *transformations* e sub *jobs* interconectados para realizar toda a extração, transformação e carga do conhecimento pretendido. Além desses componentes, são definidas também as características e os parâmetros a serem compartilhados entre passos do *job*. Para a dissertação, como exemplo, códigos dos cursos são compartilhados entre todos os passos do *job* que é a referência para todo o processo ETL criado. A Figura 23 é a representação gráfica do *job* resultante da transformação efetuada para o ETL dos dados dos sistemas do IFSC para serem aplicados na validação da hipótese pelo protótipo. É importante ressaltar que um *job* faz

chamadas de outros *jobs* e *transformation* e a visão apresentada na figura é limitada ao *job* principal, abstraindo as 30 *transformations* que foram necessárias para a extração.

Figura 23 - *Job* principal



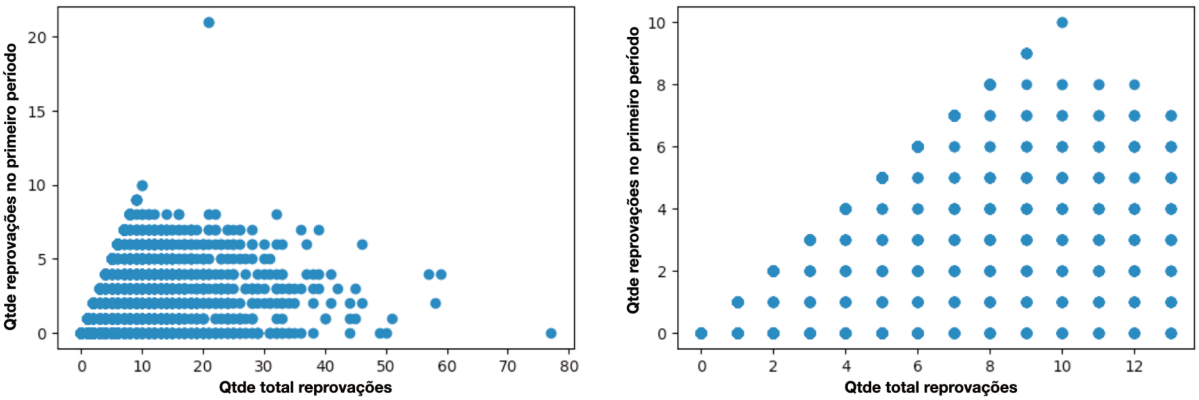
Fonte: do autor.

Em especial, na transformação que trata da distância entre residência-campus incorpora-se outras fontes de dados: APIs do Google Maps e do OpenStreetMap. Tal inclusão foi necessária para buscar os dados de latitude e longitude dos endereços de residência e do campus e, a partir desses calcular a distância em quilômetros. Por questões de arquitetura, na qual todos os endereços necessários são tratados sequencialmente, e de redução da concorrência nas chamadas de serviços externos, a distância está sendo processada separadamente, após as demais variáveis do *job* "Grupo Pessoais".

5.2 HIGIENIZAÇÃO DO DADOS

Nos dados coletados, aplica-se a análise para limpeza dos dados, identificando dados discrepantes e incorretos. Foram identificados dados inconsistentes nas seguintes variáveis conforme as figuras a seguir.

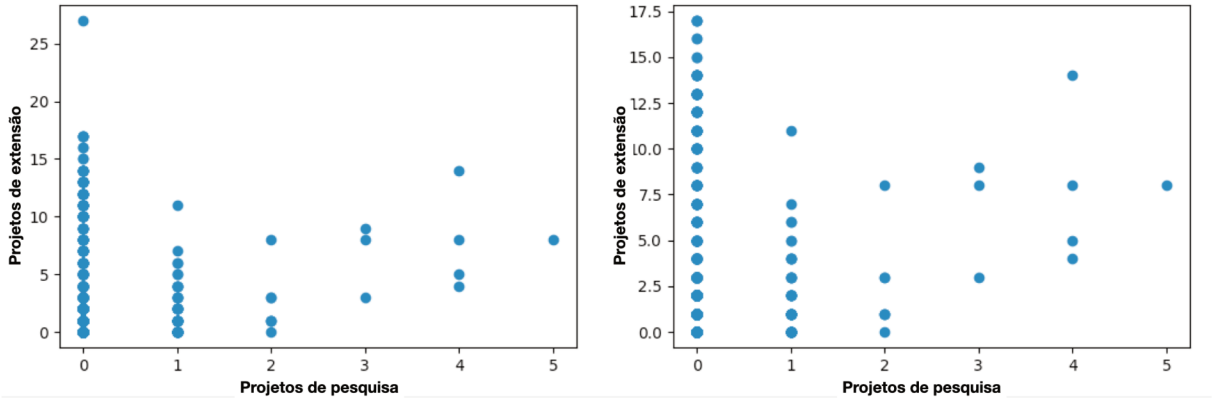
Figura 24 - Reprovações total e no primeiro período



Fonte: do autor.

A Figura 24 mostra que existiam alguns dados discrepantes nas duas variáveis (gráfico da esquerda). Constatou-se que para a variável Reprovações no primeiro período existe um dado próximo de 25 reprovações o que, na verdade, era um erro nos dados de origem. Além desse ponto, constatou-se que apenas 13 (de 2599) registros possuíam valores acima de 10 reprovações. Optou-se por eliminar estes. Quanto à reprovação total, o mesmo fato ocorre, realizando-se , assim, a remoção dos valores acima de 17 reprovações (253 registros de 2599) visto que o quartil q3 está no valor 7. O resultado é apresentado no gráfico esquerdo da mesma figura.

Figura 25 - Variáveis pesquisa e extensão

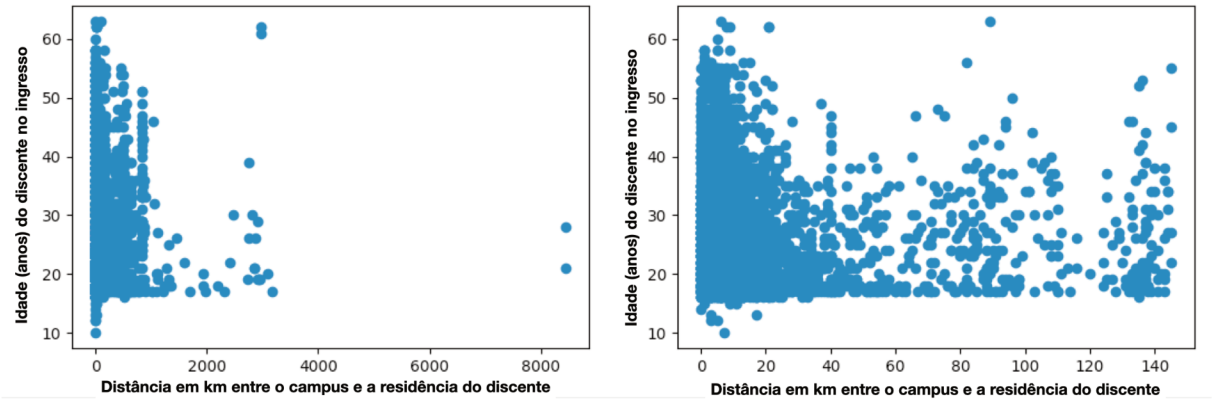


Fonte: do autor.

A variável extensão possui apenas um valor discrepante (acima de 25 projetos), como se observa na Figura 25. Ou seja, um mesmo discente participa de mais de 25 projetos de

extensão. Na análise mais aprofundada dos dados concluiu-se que era um discente que teve vários cadastros de teste no sistema de produção para "testar a emissão de certificado". Tal dado foi removido e o resultado é apresentado no gráfico da esquerda, da Figura 25.

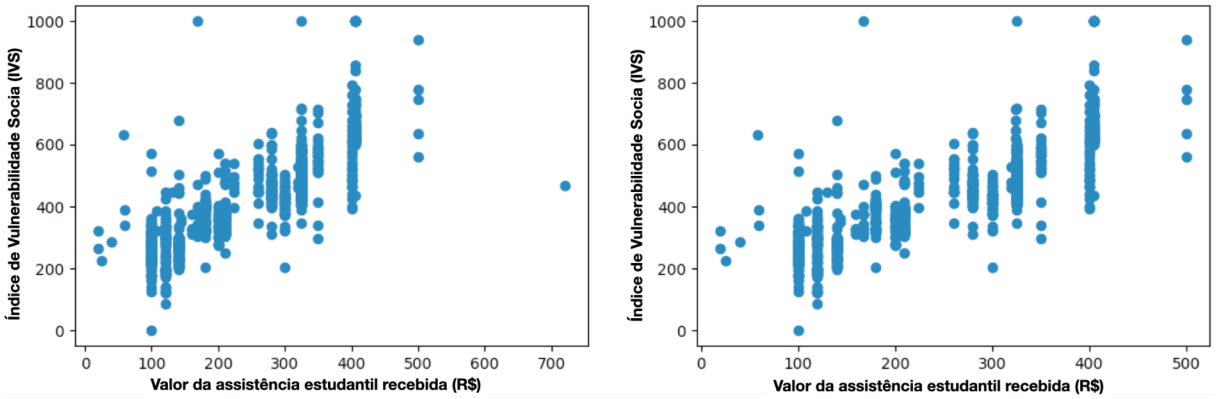
Figura 26 - Variáveis Idade no ingresso e distância do campus



Fonte: do autor.

Quanto à distância do campus, foram identificadas distâncias (em quilômetros) muito altas para cursos presenciais e semipresenciais. Em análise verificou-se erro nos dados de origem[10], que estavam incorretos. Optou-se por remover os dados de 345 (de 6814) registros cuja distância era superior a 145 km. Com isso, obteve-se uma distribuição mais homogênea e aderente ao domínio pretendido (Figura 26).

Figura 27 - Variáveis IVS e recebe assistência estudantil



Fonte: do autor.

[10] Tal dado foi adquirido por meio de APIs externas ao SIGAA, coletando latitude e longitude de cada endereço e, posteriormente, cálculo da distância

A variável “Recebe assistência estudantil” apresentou apenas 1 valor discrepante. A análise apontou para erro no dado de origem e ele foi descartado (vide Figura 27).

