Estudo sobre a percepção de aspectos que contribuíram para a formação pessoal e profissional de graduandos, em cursos tecnológicos superiores presenciais

Luciana Passos da Silva [1] , Seiji Isotani [2] , Carlos Diego Nascimento Damasceno [3]

Resumo

Diversos jovens chegam à Instituições de Ensino Superior todos os anos, mas como a formação superior está impactando na vida destes jovens? Na tentativa de encontrar quais fatores podem se tornar facilitadores na formação pessoal e profissional dos graduandos de cursos técnicos presenciais do estado de São Paulo, usaremos mineração de dados nas respostas dadas ao ENADE em 2018, procurando relações entre as condições oferecidas pelas instituições de ensino e a percepção de formação integral do graduando ao término do curso. Inicialmente utilizamos técnicas de Gain Ratio para selecionar as questões mais relevantes em relação a este item, e posteriormente a aplicação de regras de associação sobre as variáveis selecionadas foi feita com o algoritmo Apriori. Verificamos que muitas das condições associadas ao sentimento de formação integral dos graduandos tem relação com docentes que deram aulas durante o curso, com o domínio do conteúdo pelos professores e uso de tecnologia da informação aplicada à educação.

Abstract

Many young people come to Higher Education Institutions every year, but how is higher education impacting the lives of these young people? In an attempt to find out what factors can become facilitators in the personal and professional qualification of undergraduate technical courses in the state of São Paulo, we use data mining in the answers given to ENADE in 2018, looking for relationships between the conditions offered by educational institutions and the perception of qualification of the student at the end of the course. Initially, we used Gain Ratio techniques to select the most relevant questions in relation

[1] Pós-Graduanda em Computação Aplicada à Educação, USP, <lupassos@usp.br>.

[2] Seiji Isotani, <USP>, <sisotani@icmc.usp.br>.

[3] Carlos Diego Nascimento Damasceno, <USP>, <damascenodiego@alumni.usp.br>.

Cite as: Silva, L. & Isotani, S. & Damasceno, C. (2020). Estudo sobre a percepção de aspectos que contribuíram para a formação pessoal e profissional de graduandos, em cursos tecnológicos superiores presenciais. Anais dos Trabalhos de Conclusão de Curso. Pós-Graduação em Computação Aplicada à Educação Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo.

to this item, and later the application of association rules on the selected variables was done with the Apriori algorithm. We found that many of the conditions associated with integral qualification of undergraduates are related to teachers who taught during the course, with the mastery of content by teachers and the use of information technology applied to education.

1. Introdução

O desejo de cursar uma instituição de nível superior é, para muitos jovens, ainda um sonho distante, em algumas situações por condições socioeconômicas não favoráveis, que levam o jovem ao ingresso no mercado de trabalho, ou ainda por distancia da escola e em alguns casos, problemas familiares, como apresentado em [Arruda 2019].

Várias instituições públicas e privadas tentam reverter esse quadro, atuando para motivar e ajudar os jovens a permanecer na escola, e tentando fazê-los entender que a educação é, e deve ser, para todos.

Mas sobre os jovens que efetivamente chegam a essa fase, não são encontrados estudos sobre o nível de satisfação quanto a formação recebida, e se essa formação realmente possibilitou a satisfação pessoal e profissional tão esperada. Assim, almejar um diploma superior é algo que os jovens fazem, mas sem certezas se a educação irá realmente alterar sua vida.

Usualmente são mostrados casos de superação, onde estudantes com baixas condições econômicas se formam em instituições renomadas e com esse título conseguem auxiliar suas famílias. Entretanto, o que ocorre com a maioria dos jovens formados? Isso aparentemente não é objeto de muitos estudos.

O passo inicial para a caminhada do jovem no ensino superior é selecionar um curso e uma instituição de estudo, e muitas vezes, essa escolha da instituição de ensino é feita com base na proximidade da casa do estudante, outras vezes pela oferta de bolsas de estudo ou auxílios de permanência estudantil.

A pergunta que surge nesse contexto é, escolher um curso ou instituição apenas por fatores econômicos pode levar o jovem a frustração acadêmica, e consequentemente a evasão escolar?

Alguns estudos têm se debruçado sobre a questão da evasão escolar nos diversos níveis educacionais, como citaremos na Seção 3, mas a tentativa de se adiantar ao problema e tornar a educação mais satisfatória para os alunos parece ser uma forma mais simples de manter os jovens nas instituições de ensino, como sugere [Carrano, Albergaria, Infante et al 2019].

Assim, na esperança de dar meios para que o jovem faça sua escolha de instituição superior com melhores condições, decidimos buscar respostas através da base de dados do ENADE. Acreditamos que a mineração de dados pode nos dar ideias das condições mais propícias para que o graduando se sinta melhor formado pessoal e profissionalmente.

Com isso, esperamos encontrar fatores que influenciam na satisfação com a formação integral, e essa informação pode ser utilizada tanto para estudantes que vão selecionar as instituições onde desejam estudar, quanto para gestores da educação.

No caso da gestão institucional, esta pode ser uma importante ferramenta para que as instituições analisem os fatores que os estudantes apontam como mais relevantes para

sua satisfação e com isso focar seus investimentos para tornar as instituições mais atrativas.

Este trabalho está organizado da seguinte forma: Na Seção 2 temos a fundamentação teórica do estudo; a Seção 3 apresentamos alguns trabalhos relacionados ao tema; e em seguida na Seção 4 mostramos os passos de seleção, transformação e agrupamento que efetuamos na base de dados. Continuamos na Seção 5, com a avaliação dos resultados obtidos durante o estudo e discutimos possíveis entraves a validade deste estudo, na Seção 6. Finalmente na Seção 7 apontamos nossas conclusões, de acordo com os dados obtidos no trabalho.

Esperamos, com essa contribuição, que no futuro a vasta base de dados do ENADE possa ser melhor estudada para tornar os cursos superiores mais significativos para os estudantes.

2. Fundamentação Teórica

A Mineração de Dados se baseia em conceitos advindos da estatística e também de teorias de aprendizagem da inteligência artificial, tendo suas tarefas divididas nas categorias de Predição e Descrição [Campos Neto 2016].

Prosseguindo com conceitos de [Campos Neto 2016], por um lado, as tarefas preditivas adquirem conhecimento ao analisar um banco de dados históricos, e a partir desse conhecimento, buscam prever o comportamento em novas amostras de dados. Por outro lado, as tarefas descritivas se focam na identificação de padrões de comportamento nas relações entre os dados. Entretanto, não existe uma clara separação entre essas tarefas, uma vez que modelos preditivos podem ser considerados também como descritivos em determinados momentos.

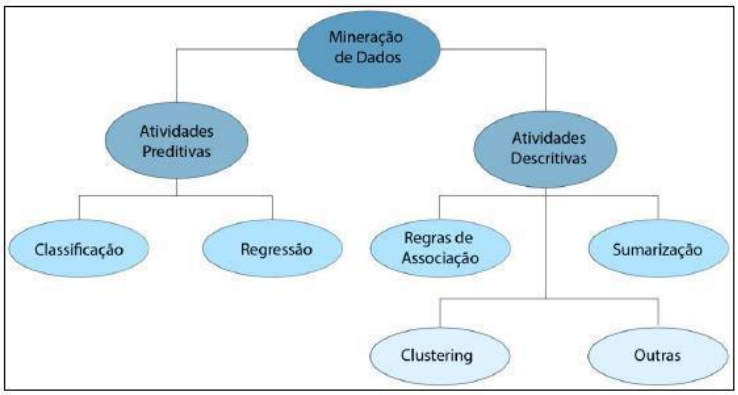


Figura 2.1. Tarefas de Mineração de Dados [Campos Neto 2016]

Das atividades preditivas, a Classificação é utilizada para estimar modelos que descrevam regras entre os dados existentes (também chamado de conjunto de treinamento). Diz-se que este é um aprendizado supervisionado porque os dados devem ser previamente categorizados para uso. A Regressão é similar à Classificação, porém utilizada para prever dados representados por valores numéricos contínuos.

Clustering é um método que busca identificar classes similares, ou seja, classes que tem propriedades similares entre si e diferente de outros grupos. Essa atividade não dispõe de dados categorizados. Quando estes existem, é sempre melhor buscar a Classificação.

Sobre as atividades preditivas, as Regras de Associação buscam identificar os atributos relacionados, demonstrando o quanto a existência de um conjunto de dados implica na presença de um outro conjunto de dados, apresentando relações na forma: SE (existe X) ENTÃO (existe Y). Enquanto isso, a Sumarização busca uma relação compacta no conjunto de dados, por exemplo a média ou desvio padrão.

A utilização da Mineração de Dados em contexto educacional é feita, de acordo com [Carrano, Albergaria, Infante et al. 2019] principalmente para tentar estudar a evasão de estudantes. Diversas técnicas já foram empregadas para buscar soluções neste sentido. Normalmente o objetivo destes trabalhos é de predição de indícios que levam os jovens ao abandono da escola.

De acordo com os estudos de [Lima, Ambrósio, Ferreira et al 2019], análises das bases de dados do ENADE e ENEM são feitos principalmente sobre dados socioeconômicos e notas dos estudantes.

Na vertente educacional, como visto em [Elias, Isotani and Penteado 2017], muitos dados podem ser obtidos de diversas fontes, tais como instituições de ensino, provas de aplicação regional ou nacional, sistemas de informação de escolas, ambientes virtuais de aprendizagem, e também dados disponibilizados pelo Ministério da Educação.

A fonte de dados sobre alunos concluintes do ensino superior é o Exame Nacional de Avaliação de Desempenhos do Estudantes (ENADE). O ENADE é uma prova aplicada a alunos de diversas instituições de ensino, de todo o Brasil, e tem como objetivo apurar os conhecimentos específicos e gerais obtidos pelos estudantes durante sua formação, e com isso atribuir notas aos diversos cursos existentes.

Esta avaliação, segundo [Silva, Rocha and Fagundes 2017] é muito importante no cenário brasileiro porque suas questões são desenvolvidas por profissionais de variados domínios educacionais, e levam em conta características do curso, não dos alunos.

Mas, a partir de tantos dados existentes e disponíveis, como transformá-los conhecimento? Uma opção é a utilização da técnica de Descoberta de Conhecimentos em Base de Dados (*Knowledge Discovery in Databases* – KDD). Esse processo é um conjunto de atividades, não triviais e contínuas, onde a última etapa abrange o tratamento do conhecimento, viabilizando o conhecimento descoberto.

De acordo com [Fayyad et al 1996], o processo é composto pelas etapas de: seleção dos dados; pré-processamento e limpeza dos dados; transformação dos dados; Mineração de Dados (*Data Mining*) e por último, a interpretação e avaliação do resultado. Através desses passos, o conjunto de dados se transforma em conhecimento útil, e pode ser utilizado como auxiliar na tomada de decisões.