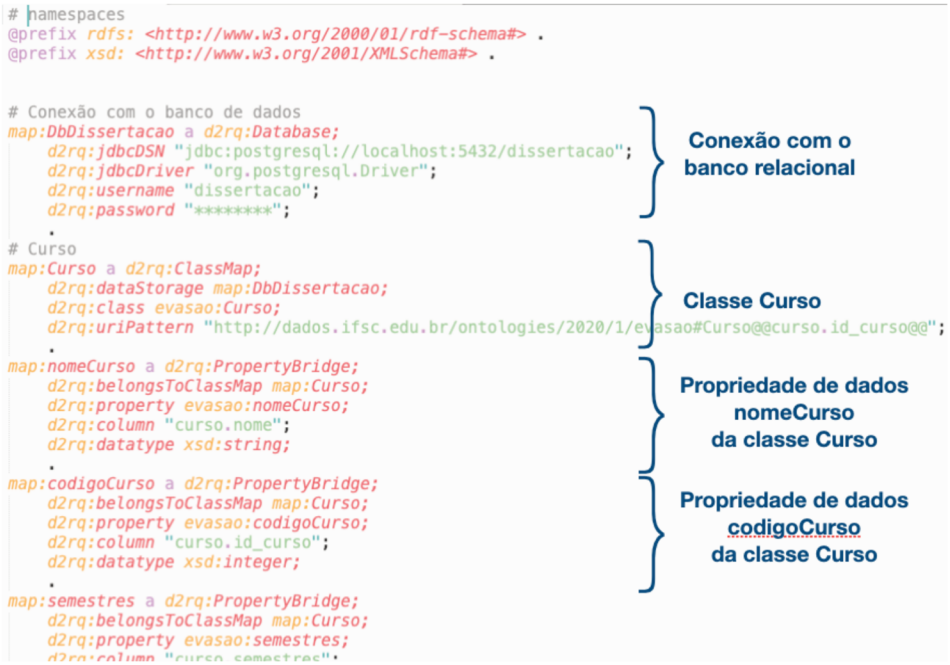
Com os dados higienizados, é feita a sua publicação em formato de dados abertos.

5.3 PUBLICAÇÃO DOS DADOS DE EVASÃO

Como complemento para a ferramenta Pentaho Data-integration, utiliza-se uma segunda ferramenta chamada D2RQ para a efetiva publicação. Esta ferramenta facilita a conversão dos dados em modelo relacional para dados abertos no formato RDF/XML. Isso é necessário pois o protótipo faz uso da Ontologia e das instâncias para a representação do conhecimento e disponibilização dele para análise da hipótese. Assim, entende-se ser possível verificar quais fatores de evasão estão presentes no padrão do IFSC.

A ferramenta D2RQ se baseia em um arquivo de mapeamento utilizando a linguagem *D2RQ Mapping Language* que segue a sintaxe *Turtle*. Neste arquivo são definidos os *namespaces* utilizados, a conexão com o banco de dados relacional de origem, as classes, propriedades de dados e propriedades de objetos. No caso da dissertação, seguindo o que está representado na Ontologia, foi implementado o mapeamento aderente, como mostra parcialmente o conteúdo do arquivo na Figura 28.

Figura 28 - D2RQ - Arquivo de mapeamento parcial




```
# namespaces
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

# Conexão com o banco de dados
map:DbDissertacao a d2rq:Database;
    d2rq:jdbcDSN "jdbc:postgresql://localhost:5432/dissertacao";
    d2rq:jdbcDriver "org.postgresql.Driver";
    d2rq:username "dissertacao";
    d2rq:password "*******";
    .
# Curso
map:Curso a d2rq:ClassMap;
    d2rq:dataStorage map:DbDissertacao;
    d2rq:class evasao:Curso;
    d2rq:uriPattern "http://dados.ifsc.edu.br/ontologies/2020/1/evasao#Curso@@curso.id_curso@@";
    .
map:nomeCurso a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:Curso;
    d2rq:property evasao:nomeCurso;
    d2rq:column "curso.nome";
    d2rq:datatype xsd:string;
    .
map:codigoCurso a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:Curso;
    d2rq:property evasao:codigoCurso;
    d2rq:column "curso.id_curso";
    d2rq:datatype xsd:integer;
    .
map:semestres a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:Curso;
    d2rq:property evasao:semestres;
```

Fonte: do autor.

Uma das vantagens em utilizar o D2RQ ao longo do desenvolvimento da pesquisa é a possibilidade de uso do recurso d2rq-server que faz a conversão relacional para RDF/XML de forma on-line. Isso permite que alterações feitas na Ontologia ao longo do seu processo de maturação e que são refletidas no processo – consequentemente no banco relacional intermediário – sejam facilmente representadas em RDF/XML. Para isso, basta realizar as adequações no arquivo de mapeamento conforme definido na Ontologia.

Chegando ao ponto de estar suficiente, é realizado o passo final do ETL utilizando um terceiro recurso do D2RQ: d2rq-dump. Esta ferramenta faz a geração de um arquivo RDF/XML dos dados relacionais utilizando como diretriz o arquivo de mapeamento. Assim, é possível carregar tais instâncias da ontologia no Virtuoso.

5.4 EXTRAÇÃO DO CONHECIMENTO

Berry e Linoff (2004, p. 9) definem que aplicar o processo de classificação requer um exame das características de determinado objeto de estudo, atribuindo a este uma ou mais classes. Exemplificando, para classificar a maioridade de uma pessoa, podemos analisar o

atributo idade, verificando que se for maior de 18 anos receberá a classe “maior de idade”. Para os demais, “menor de idade”.

Portanto, classificar algo requer atribuir uma classe para um objeto com base em um conjunto de característica deste mesmo objeto. Normalmente isso ocorre com base em uma tabela de banco de dados ou arquivo, onde as colunas representam os valores para as características dos objetos classificados e uma última coluna, a sua respectiva classe. Quando tal conjunto (objeto com os valores de atributos e classe) possui a classe pré-definida, definimos como conjunto de treino. Sendo assim, com base nesse conjunto de treino, é possível classificar novas entradas baseadas nos exemplos conhecidos (BERRY; LINOFF, 2004, p. 9).

Para casos complexos de tomada de decisão, a classificação por si só não é suficiente,

> mas é um excelente norteador para tomada de decisão em atividades intensivas de conhecimento. Assim, ao buscar identificar o risco do cliente em cumprir suas obrigações, a técnica busca predizer o futuro. Por exemplo: com base na experiência passada, poderá estabelecer numa instituição financeira quais riscos/confiança para receber empréstimos concedidos (DA SILVA, *et al.*, 2018, p. 614).

5.4.1 Análise exploratória dos dados

A partir dos dados abertos publicados, realiza-se uma análise exploratória para identificar indícios de novos conhecimentos, baseada no comportamento das variáveis, a serem extraídas. Inicialmente a pergunta que se pretende responder é qual a relação, se é que ela existe, entre cada variável? Essa pergunta é necessária a fim de verificar se as variáveis apontadas pelos estudos aplicadas sobre os dados disponíveis na instituição possuem uma interação comportamental entre si. Para tal, é utilizado um conceito de estatística que trata do coeficiente de correlação.

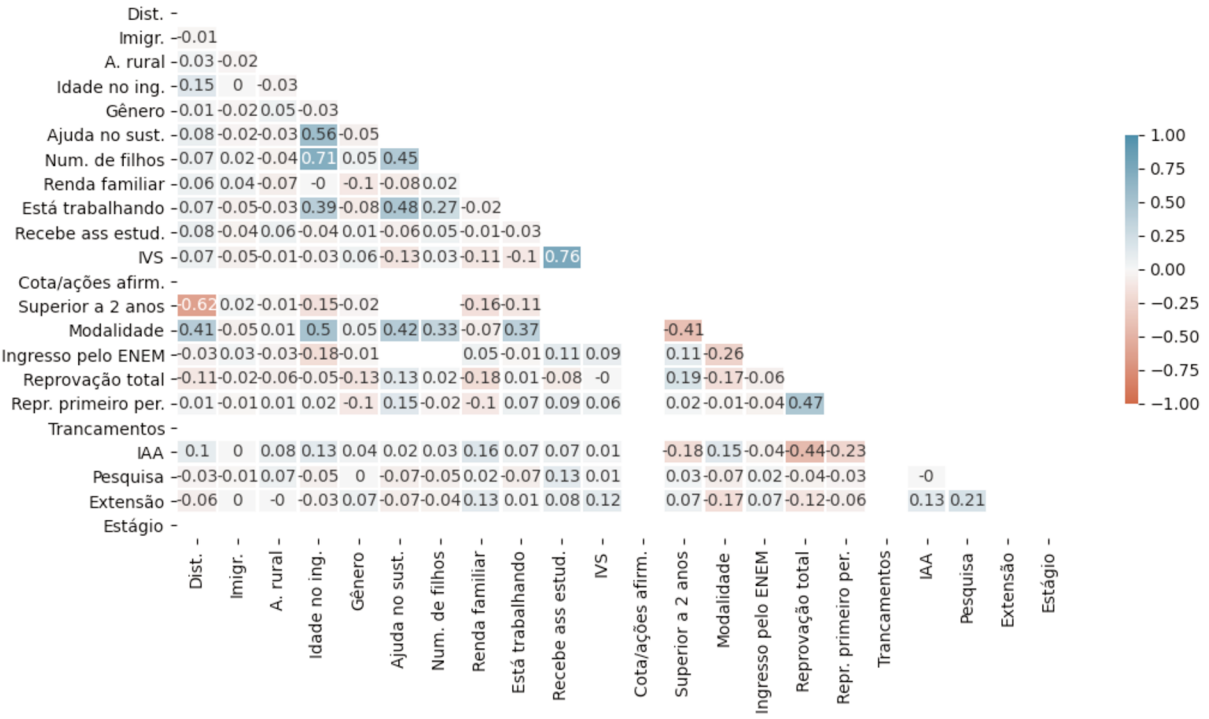
Partindo do conceito de matriz de correlação, avalia-se as variáveis para os grupos de discentes em situação de SUCESSO e EVADIDO, conforme Figura 29 e Figura 30. Foram utilizadas as bibliotecas *Pandas*[11], *statsmodels*[12] e *Matplotlib*[13] para Python que possuem as ferramentas necessárias para tal análise. É importante ressaltar que todas as variáveis possuem

[11] https://pandas.pydata.org
[12] https://www.statsmodels.org/stable/index.html
[13] https://matplotlib.org

valores dicotômicos ou ordinais e que as ferramentas tratam adequadamente a correlação entre tais variáveis, aplicando a análise mais adequada para cada caso:

- Ponto Bisserial: para correlação entre variáveis dicotômicas e dicotômicas ou dicotômicas e ordinais;
- Person para os demais casos.