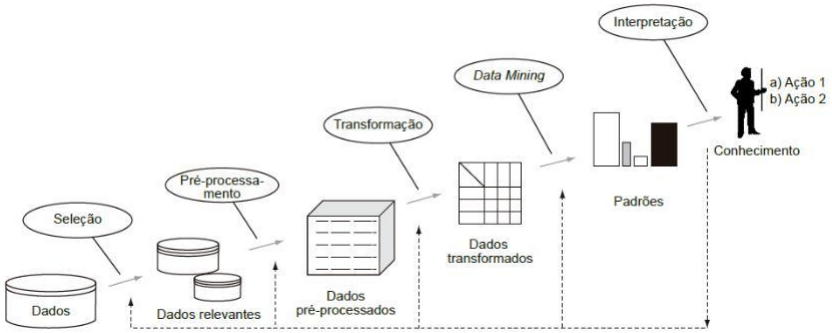


Figura 2.2. Etapas do Processo KDD [Fayyad et al 1996]

Como vimos anteriormente, obter conhecimento sobre um banco de dados é um processo extenso, que pode utilizar diversas técnicas e combinações de transformações e algoritmos, mas que ao final do caminho nos aponta direções a seguir.

Uma das análises possíveis durante o processo KDD é o uso de árvores de decisão, que são diagramas que representam as várias alternativas que podem resultar de uma classe, com as probabilidades dessa alternativa ocorrer. As árvores de decisão são usualmente aplicadas a grandes bancos de dados, para a descoberta de regras expressas em linguagem natural, utilizando conhecimentos de inteligência artificial e estatística, de acordo com [Basgalupp 2010].

Ainda segundo [Basgalupp 2010], o método de Info Gain usa a entropia como medida de impureza, ou seja, para determinar quanto uma condição de teste é boa, compara-se o valor da entropia do nó-pai com o grau de entropia dos nós-filhos, sendo que a maior diferença é a selecionada para condição teste; porém esse método tem como problematizador a preferência que dá a atributos com muitos valores possíveis. Para solucionar o problematizador do método *Info Gain*, definiu-se o *Gain Ratio*, que utiliza um ganho relativo como critério de avaliação, e com isso traz melhor acurácia nas árvores de decisão geradas.

Após a seleção das variáveis de estudo, é possível utilizar uma das técnicas de associação disponível. Em nosso estudo foi usado o algoritmo Apriori, que realiza duas etapas de estudo, a geração e poda, de acordo com [Destefani, Motyczka, Sausen et al 2017]. Inicialmente o algoritmo varre o banco de dados e cria conjuntos de regras entre os valores existentes, em seguida seleciona os conjuntos de dados com uma menor frequência pré-fixada no valor mínimo, valor este fixado na ferramenta de Mineração de Dados (MD).

3. Trabalhos Relacionados

No trabalho de [Calixto, Segundo and Gusmão 2017] o objetivo foi identificar as variáveis relevantes para os índices de evasão escolar, tomando como base os dados do censo educacional de 2014, 2015 e 2016 dos estados do Ceará e Sergipe. Nesse estudo foi utilizada a metodologia CRoss-Industry Standard Process for Data Mining (CRISP-DM) para a parte inicial do processamento dos dados, e posteriormente as análises foram feitas com técnicas de Indução de Regras e Regressão Logística. Foram identificados

como variáveis influentes na evasão escolar a idade, etapa de ensino, modalidade de ensino, existência de laboratórios e localização da escola.

Em [Carrano, Albergaria, Infante et al 2019] diversas técnicas de mineração de dados são utilizadas para tentar identificar e prevenir a evasão estudantil nas universidades públicas. Os dados utilizados correspondem a dados pessoais, socioeconômicos e acadêmicos de alunos da Universidade Federal de São João del-Rei e identificou-se que os indicadores acadêmicos foram mais relevantes para a evasão escolar.

O trabalho de [Lima, Ambrósio, Ferreira et al 2019] faz uma análise sobre como os dados públicos referentes ao Enade e Enem têm sido utilizados ao longo do tempo. Neste estudo verificou-se que a maior parte das análises tem se limitado a utilização de estatística descritiva e focam geralmente nos dados socioeconômicos e notas dos alunos no exame. Este trabalho sugere um uso mais amplo dessa base de dados.

Em [Neves Junior, Nascimento, Fagundes et al 2019] a base de dados utilizada é do INEP, e apresenta um modelo preditivo para a aprovação e reprovação escolar no ensino médio, e aponta que regressão quantílica com otimização de parâmetros obteve menor erro na predição.

Tendo como ponto de partida estes trabalhos relacionados, vemos que muitas técnicas de mineração de dados são utilizadas para análise principalmente de contexto socioeconômico ou predição de evasão escolar, mas não encontramos nenhum estudo sobre a percepção do aluno com sua formação após a conclusão de seu curso de graduação. Por isso acreditamos que tentar entender os fatores relevantes para que um aluno se sinta melhor formado, como cidadão e profissional, pode ajudar tanto na gestão escolar, para prevenção da evasão, quanto pode ser um roteiro para que o estudante selecione uma instituição de ensino adequada a sua necessidade, e assim ingresse em um curso que irá formá-lo de forma mais completa.