Figura 29 - Correlação entre as variáveis para Sucesso



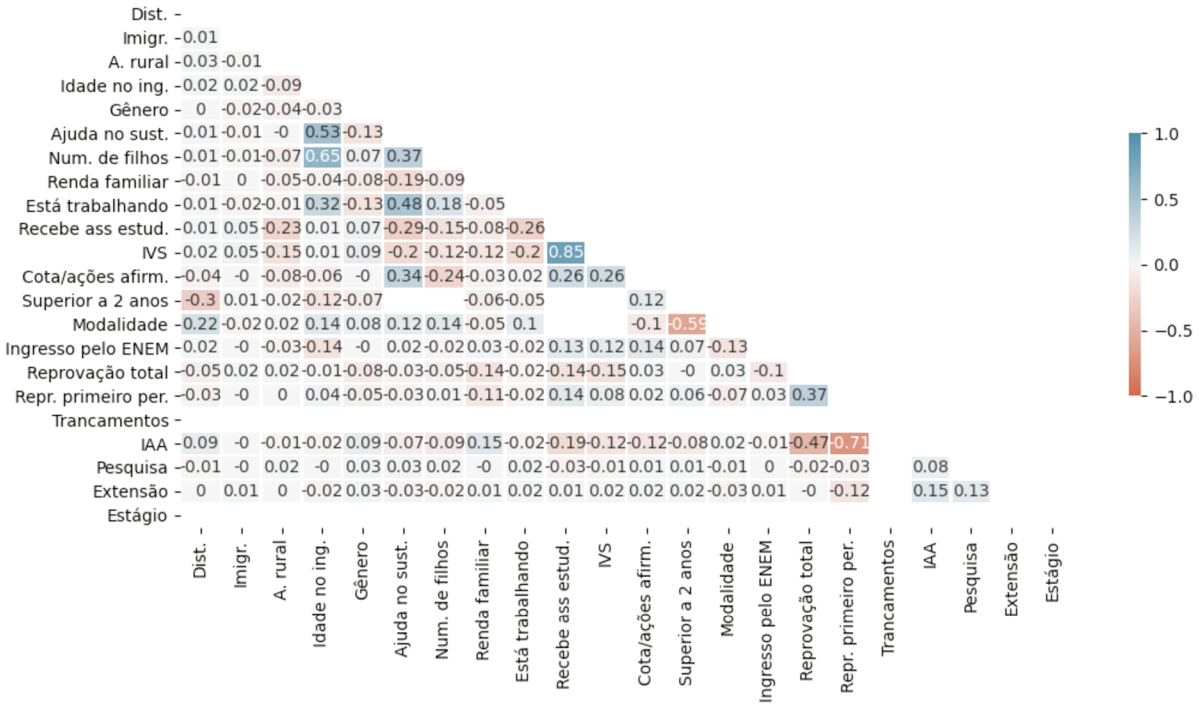
Fonte: do autor.

No que diz respeito às correlações entre os concluintes (sucesso), e de acordo com os níveis apresentados no "Quadro 4 – Nível de correlação por faixa de valor", existem 12 variáveis que possuem alguma correlação com outra variável de, no mínimo, nível médio (> 0,29): curso superior a 2 anos, modalidade, distância, ajuda no sustento da família, número de filhos, está trabalhando, idade no ingresso, índice de vulnerabilidade social (IVS), recebe assistência estudantil, reprovações no primeiro período e índice de aproveitamento acadêmico (IAA).

Para as "variáveis cotas/ações afirmativas" e "trancamentos" na visão dos discentes que tiveram sucesso (formados) a ferramenta não concluiu qualquer correlação. Por este motivo

tais dados estão em branco. Pontualmente o mesmo ocorreu com “superior a dois anos” e “ingresso pelo ENEM” que concluiu qualquer nível de correlação com “ajuda no sustento da família” e “número de filhos”.

Figura 30 - Correlação das variáveis para Evasão



Fonte: do autor.

Do ponto de vista da correlação dos evadidos (Figura 30), são 13 variáveis: curso superior a 2 anos, modalidade, distância, ajuda no sustento da família, número de filhos, está trabalhando, idade no ingresso, recebe assistência estudantil, cotas / ações afirmativas, índice de vulnerabilidade social (IVS), reprovações no primeiro período, reprovações totais e índice de aproveitamento acadêmico (IAA).

Depreende-se da análise das figuras presença de conjunto variáveis correlacionadas praticamente igual para os dois grupos: sucesso e evadido. A única diferença está na evasão que apresenta na sua matriz de correlação mais duas variáveis:

- cotas / ações afirmativas: tem correlação positiva com ajuda no sustento da família;
- reprovações totais: com o IAA (correlação negativa) e as reprovações no primeiro período (correlação positiva).

Além da variável “trancamentos”, cuja análise de correlação foi inconclusiva com qualquer outra variável, pontualmente “curso superior a 2 anos” e “modalidade” não possuem um nível de correlação com “número de filhos”, “ajuda no sustento da família”, “recebe assistência estudantil” e “IVS”. A modalidade teve pontualmente tais condições com as “recebe assistência estudantil” e “IVS”.

Logo, há uma pequena diferença entre as correlações nos dois grupos. No geral, a análise mostra que tem certo grau geral de correlação entre as variáveis para a instituição analisada.

5.4.2 *Machine Learning* para predição de evasão

Como prova de conceito no intuito de verificar a capacidade dos algoritmos de classificação de predizer se um discente irá ou não evadir a partir das variáveis identificadas, os dados são submetidos a algoritmos de classificação. Foram selecionados seis algoritmos disponíveis na biblioteca *sklearn*[14] que possui, não só esses, mas uma série de implementações voltadas para *machine learning*: *DecisionTreeClassifier* (árvore de decisão), *MultinomialNB* (estatístico utilizando *Naive Baÿes*), *LogisticRegression* (regressão logística multinominal), SVC (que utiliza *support vector machine*), *KNeighboardClassifier* (distância entre instâncias) e *MPLClassifier* (rede neural).

Além dos algoritmos, foram aplicados a estes apenas os valores das treze variáveis que apresentaram alguma correlação de, no mínimo, média (>0,29 conforme Quadro 4 – Nível de correlação por faixa de valor). Nos dados históricos de alunos formados e evadidos conhecidos – que totalizam 11.110 instâncias – foi utilizada a técnica de particionamento de dados para treino e teste dos algoritmos de duas formas:

- particionamento de 66% (7.332 instâncias) para treito e 33% (3.367 instâncias) para teste para matriz de confusão, curva ROC e AUC;
- validação cruzada particionada em 5 grupos (2.222 instâncias cada grupo) para verificação da acurácia, precisão, *recall* e *fscore*.

[14] https://scikit-learn.org/stable/

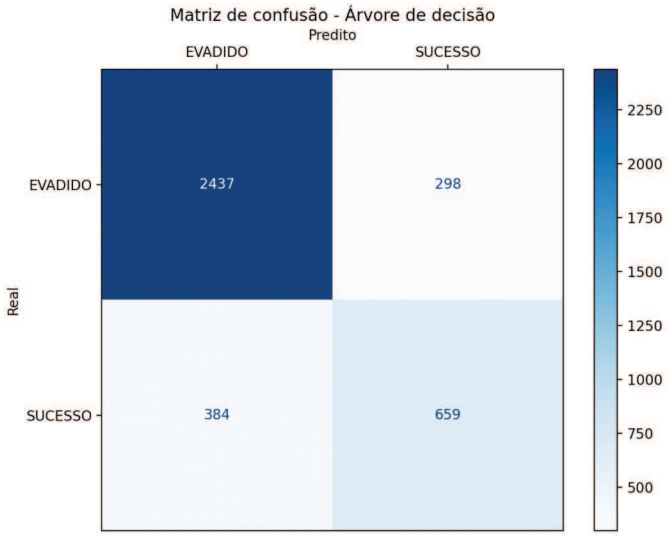
Na validação cruzada, supondo 3 partições, os dados são divididos em 3 partições de tamanho aproximadamente igual. Nesse cenário, dois terços dos dados serão inicialmente utilizados como massa de treino e um terço massa de teste, repetindo o procedimento e permutando as partes três vezes. Ao final do processamento da validação cruzada, cada instância do conjunto total terá sido usada exatamente uma duas vezes para treino e uma vez para teste (WITTEN; FRANK, 2002, p. 153).

5.4.2.1 Matriz de confusão

Após o processo de treino e teste de cada um dos algoritmos selecionados, obteve-se a matriz de confusão que possibilitará a realização das métricas de acurácia, precisão, *recall* e *fscore*. Para o estudo em questão, verdadeiro positivo é o discente ter realmente evadido e, ao submeter para a predição do algoritmo, foi classificado como evadido (quadrante superior esquerdo das figuras). Os verdadeiros negativos são os casos em que o discente não evadiu (sucesso) e o algoritmo classificou como sucesso (quadrante inferior direito das figuras).

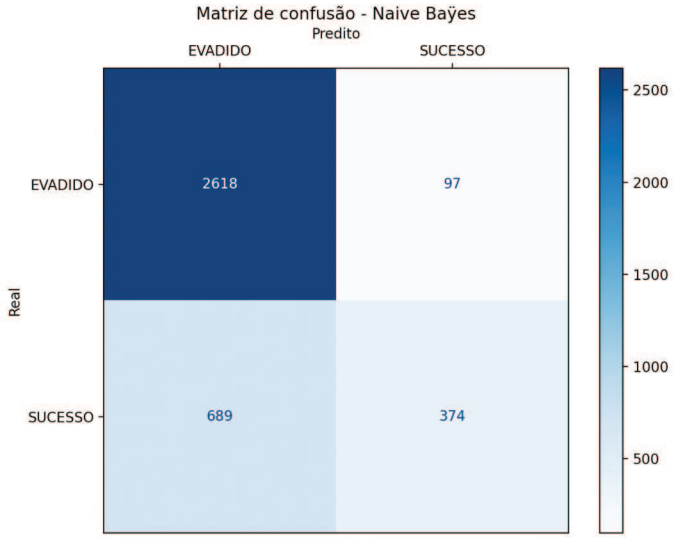
Das Figura 31 até a Figura 36, apresentam-se as matrizes de confusão de cada algoritmo, especificamente.

Figura 31 – Matriz de confusão - *DecisionTreeClassifier*



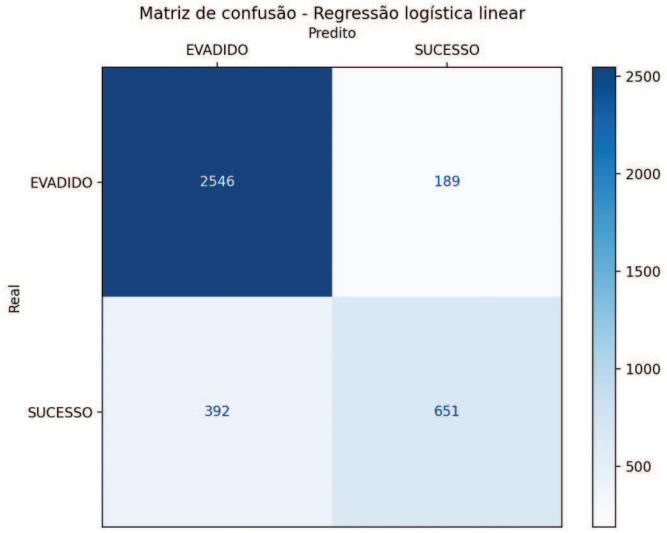
Fonte: do autor.