Também somos motivados pela percepção que a ampla base de dados obtidos no ENADE possa ser melhor aproveitada para este contexto, assim pretende-se estimular o início de uma melhor distribuição de estudos de mineração de dados pelas diversas questões da prova.

4. Metodologia

Para que a base de dados do ENADE 2018 pudesse se transformar em conhecimento e com isso auxiliar na tomada de decisões de gestores da educação e alunos à procura de um curso superior que os traga maior satisfação pessoal e profissional, foram necessários alguns passos, que serão descritos a partir de agora, conforme o método KDD.

4.1. Seleção de dados

Inicialmente, selecionamos a base de dados (ENADE 2018) por ser a mais atual disponível no Portal de Microdados do Inep. Essa base de dados é composta por 136 questões distribuídas conforme tabela 4.1.1:

Tabela 4.1.1. Questões ENADE 2018

Parte 1 - informações da instituição de ensino superior e do curso	9 questões
Parte 2 - informações do estudante	7 questões
Parte 3 - número de itens da parte objetiva	8 questões
Parte 4 – vetores de notas	8 questões
Parte 5 - tipos de presença	6 questões
Parte 6 - tipos de situação das questões da parte discursiva	5 questões
Parte 7 - notas na formação geral e componente específico	16 questões
Parte 8 - questionário de percepção da prova	9 questões
Parte 9 - questionário do estudante	26 questões
Parte 10 – questões relacionadas as condições acadêmicas	42 questões
Total	136 questões

Uma visão inicial dos dados mostrou que 548.127 alunos participaram do exame, sendo estudantes de todo o Brasil, de diversos cursos, e que havia diversos tipos de participação na prova (ausentes, presentes, com resultados válidos ou não, etc.), dessa forma, antes de prosseguir com o estudo, optou-se por selecionar o conjunto a ser usado da seguinte forma:

- Diminuir o conjunto de dados a estudantes apenas do Estado de São Paulo, uma vez que este estado apresenta a maior concentração de graduandos;
- Utilizar dados de graduandos de cursos técnicos, pois estes cursos agrupam diversos cursos distintos, e essa variedade de cursos pode auxiliar nas associações buscadas no estudo;
- Trabalhar apenas com cursos feitos de forma presencial, pois a estrutura acadêmica oferecida em cursos à distância (EaD) é diferente, e considerando a maior oferta de cursos presenciais, essa modalidade apresenta maior quantidade de dados;
- Utilizar apenas dados de alunos presentes com presença válida na prova.

Na Tabela 4.1.2., apresentamos um resumo dos dados na base de dados do ENADE, e as quantidades restantes após as filtragens efetuadas.

Tabela 4.1.2. Resumo de dados

	Total de Participantes na Prova	Percentual
ENADE	548.127	100%
São Paulo	141.641	25,84% do ENADE
Graduandos em cursos presenciais	117.310	82,82% do total de SP
Graduandos em cursos técnicos presenciais	24.537	20,92% do total de graduandos em SP
Com Presença Válida na prova	18.783	76,55% do total de graduandos em cursos técnicos presenciais em SP

4.2. Pré-processamento e transformação dos dados

Os dados foram filtrados e tratados no software Excel (editor de planilhas produzido pela Microsoft). O tratamento incluiu a remoção de caracteres especiais, para utilização no programa Waikato Environment for Knowledge Analysis (WEKA).

O WEKA foi desenvolvido na Universidade de Waikato, Nova Zelândia, e é composto de uma série de algoritmos para solucionar problemas de Mineração de Dados. É um programa muito utilizado por tratar-se de programa de domínio público, cuja interface é amigável e simples para o usuário final, segundo [Destefani, Motyczka, Sausen et al 2017].

Após a filtragem inicial, verificou-se que ainda restavam campos cuja resposta do estudante era nula, assim todos os dados de estudantes com ao menos uma resposta "nula" foram descartados do estudo.

Encontramos ainda respostas do tipo "Não se aplica" e "Não sei responder" que foram descartados no processo de limpeza de dados.

Após a etapa acima descrita, restaram 7.168 graduandos com todas as respostas válidas para análise.

A partir da base de dados restante, o campo da "Idade" dos estudantes foi agrupado em faixas etárias.

Após todo esse pré-processamento e transformação, para tentar identificar através da mineração de dados os itens que seriam mais importantes para o estudante sentir-se melhor formado, como cidadão e profissional, definiu-se, dentre as respostas dos estudantes às questões relacionadas as condições acadêmicas durante a graduação, os campos a analisar, excluindo-se perguntas cuja resposta esperada seria não mensurável (por exemplo, melhoria da capacidade de reflexão, de pensamento, etc).

Os dados socioeconômicos também foram excluídos, pois como apresentado na Seção 3, existem outros estudos que se focam nesta área do banco de dados.

As informações sobre o tipo de organização acadêmica, nota da prova e sexo foram mantidos, para verificar se são relevantes no conjunto dos dados, juntamente com as condições acadêmicas.

Assim, os campos utilizados neste estudo, juntamente com a parte do ENADE ao qual se referem, são apresentados na Tabela 4.2.1.

Tabela 4.2.1. Questões utilizadas neste estudo

Parte 1	Q.1. Organização acadêmica da IES
	Q.2. Área de enquadramento do curso no Enade
Parte 2	Q.3. Faixa etária
	Q.4. Sexo
Parte 7	Q.5. Nota bruta da prova
Parte 9	Q.6. Que tipo de bolsa de estudos ou financiamento do curso
	Q.7. Você recebeu algum tipo de bolsa de permanência
	Q.8. Você recebeu algum tipo de bolsa acadêmica
	Q.9. Você participou de programas e ou atividades curriculares no exterior
Parte 10	Q.10. As disciplinas cursadas contribuíram para sua formação integral como cidadão e profissional
	Q.11. Os conteúdos abordados favoreceram atuação em estágios ou atividades de iniciação profissional
	Q.12. O curso propiciou experiências de aprendizagem inovadoras
	Q.13. As relações professor-aluno estimularam você a estudar e aprender
	Q.14. Os planos de ensino contribuíram para as atividades acadêmicas e para seus estudos
	Q.15. Houve oportunidades para superar dificuldades relacionados ao processo de formação
	Q.16. A coordenação do curso esteve disponível para orientação acadêmica
	Q.17. Houve oportunidades para participar de programas, projetos ou atividades de extensão universitária
	Q.18. Houve oportunidades para participar de projetos de iniciação científica ou atividades de investigação acadêmica
	Q.19. Ofereceu condições para participar de eventos internos e externos a instituição
	Q.20. Ofereceu oportunidades para atuar como representantes em órgãos colegiados
	Q.21. O curso favoreceu a articulação do conhecimento teórico com atividades práticas
	Q.22. As atividades práticas foram suficientes para relacionar os conteúdos com a prática
	Q.23. O curso propiciou acesso a conhecimentos atualizados e contemporâneos em sua área de formação
	Q.24. O estágio supervisionado proporcionou experiências diversificadas
	Q.25. As atividades durante trabalho de conclusão contribuíram para qualificar sua formação profissional
	Q.26. Houve oportunidades para os estudantes realizarem intercâmbios ou estágios no país
	Q.27. Houve oportunidades para os estudantes realizarem intercâmbios ou estágios fora do país
	Q.28. Os estudantes participaram de avaliações periódicas do curso
	Q.29. As avaliações de aprendizagem realizadas foram compatíveis com os conteúdos ou temas trabalhados
	Q.30. Os professores apresentaram disponibilidade para atender fora do horário das aulas
	Q.31. Os professores demonstraram domínio dos conteúdos abordados
	Q.32. Os professores utilizaram tecnologias da informação e comunicação (TICs) como estratégia de ensino
	Q.33. A instituição dispôs de quantidade suficiente de funcionários para o apoio administrativo e acadêmico
	Q.34. O curso disponibilizou monitores ou tutores para auxiliar os estudantes
	Q.35. As condições de infraestrutura das salas de aula foram adequadas

	Q.36. Os equipamentos e materiais disponíveis para as aulas práticas foram adequados para a quantidade de estudantes
	Q.37. Os ambientes e equipamentos destinados as aulas práticas foram adequados ao curso
	Q.38. A biblioteca dispôs das referências bibliográficas que os estudantes necessitaram
	Q.39. A instituição contou com biblioteca virtual ou conferiu acesso a obras disponíveis em acervos virtuais

Para que as respostas fossem submetidas ao processo de Mineração de Dados, os dados obtidos após todo o processamento acima detalhado foi salvo em planilha de dados separados por vírgulas, e posteriormente esse arquivo foi aberto no programa Bloco de Notas (editor de texto simples existente no Microsoft Windows) e ao fim, salvo com extensão .arff, para carregamento no programa WEKA.

A Q.9 e todas as questões da parte 10 tinham 6 opções de resposta possíveis, sendo 3 níveis para “Concordo” (Concordo Parcialmente, Concordo e Concordo Totalmente) e mais 3 para “Discordo” (Discordo Parcialmente, Discordo e Discordo Totalmente), como se pode verificar na Figura. 4.2.1., contendo a distribuição de respostas na tela do programa WEKA, com base na Q.10.

A Q.10 foi utilizada como base para as outras respostas por ser a questão que desejamos encontrar as associações, ou seja, é a questão onde os alunos responderam sobre sua formação integral como cidadão e profissional.