Figura 32 – Matriz de confusão - *MultinomialNB*



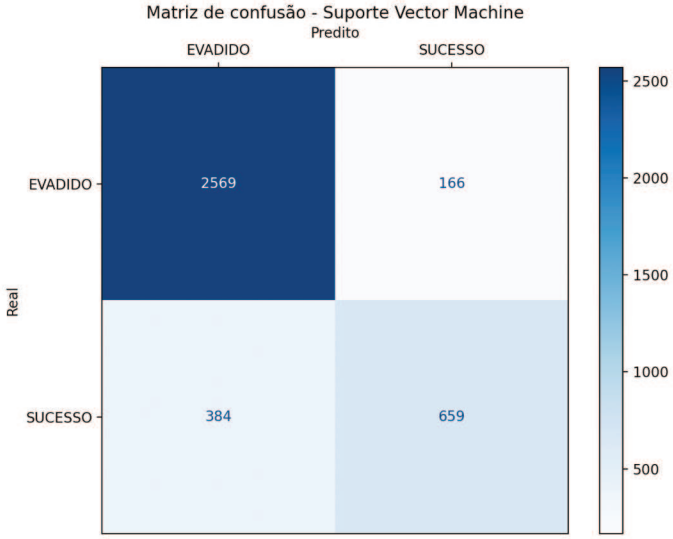
Fonte: do autor.

Figura 33 – Matriz de confusão - *LogisticRegression*



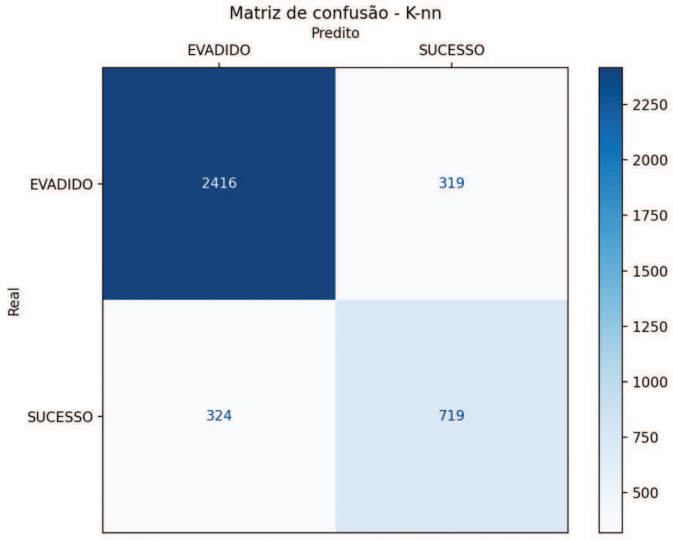
Fonte: do autor.

Figura 34 – Matriz de confusão - SVC



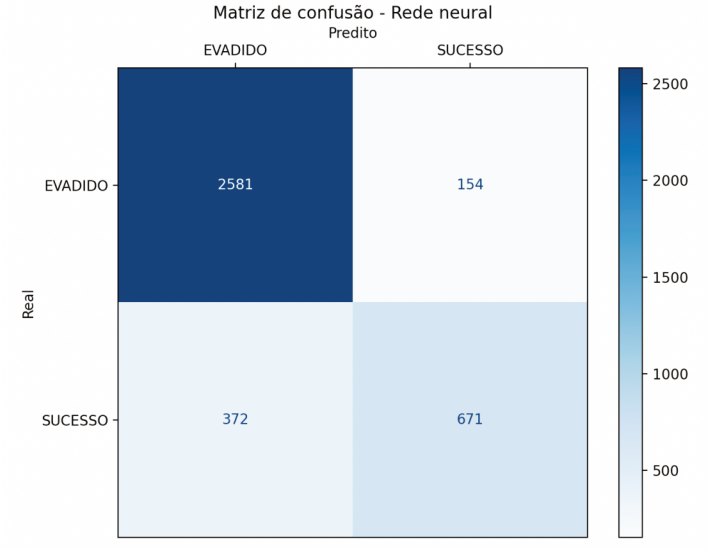
Fonte: do autor

Figura 35 – Matriz de confusão - *KNeighboardsClassifier*



Fonte: do autor.

Figura 36 – Matriz de confusão - *MPLClassifier*



Fonte: do autor.

As matrizes apresentam um comportamento geral da confiança de cada algoritmo quanto à predição. No entanto, somente tais dados não sugerem necessariamente qual deles tem uma maior confiança para predição diante da população (instâncias) utilizada para treino e testes.

O objetivo da análise é avaliar o quão confiáveis os algoritmos são em predizer que um discente possui tendência a evadir do curso, possibilitando que as áreas interessadas atuem antes que isto ocorra. Depreende-se das matrizes de confusão (da Figura 31 até a Figura 36), que no geral os algoritmos têm uma boa predição de evadidos frente aos falsos negativos para evadidos.

É possível perceber dentre os algoritmos os melhores, avaliando apenas as quantidades de:

- Melhor verdadeiro positivo: *MPLClassifier* (2581)
- Melhor verdadeiro negativo: *KNeighboardsClassifier* (719)
- Melhor falso positivo: *MPLClassifier* (324)
- Melhor falso negativo: *MultinomialNB* (97)

Mas é necessário verificar as demais métricas de qualidade que são derivadas da matriz.

5.4.2.2 Acurácia, precisão, *recall* e *f1-score*

A Tabela 5 apresenta as medidas de acurácia, precisão, *recall* e *f-score* para cada um dos algoritmos.

Tabela 5 – Acurácia, Precisão, Recall e F-score

	Acurácia	**Precisão**	**Recall**	**F-score**
DecisionTreeClassifier	0,82	0,79	0,76	0,76
MultinominalNB	0,64	0,67	0,70	0,62
LogisticRegression	0,84	0,83	0,77	0,79
SVC	0,85	0,85	0,78	0,80
KNeighboardClassifier	0,82	0,78	0,77	0,77
MPLClassifier	0,86	0,86	0,79	0,81

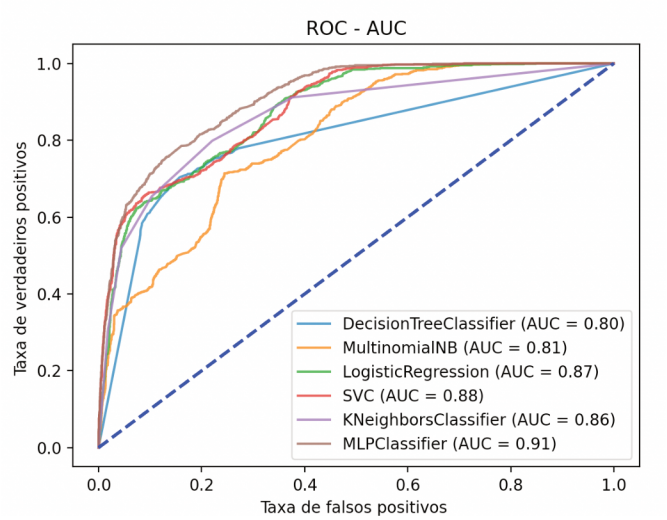
Fonte: do autor.

Apreciando os dados da Tabela 5, pode-se perceber que em termos de acurácia, todos, exceto o algoritmo baseado em *Naive Baÿes* obtiveram um valor expressivo próximo do 0,85. No quesito precisão (quando de VPs o algoritmo é capaz de prever frente aos VPs e FPs) três se destacam: *LogisticRegression, SVC* e *MPLClassifier*. Isto mostra que olhar apenas para a acurácia pode levar à seleção não muito adequada ao propósito. No quesito *recall*, novamente apenas o *MultinominalNB* ficou afastado dos demais. E, por fim a concordância específica (*fscore*) nos mostra, também, que *LogisticRegression, SVC* e *MPLClassifier* parecem mais adequados.

5.4.2.3 AUC e curva ROC

Além de verificar as quatro métricas acima, plota-se a curva de ROC e mede-se a AUC de cada algoritmo, como mostra a Figura 37.

Figura 37 – Curva ROC e AUX



Fonte: do autor.

A partir da análise anterior, a curva ROC corrobora com os dados mostrados pelas demais métricas. No entanto, visualmente pode-se perceber o destaque do algoritmo *MPLClassifier* frente aos demais. Tal algoritmo com a população de treino utilizada é capaz de alcançar uma taxa de verdadeiros positivos melhor que os demais, como resume o valor da área abaixo da curva (AUC) de 0,91, 0,03 (ou 3%) pontos maior que o segundo melhor algoritmo (*SVC*).

Tais conhecimentos extraídos – correlação e confiança na predição – passam a compor o modelo.

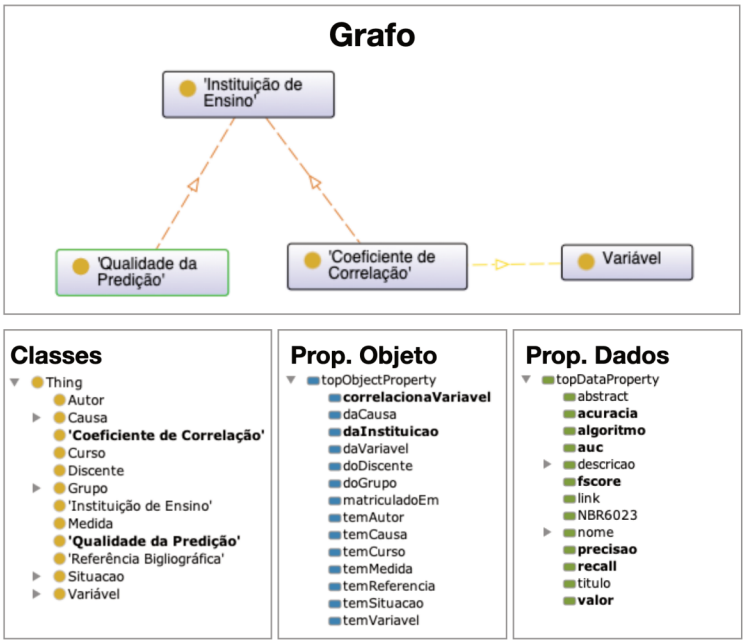
5.4.3 Representação do conhecimento extraído

As instâncias das medidas das 13 variáveis selecionadas foram capazes de dar os subsídios minimamente necessários para avaliar a correlação entre elas e, para a maioria dos algoritmos, um bom grau de confiança para predição da evasão.

Esses conhecimentos extraídos para a instituição de ensino pesquisada, no que tange ao coeficiente de correlação entre variáveis e os valores de acurácia, precisão, *recall*, *fscore* e AUC, são alimentados no modelo por meio de uma ontologia adicional (Figura 38). Os coeficientes de correlação são instâncias da classe "Coeficiente de Correlação" para cada par das instâncias "Variável" do modelo proposto. As demais métricas são representadas por

instâncias da classe “Qualidade de Predição”, sendo uma instância para cada um dos algoritmos com os valores das respectivas métricas alcançadas.