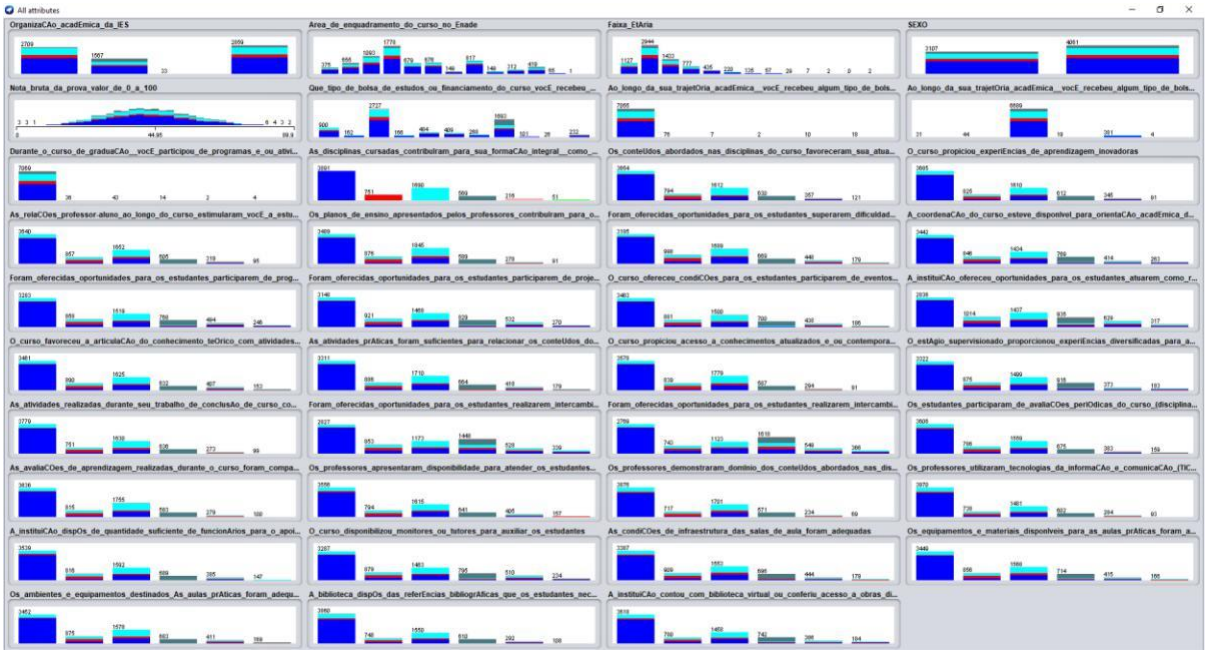


Figura. 4.2.1. Tela do WEKA – Distribuição das 6 respostas possíveis, com base na questão 10

A partir da visualização da distribuição de respostas acima, acreditamos que por não haver grande distribuição de dados entre as 6 respostas possível (3 níveis para “Concordo” e 3 para “Discordo”) o agrupamento destas respostas poderia auxiliar a associação de dados posterior.

E ainda para termos condições de analisar comparativamente os resultados obtidos, definimos uma segunda etapa de Mineração de Dados, desta vez com o agrupamento de respostas positivas (variações de “concordo”) e respostas negativas (variações de “discordo”) às questões nº 9 a nº 39.

Após o agrupamento feito no software Excel (editor de planilhas produzido pela Microsoft), foi feita nova planilha de dados separados por vírgulas, salva posteriormente com extensão .arff, para carregamento no programa WEKA.

Neste segundo carregamento, obtivemos a distribuição de respostas constante na Figura. 4.2.2., também levando em consideração a questão 10.

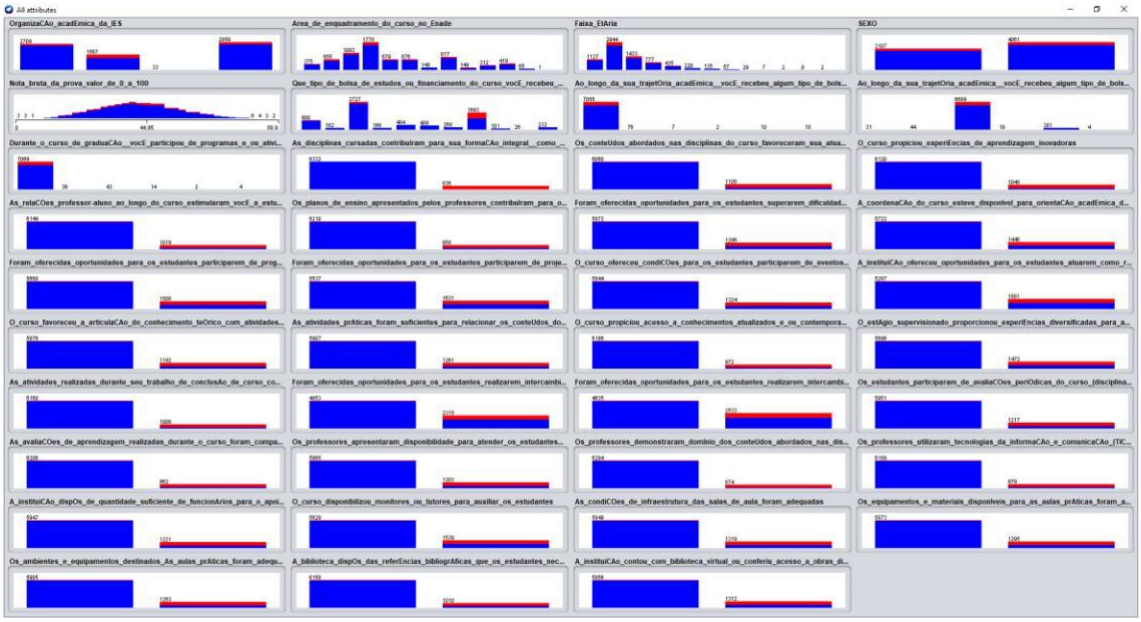


Figura. 4.2.2. Tela do WEKA – Distribuição das 2 respostas possíveis, com base na questão 10

A intenção é verificar se o seleção de variáveis gerada pelo programa WEKA será alterada de acordo com a quantidade de respostas possíveis, e com isso encontrar maior número de elementos relevantes para a formação dos graduandos.

Duas possibilidades existentes no programa Weka para seleção de atributos relevantes são os métodos de Info Gain e Gain Ratio. Segundo [Lopes 2013], o método Info Gain mede a proporção de ganho de informação de cada atributo, considerando a classe, enquanto o método Gain Ratio utiliza uma técnica chamada de informação de divisão e assim os atributos com maior ganho de informação serão selecionados, enquanto aqueles não relevantes são descartados.