Figura 38 - Ontologia adicional: coeficiente de correlação e qualidade de predição da instituição



Fonte: do autor.

Com tal representação, o conhecimento é representado e compartilhado com a IES analisada, além de poder ser utilizado por outras IES considerando uma publicação dos dados abertos de forma mais ampla.

5.5 CONSIDERAÇÕES

O conhecimento representado pelo modelo possibilita um bom ponto de partida às instituições de ensino que buscam compreender a evasão. Por ser composto por estudos já realizados e publicados sobre o tema, suas causas e variáveis norteiam, por meio do modelo, as IES na coleta de dados relacionados ao que leva à evasão real.

Como prova de conceito, ao instanciar os indivíduos no modelo e aplicá-los aos algoritmos de *machine learning*, obtém-se resultados expressivos de métricas de qualidade na

predição da evasão. Isto se deve à boa aderência da seleção adequada das variáveis por meio da revisão bibliográfica, cujo conhecimento fora compilado no modelo.

Como evoluções importantes a serem buscadas para melhoria da extração de conhecimentos da instituição, atuar na qualidade dos dados, não somente quanto à "sujeira", se mostra relevante e necessário. Alguns pontos foram tratados – vide 5.2 Higienização do dados.

Ainda há dados parciais, ou de qualidade duvidosa, que devem ser avaliados pela instituição pesquisada no intuito de planejar estudos mais aprofundados e direcionamento para a correção. A evolução deve passar, também, pela busca de aquisição dos dados para as variáveis que a instituição não possui ainda.

No tocante ao compartilhamento do conhecimento, a ontologia irá contribuir para isto representando o conhecimento geral e, para a retroalimentação do conhecimento descoberto por meio do KDD específico para a instituição. Instituições podem fazer uso do modelo, seja para entender as principais causas de evasão como mostram os estudos contemplados, ou ir além instanciando os indivíduos e extraindo novos conhecimentos sobre a evasão na instituição.

Na instituição de ensino pesquisada, uma extensão da ontologia modelo foi definida para representar os novos conhecimentos extraídos das análises exploratórias e algoritmos de *machine learning*. Assim, agora, além do modelo teórico, a instituição tem também a correlação das variáveis e as alternativas de *machine learning* que podem ser aplicadas em um futuro SBC.

Durante a realização da prova de conceito, notou-se a necessidade de adicionar ao modelo a estrutura de dados intermediária para ser povoada por meio de ETL e o mapeamento do D2RQ para transformar os dados relacionais em triplas RDF e publicação no formato de dados ligados. Sendo assim, outra instituição que tenha interesse em fazer uso do conhecimento sintetizado no modelo teria que realizar os passos do KDD relacionados à seleção e limpeza dos seus dados – seguindo o modelo proposto –, alimentar a estrutura intermediária que comporá o modelo e utilizar o mapeamento existente para publicação – uso do D2RQ para converter relacional para dados abertos em RDF.

6 CONCLUSÕES E RECOMENDAÇÕES

A presente dissertação traz o tema evasão e suas causas, segundo estudos realizados de 2015 até 2019. Foi possível perceber a importância do tema para o Brasil e para outros países. No entanto, a baixa quantidade de produções no Brasil com este foco demonstra, também, que existe um campo amplo a ser explorado.

A partir do conhecimento das causas de evasão, faz-se uso da Engenharia do Conhecimento para propor um modelo que possibilite às instituições de ensino entender este processo e extrair novos conhecimentos de acordo com suas realidades.

O modelo proposto mostrou-se, por meio da prova de conceito realizada, uma representação consistente do tema no domínio pesquisado. De forma padronizada, por meio de ontologia, as instituições têm disponível a compreensão das principais causas que levam à evasão e, principalmente, quais são as variáveis – idade, reprovações, distância residência-campus, bolsa de estudo, entre outras – que estão relacionadas com estas causas. Definir as variáveis é importante para aplicá-las em processamento computadorizado, o que permite a extração de novos conhecimentos. Sendo assim, o modelo facilita o trabalho das instituições ao iniciar uma análise de evasão, pois as variáveis do modelo direcionam e agilizam a fase de seleção dos dados no processo KDD.

A revisão bibliográfica foi suficiente para o modelo proposto quanto à compreensão das causas e o apontamento das variáveis. Destaca-se ainda o quanto alguns estudos possuem aderência de causas entre si, como mostra o Quadro 6, que sugere serem as causas mais frequentes e, possivelmente, presentes em diversas instituições de ensino.

Sendo a revisão bibliográfica base para o modelo, o proposto fez uso de ontologia para a representação e compartilhamento do conhecimento adquirido nas publicações. Com as classes, propriedades, relações, restrições e instâncias-base, possibilita aos interessados a compreensão das causas da evasão e a sua aplicação em processamentos computacionais para identificação de padrões e predição de evasão.

Destaca-se, ainda, a possibilidade de expansão do conhecimento representado pelo modelo à medida que novas perguntas são identificadas, avaliadas e respondidas e novos estudos são publicados. Para conhecimentos descobertos e que são específicos da instituição, possibilita a especialização do modelo para a sua representação. Eventualmente, caso seja

identificado como um conhecimento de cunho comum ao tema, pode ser incorporado ao modelo comum, no futuro.

Ao longo do desenvolvimento da pesquisa verificou-se que a utilização de dados no formato aberto do modelo pode viabilizar o compartilhamento de informações com outras instituições e pesquisadores. Dentro dos limites da LGPD[15], tornar dados e conhecimentos uma fonte de dados abertos de informações para qualquer interessado realizar pesquisas sobre o tema se mostra promissor para o tema. Consequentemente, a publicação de todos os dados em formato aberto por várias instituições, viabilizará o uso não só do conhecimento modelado na ontologia, mas dos conhecimentos extraídos para cada instituição e compará-los, replicá-los e avaliá-los em populações diferentes por meio de meta-análises. Espera-se com isso promover a melhoria da compreensão do fenômeno e gerar novos conhecimentos.

Por fim, a prova de conceito aplicada a uma instituição de ensino mostrou que o modelo é um bom norteador para compreender as causas e ações de análise, partindo dos bancos de dados estruturantes da instituição e direcionando as análises a partir dele. No caso da PoC, as variáveis previamente selecionadas nos estudos são relevantes para análises de evasão, mesmo que a instituição não possua todas as informações disponíveis para alimentar o modelo na íntegra – foram utilizadas 23 variáveis do total de 37 disponíveis no modelo.

Por meio do KDD, a instanciação dos indivíduos prevista no modelo e a publicação padronizada no formato aberto facilita o compartilhamento de todo o conhecimento modelado. Sendo assim, de modo geral, as análises exploratórias e de *machine learning* apontam para uma boa aderência entre o que está publicado nos estudos inclusos nesta pesquisa e a realidade da instituição. As correlações entre as variáveis dos grupos “sucesso” e “evadido” dá à instituição a visão de como tais variáveis interagem no seu domínio. Já os algoritmos de *machine learning* mostraram confiança suficiente para, como trabalho futuro, a modelagem e desenvolvimento de um SBC para predição de evasão.

[15] Lei Geral de Proteção de Dados

6.1 RECOMENDAÇÕES

No geral, o modelo proposto atendeu as expectativas de ser suporte à análise das causas da evasão. No entanto, ressalta-se que um modelo "é apenas uma aproximação da realidade. Em princípio, o processo de modelagem é infinito, pois é uma atividade incessante com o objetivo de aproximar o comportamento pretendido" (STUDER, BENJAMINS; FENSEL, 1998, p. 3).

Como trabalho futuro, aproveitando-se dos dados no formato aberto, pretende-se propor a publicação do modelo, dos conhecimentos e os dados abertos, deixando acessíveis a outros pesquisadores e instituições. Assim, espera-se que instituições possam fazer a análise da evasão a partir do modelo e pesquisadores fazer uso de métodos, técnicas e ferramentas para descoberta de novos conhecimentos. Indo além, ao considerar o fato de que mais instituições façam compartilhamento de conhecimento e dados por meio do modelo, percebe-se que isto pode fomentar mais pesquisas e gerar novos conhecimentos. Desta forma, pode-se correlacionar dados entre instituições, comparar comportamentos entre elas e, até mesmo avaliar o comportamento de evasão em uma determinada região com 2 ou mais instituições que faça uso do modelo.

Quanto aos resultados de qualidade alcançados com os modelos preditivos, tal expressividade revelou que há campo para evolução e melhoria do modelo, além de evolução na predição com consequente implantação como um SBC à gestão de evasão na IES. Essa tendência vai ao encontro do desejo externado pela instituição sobre o tema (IFSC, 2018), o que é um grande fator motivacional.

Ainda como trabalhos futuros, avaliar as mudanças nas variáveis ao longo da vida acadêmica se faz necessário. Conseguir ou perder o emprego durante o curso, o aumento ou redução de renda, o nascimento de filho durante o curso, são alguns exemplos de mudanças que, hoje, não são identificadas formalmente e não são acompanhadas pela instituição, tornando tais variáveis pouco precisas perto do que poderiam ser.

Durante a escrita deste documento, estamos em um momento de pandemia, ocasionado pela COVID-19, fato que por certo irá mudar o comportamento dos discentes em relação à dedicação e continuidade dos seus estudos, alterando o "padrão de evasão". Isto reforça a afirmação de autores sobre a necessidade de rever o modelo à medida que novas observações são realizadas, fazendo com que o modelo esteja em constante aprimoramento

(STUDER; BENJAMINS; FENSEL, 1998) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Como consideração final, entende-se ser este trabalho uma contribuição importante, tanto para os estudos relacionados ao tema quanto para as instituições de ensino superior na compreensão e direcionamento de suas análises e ações.

REFERÊNCIAS

ADACHI, Ana Amélia Chaves Teixeira. Evasão e evadidos nos cursos de graduação da UFMG, Dissertação, 2009.

APOLINÁRIO, Fábio. **Metodologia da Ciência:** Filosofia e Prática da Pesquisa. 2. Edição. ed. São Paulo: Cengage Learning, 2012.

BARROS, Aparecida da Silva Xavier. Expansão da educação superior no Brasil: limites e possibilidades. **Educação & Sociedade**, v. 36, n. 131, p. 361-390, 2015.

BERNERS-LEE, Tim; HENDLER, James; LARISSA, Ora. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, 2002.

BERRY, Michael J.A.; LINOFF, Gordon S. **Data mining techniques:** for marketing, sales, and custumer relationship management. 2. Ed. ed. Indianapolis: Wiley Publishing, 2004.

BODIN, Romuald; ORANGE, Sophie. Access and retention in French higher education: student drop-out as a form of regulation. **British Journal of sociology of Education**, v. 39, n. 1, p. 126-143, 2018.

BORST, Nico Willem. Construction of engineering ontologies for knowledge sharing and reuse., 1999.

BOSAK, Jon; BRAY, Tim. XML and the Second-Generation Web. **Scientific Americam**, v. 280, n. 5, p. 89-93, 1999.

CAMPOS, Larissa Couto et al. Cotas sociais, ações afirmativas e evasão na área de Negócios: análise empírica em uma universidade federal brasileira. **Revista Contabilidade & Finanças**, v. 28, n. 73, 2017. p. 27-42.

CARREIRA, Pedro; LOPES, Ana Sofia. Drivers of academic pathways in higher education: traditional vs. non-traditional students. **Studies in Higher Education**, p. 1-16, 2019.

CISLAGHI, Renato. **Um modelo de sistema de gestão do conhecimento em um framework para a promoção da permanência discente no ensino de graduação**. Universidade Federal de Santa Catarina. Florianópolis. 2008.

COMARELLA, Rafaela Lunardi. **Educação superior a distância: evasão discente**. Universidade Estadual de Santa Catarina. Florianópolis. 2009.

COSTA, Francisco José da; BISPO, Marcelo de Souza; PEREIRA, Rita de Cássia de Faria. Dropout and retention of undergraduate students in management: a study at a Brazilian Federal University. **RAUSP Management Journal**, v. 53, n. 1, p. 74-85, 2018.

COSTA, Oberdan Santos da; GOUVEIA, Luis Borges. Modelos de Retenção de Estudantes: abordagens e perspectivas. **REAd: Revista Eletrônica de Administração**, Porto Alegre, v. 24, n. 3, p. 155-182, Dezembro 2018.

DA SILVA, Sérgio Nicolau et al. Brazil's Universoty Ranking: a prediction study with machine learning. **13th Internationa Forum on Knowledge asset dynamics**, Delft, v. 1, p. 609-620, 2018.

DE GUIMARÃES, Julio Cesar Ferro et al. A influência da inovação no ensino, qualidade e comprometimento sobre a retenção de alunos no ensino superior. **Revista Gestão Universitária na América Latina-GUAL**, v. 12, n. 1, p. 249-269, 2019.

DE LIMA, Franciele Santos; ZAGO, Nadir. Desafios conceituais e tendências da evasão no ensino superior: a realidade de uma universidade comunitária. **Revista Internacional de Educação Superior**, v. 4, n. 2, p. 366-286, 2018.

DIAS-SOBRINHO, José. Paradigmas e políticas de avaliação da educação superior: autonomia e heteronomia. **CLACSO: Consejo Latinoamericano de Ciencias Sociales**, Buenos Aires, v. 16, n. 2, p. 169-192, 2006.

DIOGO, Maria Fernanda et al. Percepções de coordenadores de curso superior sobre evasão, reprovações e estratégias preventivas. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, v. 21, n. 1, p. 125-151, 2016.

DORE, Rosemary; SALES, Paula Elizabeth Nogueira; CASTRO, Tatiana Lage de. Evasão nos cursos técnicos de nível médio da rede federal de educação profissional de Minas Gerais. In: _____ **Evasão na educação:** estudos, políticas e propostas de enfrentamento. Brasília: RIMEPES, 2014. p. 379-413.

EDUCAÇÃO. Lei No 9.394, de 20 de dezembro de 1996. **Estabelece as diretrizes e bases da educação nacional**, 1996.

FAYYAD, M. Usama. Data mining and knowledge discovery: making sense out of data. **IEEE Expert**, v. 11, n. 5, p. 20-25, 1996.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996.

FERNÁNDEZ, Mariano; GÓMEZ-PÉREZ, A.; JURISTO, Natalia. METHONTOLOGY: From Ontological Art Towards Ontological Engineering. **AAAI Technical Report**, p. 33-40, 1997.

FIGUEIREDO FILHO, Dalson Britto; SILVA JÚNIOR, José alexandre da. Desvendando os Mistérios do Coeficiente de Correlação de Pearson, 2009. Disponivel em: <http://bibliotecadigital.tse.jus.br/xmlui/bitstream/handle/bdtse/2766/2009_figueiredo_desvendando_misterios_coeficiente.pdf?sequence=1>. Acesso em: nov. 2019.

GIL, Antônio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. ed. [S.l.]: Editora Atlas SA, 2008.

GIL, Antônio Carlos. **Como elabora projetos de pesquisa**. 5a ed. ed. [S.l.]: Editora Atlas, 2010.

GOOGLE DEVELOPERS. Machine learning crash course. **Google Developers**, 2019. Disponivel em: <https://developers.google.com/machine-learning/crash-course/classification/accuracy>. Acesso em: 31 mai. 2020.

GRUBER, Thomas R. A translation approach to portable ontology specifications. **Knowledge acsition**, v. 5, n. 2, p. 199-220, 1993.

GRUBER, Thomas R. Toward principles for the design of ontologies used for knowledge sharing? **International journal of human-computer studies**, v. 43, n. 5-6, p. 907-928, 1995.

GUARINO, Nicola. **Formal Ontologies and Information Systems**. FOIS'98 Conference. Trento: [s.n.]. 1998. p. 2-5.

GUARNIERI, Fernanda Vieira; MELO-SILVA, Lucy Leal. Cotas Universitárias no Brasil: Análise de uma década de produção científica. **Psicologia Escolar e Educacional**, v. 21, n. 2, p. 183-193, 2017.

GUERRA, Maria das Graças Gonçalves Vieira. **Sistema Nacional de Avaliação da Educação Superior (SINAES):** avanções na qualidade da avaliação da educação superior no Brasil. Actas do XIV Colóquio Internacional de Psicologia e Educação. [S.l.]: Edições ISPA. 2019. p. 219-299.

HASSLER, Edgar et al. Identification of SLR Tool Needs: results of a community workshop. **Information and Software Tecnology**, v. 70, p. 122-129, 2016.

HOFFMAN, Ivan Londero; NUNES, Raul Ceretta; MULLER, Felipe Martins. As informações do Censo da Educação Superior na implementação da gestão do conhecimento organizacional sobre evasão. **Gestão & Produção**, São Carlos, v. 26, n. 2, p. 1-14, mai. 2019.

IBGE. Populaç!ao nos Censos Demográficos, segundo as Grandes Regioes e as Unidades de Federação - 1872/2010. **Instituto Brasileiro de Geografia e Estatística**, 2011. Disponivel em: <ftp://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Sinopse/Brasil/sinopse_brasil_tab_1_4.zip>. Acesso em: 02 set. 2019.

IBGE. Estimativas da população residente no Brasil e unidades da federação com data de referência em 1º de julho de 2019. **Instituto Brasileiro de Geografia e Estatística**, 2019. Disponivel em: <ftp://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2019/estimativa_TCU_2019_20191031.pdf>. Acesso em: 5 nov. 2019.

IFSC. Plano Estratégico de Permanência e Êxito dos Estudantes do IFSC - PPE-IFSC, 2018. Disponivel em:

<https://www.ifsc.edu.br/documents/23567/0/Plano+de+Permanência+e+Êxito/11b7634e-0c69-4056-9034-a40275ff9a0b>. Acesso em: 14 abr. 2019.

IFSC. Reunisão da Reitoria e Trofeu Maricotinha. **Blog da Reitora**, 2019. Disponivel em: <https://www.ifsc.edu.br/postagens/-/asset_publisher/hfovc6ZpW9YV/content/id/1143184>. Acesso em: 09 mar. 2019.

INEP. Investimento público direto em educação por estudante em valores reais, por nível de ensino de (2000 a 2015). **Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira**, 2015. Disponivel em: <http://inep.gov.br/indicadores-financeiros-educacionais>. Acesso em: 10 mar. 2019.

INEP. Baixa ocupação de vagas remanescentes revelada pelo Censo da Educação Superior inspira nova política do MEC para as Universidades Federais. **Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira**, 2018. Disponivel em: <http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/baixa-ocupacao-de-vagas-remanescentes-revelada-pelo-censo-da-educacao-superior-inspira-nova-politica-do-mec-para-as-universidades-federais/21206>. Acesso em: 13 jan. 2019.

INEP. Sinopse Estatística da Educação Superior 2018. **Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira**, 2019. Disponivel em: <http://download.inep.gov.br/informacoes_estatisticas/sinopses_estatisticas/sinopses_educacao_superior/sinopse_educacao_superior_2018.zip>. Acesso em: 2019 out. 2019.

INEP. **Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira**, 2019b. Disponivel em: <http://download.inep.gov.br/informacoes_estatisticas/sinopses_estatisticas/sinopses_educacao_superior/sinopse_educacao_superior_2018.zip>. Acesso em: 01 out. 2019.

INEP. IDEB. **Índice de Desenvolvimento da Educação Básica**, 2020. Disponivel em: <http://portal.ineo.gov.br/ideb>. Acesso em: 10 fev. 2020.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. **Dados Abertos Conectados:** Em busca da Web do Conhecimento. São Paulo: Novatec Editora, 2015.

JUNIOR, Jaime Souza Sales et al. Fatores Associados à Evasão e Conclusão de Cursos de Graduação Presenciais na UFES. **Meta: Avaliação**, Rio de Janeiro, v. 8, n. 24, p. 488-514, set./dez. 2016.

JUNIOR, José da Silva Santos; REAL, Giselle Cristina Martins. A evasão na educação superior: o estado da arte das pesquisas no Brasil a partir de 1990. **Avaliação: Revista da Avaliação da Educação Superior**, Sorocaba, v. 22, n. 2, p. 385-402, mai./ago. 2017.

KAMAL, Preet; AHUJA, Sachin. An ensembled-based model for prediction of academic performance of students in undergrad professional course. **Journal od Engeneering Design and Technology**, v. 17, n. 4, p. 769-781, ago. 2019.

KIMBALL, Ralph; CASERTA, Joe. **The data warehouse ETL toolkit:** practical techniques for extracting, cleaning, conforming, and delivering data. 1 ed. ed. [S.l.]: Wiley Publishing, Inc., 2004.

KNELLER, George Frederick. **A ciência como atividade humana.** Rio de Janeiro: Zahar, 1980.

LAKATOS, Eva Maria; MARCONI, Maria de Andrade. **Fundamentos de metodologia científica**. 5a ed. ed. São Paulo: Atlas, 2003.

LASSILA, Ora; SWICK, Ralph R. Resource Description Framework (RDF) Model and Syntax Specification. **W3C**, 1998. Disponivel em: <https://www.w3.org/TR/1998/WD-rdf-syntax-19980720/>. Acesso em: 01 dez. 2019.

MARTINS, Antônio Carlos Pereira. Ensino superior no Brasil: da descoberta aos dias atuais. **Acta Cirúrgica Brasileira**, São Paulo, v. 17, p. 04-06, 2002.

MEC. Altos índices de desistência na graduação revelam fragilidade do ensino médio. **Ministerio da Educação**, 2016. Disponivel em: <http://portal.mec.gov.br/ultimas-noticias/212-educacao-superior-1690610854/40111-altos-indices-de-evasao-na-graduacao-revelam-fragilidade-do-ensino-medio-avalia-ministro>. Acesso em: 15 jul. 2019.