Optamos por utilizar o *Gain Ratio*, pois o *Info Gain* não diferencia atributos com diferentes números de categorias, como é nosso conjunto de dados. No método utilizado, o atributo com maior ganho de informação será selecionado, enquanto os atributos considerados sem ganho (ranqueados como zero) são descartados. Dessa forma, o ganho de informação é considerado um selecionador, pois aqueles que não agregam informação são descartados durante o processo.

Assim, após o carregamento da planilha de dados no Weka, selecionamos o “*GainRatioAttributeEval*”, e no método de seleção de atributos, acatamos a colocação do

programa sobre o tipo “Ranker” necessário para o método. Como classe a ser atingida, selecionamos a “Q.10 - As disciplinas cursadas contribuíram para sua formação integral como cidadão e profissional”.

Esse processo foi feito para as duas planilhas de dados (com e sem as respostas agrupadas).

Após a primeira verificação, repetimos o processo excluindo da base de dados as questões com menor ranqueamento. Verificamos que os índices permaneceram inalterados, assim deixamos de analisar novas diminuições de questões.

Verificamos que as nove variáveis melhor ranqueadas são as mesmas para as duas situações, apesar da colocação dos itens não ser o mesmo. Em cada situação, as respostas passaram a divergir a partir da 10ª colocada.

A Tabela 4.2.2. demonstra o ranqueamento em cada situação, com os índices obtidos no programa WEKA, para as 10 questões mais relevantes em cada situação.

Tabela 4.2.2. Dez Itens melhor ranqueados

Questão	Ranqueamento dos dados considerando todas as variações de "Concordo" e "Discordo"	Ranqueamento dos dados com o agrupamento das variações de "Concordo" e "Discordo"
Q.11	0,3767961 (1º)	0,45634728 (8º)
Q.12	0,3536805 (6º)	0,46406417 (5º)
Q.13	0,332569 (8º)	0,45943955 (7º)
Q.14	0,3628084 (3º)	0,50125385 (2º)
Q.21	0,3182464 (10º)	Não se aplica
Q.23	0,366885 (2º)	0,49883305 (3º)
Q.25	0,3588674 (5º)	0,46380817 (6º)
Q.29	0,3592811 (4º)	0,48945283 (4º)
Q.31	0,3507245 (7º)	0,51860239 (1º)
Q.32	0,3247102 (9º)	0,44877137 (9º)
Q.38	Não se aplica	0,39369937 (10º)