MEC. Censo da Educação 2017: Divulgação dos principais resultados. **Ministério da Educação**, 2018. Disponivel em: <http://portal.mec.gov.br/docman/setembro-2018-pdf/97041-apresentac-a-o-censo-superior-u-ltimo/file>. Acesso em: 10 jun. 2019.

MEC. Bloqueio total do MEC nas Universidades foi de 3,4%. **Ministério da Educação**, 2019. Disponivel em: <http://portal.mec.gov.br/ultimas-noticias/33381-notas-oficiais/75781-bloqueio-total-do-mec-nas-universidades-foi-de-3-4>. Acesso em: 10 jun. 2019.

MEC. Cálculo do número de vagas. **Ministério da Educação**, 2019b. Disponivel em: <http://portal.mec.gov.br/cotas/sobre-sistema.html>. Acesso em: 7 dez. 2019.

MENEGHEL, Stela Maria. Considerações sobre o atual sistema de ensino superior no Brasil. **Revista Pesquisa e Debate em Educação**, v. 7, n. 1, p. 340-348, 2018.

MORGAN, Gareth. Paradigms, metaphors, and puzzle solving in organization theory. **Administrative science quarterly**, p. 605-622, 1980.

OECD. **Rethinking Quality Assurance for Higher Education in Brazil**. OECD Publishing. Paris, p. 184. 2018.

OWL WORKING GROUP. Web Ontology Language (OWL), 2012. Disponivel em: <https://www.w3.org/2001/sw/wiki/OWL >. Acesso em: 31 mar. 2019.

PACHECO, Andressa Sasaki Vasques. Evas!ao e permanência dos estudantes de um curso de administração do sistema Universidade Aberta do Brasil: uma teoria fundamentada em fatos e na gestao do conhecimento, Tese, 2010.

PACHECO, Roberto Carlos dos Santos. Coprodução em ciência, tecnologia e inovação: Fundamentos e visões. Interdisciplinaridade. In: PEDRO, J. M.; FREIRE, P. D. S. **nterdisciplinaridade:** Universidade e Inovação Social e Tecnológica. 1a ed. ed. [S.l.]: CRV Editora, 2016.

POWERS, David Martin. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, 2011.

QEDU. Evasão escolar. **QEDU Acadeia**, 2019. Disponivel em: <https://academia.qedu.org.br/censo-escolar/evasao-escolar/>. Acesso em: 13 mai. 2019.

RAMOS, Bárbara Maria Soares. Eficácia no uso de tecnologias para alavancar o aprendizado do idioma inglês no ensino médio, Dissertação, 2007.

RAUTENBERG, Sandro; STEIL, Andrea Valéria; TODESCO, José Leomar. Modelo de conhecimento para mapeamento de instrumentos da gestão do conhecimento e de agentes computacionais da engenharia do conhecimento baseado em ontologias. **Perspectiva em Cienência da Informação**, Belo Horizonte, v. 16, n. 3, p. 26-46, jul./set. 2011b.

RAUTENBERG, Sandro; TODESCO, José Leomar; STEIL, Andrea Valéria. Uma Ontologia para instrumentosda Gestão do Conhecimento e Agentes da Engenharia do Conhecimento. **Informação & Sociedade**, v. 21, n. 1, p. 111-128, jan. 2011.

ROMERO, Cristóbal; VENTURA, Sebastián; GARCÍA, Enrique. Data mining in course management systems: Moodle case study and tutorial. **Computers & Education**, v. 51, n. 1, p. 368-384, 2008.

SACCARO, Alice; FRANÇA, Marco Túlio Aniceto; JACINTO, Paulo de Andrade. Fatores Associados à Evasão no Ensino Superior Brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de Ciência, Matemática e Computação e de Engenharia, Produção e Construção em instituições públicas e privadas. **Estudos Economicos (São Paulo)**, São Paulo, v. 49, n. 2, p. 337-373, 2019.

SCHREIBER, Guus et al. **Knowledge Engineering and Management:** The CommonKADS Methodology. [S.l.]: The MIT Press, 1999.

SHUE, Li-Yen; CHEN, Ching-Wen; SHIUE, Weissor. The development of an ontology-based expert system for corporate financial rating. **Expert Systems with Applications**, v. 36, n. 2, p. 2130-2142, 2009.

SILVA FILHO, Roberto Leao Lobo et al. A evasão no ensino superior brasileiro. **Cadernos de Pesquia**, v. 37, n. 132, p. 641-659, 2007.

SILVA, Marco Antônio da. Prova de conceito (PoC) em projetos. **PMBK**, 2014. Disponivel em: <https://pmkb.com.br/artigos/prova-de-conceito-poc-em-projetos/>. Acesso em: 3 jul. 2020.

STALLIVIERE, Luciane. **O sistema de ensino superior do Brasil características, tendências e perspectivas**. Universidade de Caxias do Sul. [S.l.], p. 22. 2007.

STOESSEL, Katharina et al. Sociodemographic diversity and distance education: Who drops out from academic programs and why? **Research in Higher Education**, v. 56, n. 3, p. 228-246, 2015.

STUDER, Rudi; BENJAMINS, V. Richard; FENSEL, Dieter. Knowledge engineering: Principles and methods. **Data & Knowledge Engineering**, v. 25, n. 1-2, p. 161-197, mar. 1998.

SURE, York; STAAB, Steffen; STUDER, Rudi. On-to-knowledge methodology (OTKM). In: _____ **Handbook on ontologies**. [S.l.]: Springer, Berlin, Heidelberg, 2004. p. 117-132.

SURE, York; STUDER, Rudi. On-To-Knowledge methodology - final version, 2002.

TINTO, Vicent. Dropout from higher education: A theoretical synthesis of recent research. **Review pf educational research**, v. 45, n. 1, p. 89-125, 1975.

TODESCO, José Leomar et al. **ontoKEM:** A web tool for ontologies' construction and documentation. IKE. [S.l.]: [s.n.]. 2009. p. 86-92.

TORRES-CORONAS, Teresa; VIDAL-BLASCO, María-Arántzazu. MOOC y modelos de aprendizaje combinado. Una aproximación práctica. **RIED. Revista Iberoamericana de Educación a Distancia**, v. 22, n. 2, p. 325-343, 2019.

TRUTA, Camelia; PARV, Luminita; TOPALA, Ioana. Academic Engagement and Intention to Drop Out: Levers for Sustainability in Higher Education. **Sustainability**, v. 10, n. 12, p. 4637, 2018.

VAN PETTEN, Adriana Maria Valladão Novais; DA COSTA ROCHA, Terezinha Cristina; BORGES, Adriana Araújo Pereira. Política de cotas na universidade federal de minas gerais: uma análise do perfil dos alunos com deficiência. **Regista Diálogos e Perspectivas em Educação Especial**, v. 5, n. 1, p. 127-140, 2018.

VASCONCELOS, Matheus Ferreira et al. Uso de datos abiertos y técnicas de minería de datos para la clasificación de estudiantes de instituciones privadas de educación superior de Belem-PA. **Research in Computiong Science**, v. 147, p. 27-36, 2018.

VENEGAS-MUGGLI, Juan I. Higher education dropout of non-traditional mature freshmen: the role of sociodemographic characteristics. **Studies on Continuing Education**, p. 1-17, 2019.

VIANNA, Cléverson Tabajara. Classificação das pesquisas científicas - Notas dos alunos, 2013. Disponivel em: <http://www.tabajara.tv/wp/wp-content/uploads/2016/01/MY-Classificação-dos-tipos-de-pesquisa-QUADRO-RESUMO-V31.pdf>. Acesso em: 13 jun 2019.

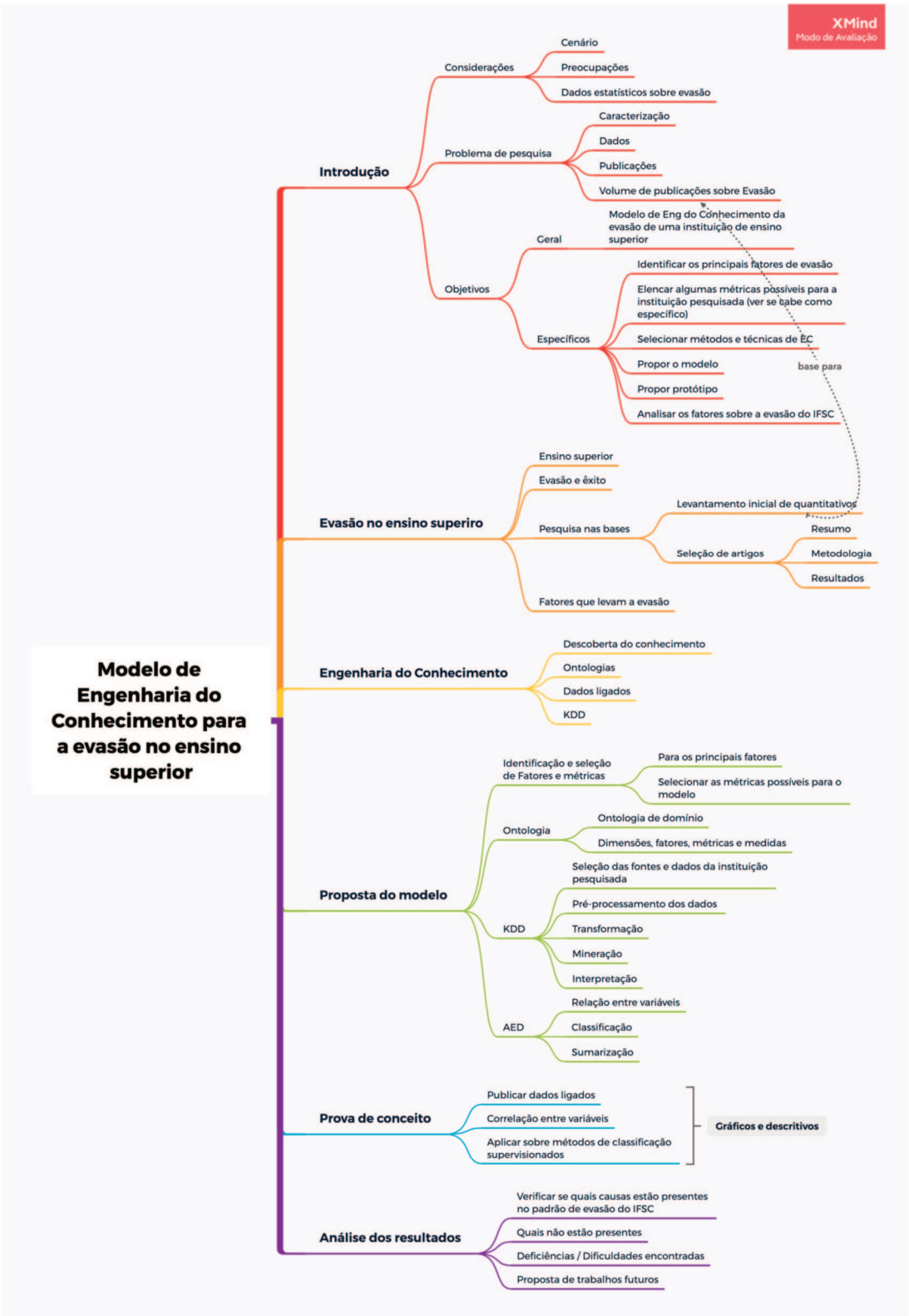
W3C. OWL Web Ontology Language Overview. **W3C**, 2004. Disponivel em: <https://www.w3.org/TR/2004/REC-owl-features-20040210/>. Acesso em: 10 jul. 2019.

W3C. Web Semântica. **W3C**, 2011. Disponivel em: <https://www.w3c.br/Padroes/WebSemantica>. Acesso em: 03 nov. 2019.

WITTEN, Ian H.; FRANK, Eibe. Data mining: practical machine learning tools and techniques with Java implementations. **Acm Sigmod Record**, v. 31, n. 1, p. 76-77, 2002.

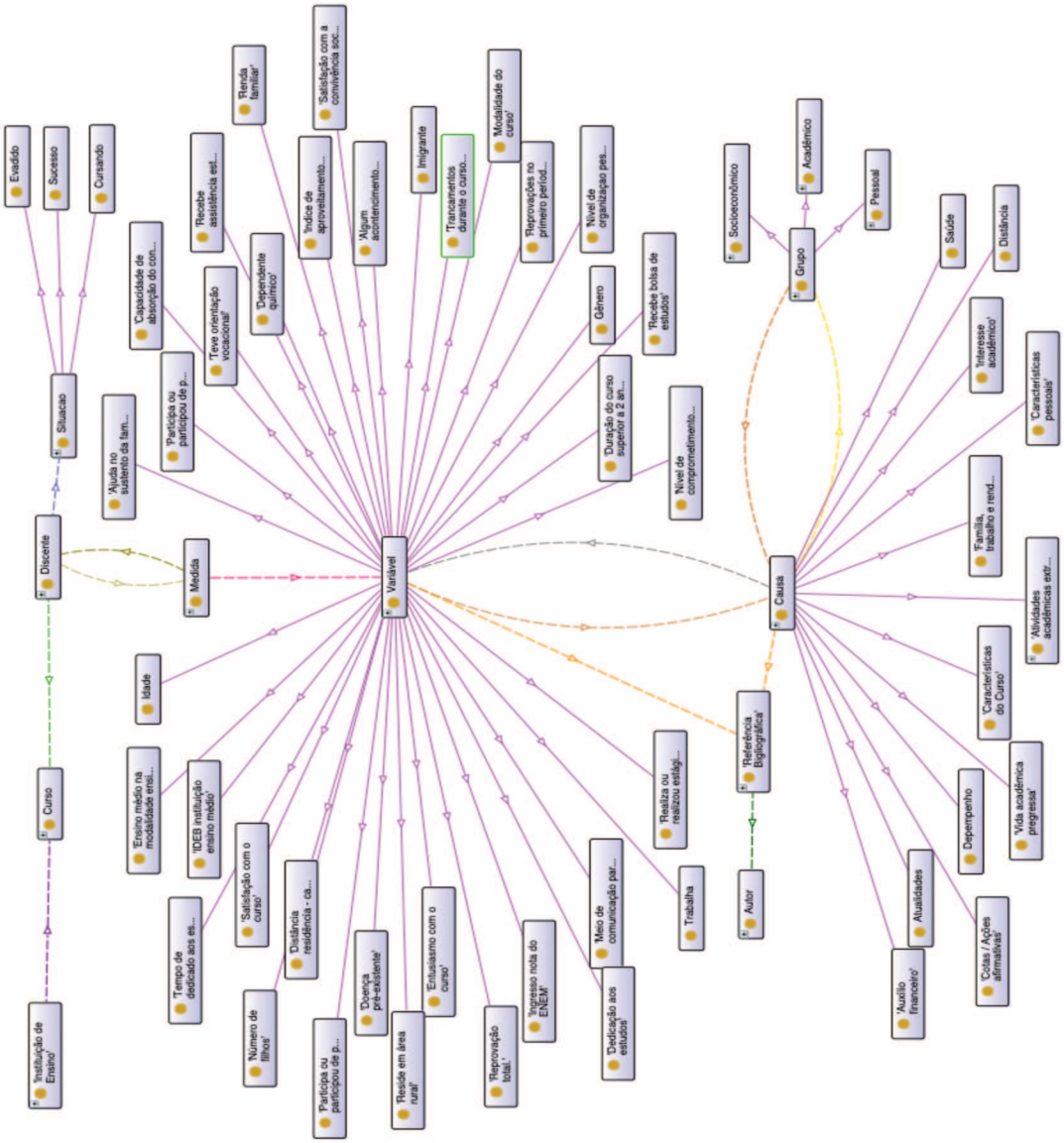
WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A. **Data Mining - Practical Machine Learning Tools and Techniques**. 3 ed. ed. [S.l.]: Elsevier, 2011.

ANEXO A – Mapa mental



ANEXO B – Ontologia completa

Arquivo owl disponível: https://www.dropbox.com/s/umox185e9ueg4ni/modelo-evasao-final.owl?dl=0.



ANEXO C – Arquivo de Mapeamento Relacional para RDF

```
# D2RQ Namespace
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
@prefix d2r: <http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2r-server/config.rdf#> .

# JDBC Namespace
@prefix jdbc: <http://d2rq.org/terms/jdbc/> .

# Namespace da ontologia
@prefix evasao: <http://dados.ifsc.edu.br/ontologia/evasao#> .

# Namespace para o arquivo de mapeamento (utilizado apenas para a carga, não é incluído na
ontologia)
@prefix map: <file:///Users/sergions/Dropbox/PPGEGC/Dissertação/workspace/d2rq/modelo-evasao-
carga.ttl#> .

# RDF Schema amespace
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

# XSD Schema namespage
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

# Conexão com o banco de dados
map:DbDissertacao a d2rq:Database;
  d2rq:jdbcDSN "jdbc:postgresql://localhost:5432/evasao";
  d2rq:jdbcDriver "org.postgresql.Driver";
  d2rq:username "dissertacao";
  d2rq:password "senha1234";
  jdbc:currentSchema "evasao";
  .
# Situacao
map:Situacao  a d2rq:ClassMap;
  d2rq:dataStorage map:DbDissertacao;
  d2rq:class evasao:Situacao;
  d2rq:uriPattern "http://dados.ifsc.edu.br/ontologia/evasao#@@evasao.situacao.identificador@@";
  d2rq:condition "evasao.situacao.identificador is not null";
  .
map:nomeSituacao a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:Situacao;
  d2rq:property evasao:nomeSituacao;
  d2rq:column "evasao.situacao.situacao_consolidada";
  d2rq:datatype xsd:string;
  .
#Variavel
map:Variavel a d2rq:ClassMap;
  d2rq:dataStorage map:DbDissertacao;
  d2rq:class evasao:Variavel;
```

```
  d2rq:uriPattern "http://dados.ifsc.edu.br/ontologia/evasao#@@evasao.variavel.identificador@@";
  .
map:nomeVariavel a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:Variavel;
  d2rq:property evasao:nomeVariavel;
  d2rq:column "evasao.variavel.nome";
  d2rq:datatype xsd:string;
  .
# Instituição
map:InstituicaoEnsino a d2rq:ClassMap;
  d2rq:dataStorage map:DbDissertacao;
  d2rq:class evasao:InstituicaoEnsino;
  d2rq:uriPattern "http://dados.ifsc.edu.br/ontologia/evasao#@@evasao.instituicao.identificador@@";
    .
map:temCurso a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:InstituicaoEnsino;
  d2rq:property evasao:temCurso;
  d2rq:refersToClassMap map:Curso;
  d2rq:join "evasao.instituicao.id => evasao.curso.id_instituicao";
  .
map:nomeInstituicao a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:InstituicaoEnsino;
  d2rq:property evasao:nomeInstituicao;
  d2rq:column "evasao.instituicao.nome";
  d2rq:datatype xsd:string;
  .

# Curso
map:Curso a d2rq:ClassMap;
  d2rq:dataStorage map:DbDissertacao;
  d2rq:class evasao:Curso;
  d2rq:uriPattern "ehttp://dados.ifsc.edu.br/ontologia/evasao#@@evasao.curso.identificador@@";
  .
map:nomeCurso a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:Curso;
  d2rq:property evasao:nomeCurso;
  d2rq:column "evasao.curso.nome";
  d2rq:datatype xsd:string;
  .

# Discente
map:Discente a d2rq:ClassMap;
  d2rq:dataStorage map:DbDissertacao;
  d2rq:class evasao:Discente;
  d2rq:uriPattern "http://dados.ifsc.edu.br/ontologia/evasao#@@evasao.discente.identificador@@";
  d2rq:join "evasao.discente.id_curso = evasao.curso.id";
  d2rq:join "evasao.curso.id_tipo = evasao.tipo_curso.id";
```

```
  d2rq:join "evasao.discente.id_situacao => evasao.situacao.id";
  d2rq:condition "evasao.situacao.identificador is not null";
    .
map:temSituacao a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:Discente;
  d2rq:property evasao:temSituacao;
  d2rq:refersToClassMap map:Situacao;
  d2rq:join "evasao.discente.id_situacao => evasao.situacao.id";
  .

#Medida
map:Medida a d2rq:ClassMap;
  d2rq:dataStorage map:DbDissertacao;
  d2rq:class evasao:Medida;
  d2rq:uriPattern "http://dados.ifsc.edu.br/ontologia/evasao#Medida@@evasao.medicoes.id@@";
  d2rq:join "evasao.medicoes.id_discente = evasao.discente.id";
  d2rq:join "evasao.discente.id_curso = evasao.curso.id";
  d2rq:join "evasao.discente.id_situacao => evasao.situacao.id";
  d2rq:join "evasao.curso.id_tipo = evasao.tipo_curso.id";
  d2rq:condition "evasao.situacao.identificador is not null";
  .
map:doDiscente a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:Medida;
  d2rq:property evasao:doDiscente;
  d2rq:refersToClassMap map:Discente;
  d2rq:join "evasao.medicoes.id_discente => evasao.discente.id";
  .
map:daVariavel a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:Medida;
  d2rq:property evasao:daVariavel;
  d2rq:refersToClassMap map:Variavel;
  d2rq:join "evasao.medicoes.id_variavel => evasao.variavel.id";
  .
map:valor a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:Medida;
  d2rq:property evasao:valor;
  d2rq:column "evasao.medicoes.valor";
  d2rq:datatype xsd:double;
  .
```