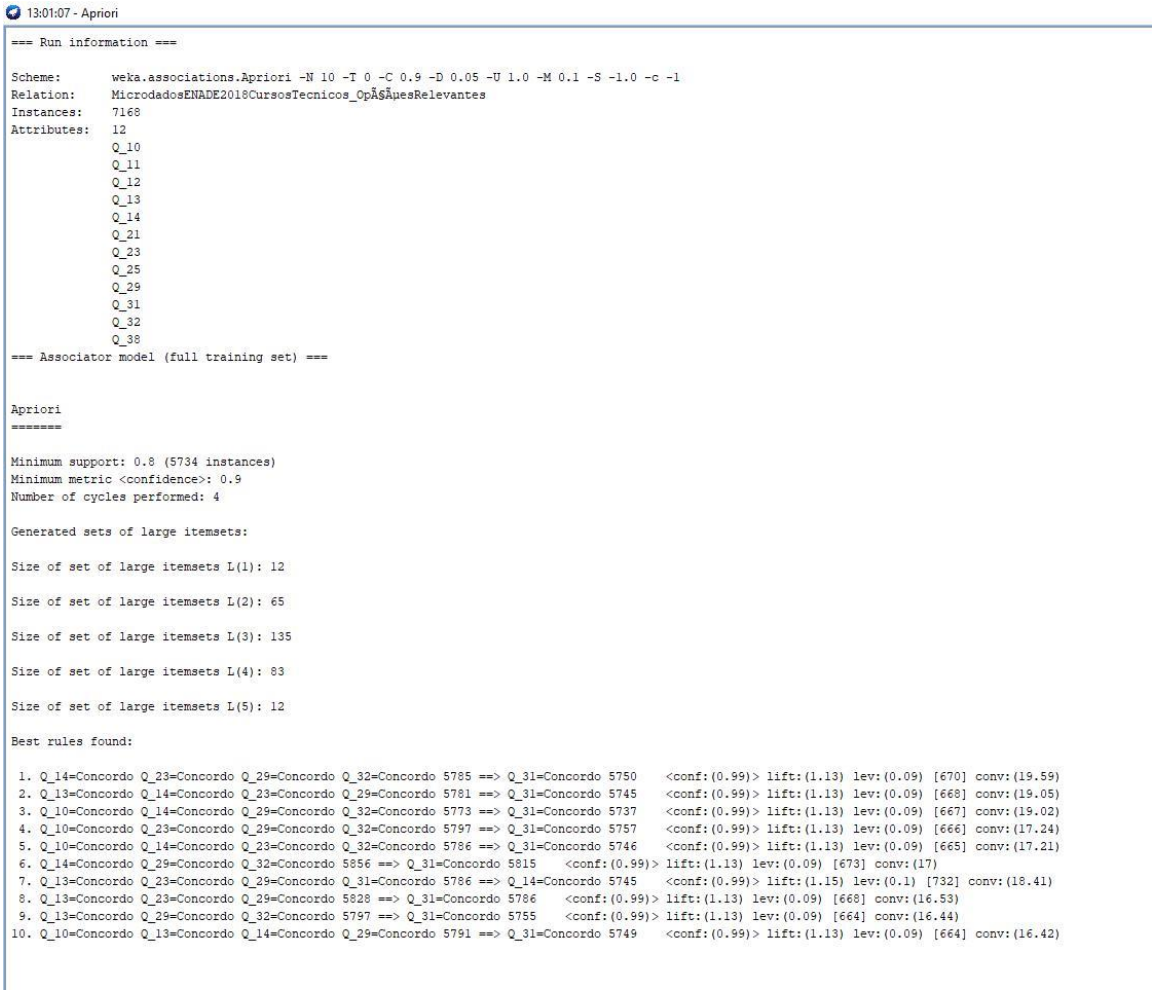
5. Avaliação

Após o ranqueamento dos critérios, definimos manter as 9 questões melhor ranqueadas (considerando os dois métodos utilizados), e também a resposta ranqueada na 10ª posição, em cada método. Acrescentamos ainda a questão 10, sobre a formação pessoal e profissional dos graduandos para submissão as regras de associação do algoritmo “Apriori” do programa WEKA.

Nesse momento, demos preferência as respostas agrupadas (Concordo e Discordo), para facilitar entendimento das regras geradas.

Os parâmetros usados para geração de regras de associação foram N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1.

As regras geradas podem ser verificadas na Figura 5.1.



```
13:01:07 - Apriori
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     MicrodadosENADE2018CursosTecnicos_OpÃ§Ã£esRelevantes
Instances:    7168
Attributes:   12
              Q_10
              Q_11
              Q_12
              Q_13
              Q_14
              Q_21
              Q_23
              Q_25
              Q_29
              Q_31
              Q_32
              Q_38
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.8 (5734 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 4

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 65

Size of set of large itemsets L(3): 135

Size of set of large itemsets L(4): 83

Size of set of large itemsets L(5): 12

Best rules found:

 1. Q_14=Concordo Q_23=Concordo Q_29=Concordo Q_32=Concordo 5785 ==> Q_31=Concordo 5750    <conf:(0.99)> lift:(1.13) lev:(0.09) [670] conv:(19.59)
 2. Q_13=Concordo Q_14=Concordo Q_23=Concordo Q_29=Concordo 5781 ==> Q_31=Concordo 5745    <conf:(0.99)> lift:(1.13) lev:(0.09) [668] conv:(19.05)
 3. Q_10=Concordo Q_14=Concordo Q_29=Concordo Q_32=Concordo 5773 ==> Q_31=Concordo 5737    <conf:(0.99)> lift:(1.13) lev:(0.09) [667] conv:(19.02)
 4. Q_10=Concordo Q_23=Concordo Q_29=Concordo Q_32=Concordo 5797 ==> Q_31=Concordo 5757    <conf:(0.99)> lift:(1.13) lev:(0.09) [666] conv:(17.24)
 5. Q_10=Concordo Q_14=Concordo Q_23=Concordo Q_32=Concordo 5786 ==> Q_31=Concordo 5746    <conf:(0.99)> lift:(1.13) lev:(0.09) [665] conv:(17.21)
 6. Q_14=Concordo Q_29=Concordo Q_32=Concordo 5856 ==> Q_31=Concordo 5815    <conf:(0.99)> lift:(1.13) lev:(0.09) [673] conv:(17)
 7. Q_13=Concordo Q_23=Concordo Q_29=Concordo Q_31=Concordo 5786 ==> Q_14=Concordo 5745    <conf:(0.99)> lift:(1.15) lev:(0.1) [732] conv:(18.41)
 8. Q_13=Concordo Q_23=Concordo Q_29=Concordo 5828 ==> Q_31=Concordo 5786    <conf:(0.99)> lift:(1.13) lev:(0.09) [668] conv:(16.53)
 9. Q_13=Concordo Q_29=Concordo Q_32=Concordo 5797 ==> Q_31=Concordo 5755    <conf:(0.99)> lift:(1.13) lev:(0.09) [664] conv:(16.44)
10. Q_10=Concordo Q_13=Concordo Q_14=Concordo Q_29=Concordo 5791 ==> Q_31=Concordo 5749    <conf:(0.99)> lift:(1.13) lev:(0.09) [664] conv:(16.42)
```

Figura 5.1. Saída do Programa WEKA, para Associação Apriori

A expectativa era encontrar associações do tipo SE (XXXXX) ENTÃO (Q.10), mas essa regra não foi gerada. Assim, como nosso objeto de estudo está ligado à questão 10, analisaremos apenas as regras que tem essa questão em seu conjunto, conforme dados da Tabela 5.2.

Tabela 5.1. Regras ligadas à Satisfação com a Formação do Graduando

Número da Regra	Regra	Conf	Lift	Conv
3	Q_10=Concordo Q_14=Concordo Q_29=Concordo Q_32=Concordo 5773 ==> Q_31=Concordo 5737	0.99	1.13	19.02
4	Q_10=Concordo Q_23=Concordo Q_29=Concordo Q_32=Concordo 5797 ==> Q_31=Concordo 5757	0.99	1.13	17.24
5	Q_10=Concordo Q_14=Concordo Q_23=Concordo Q_32=Concordo 5786 ==> Q_31=Concordo 5746	0.99	1.13	17.21

10	Q_10=Concordo Q_13=Concordo Q_14=Concordo Q_29=Concordo 5791 ==> Q_31=Concordo 5749	0.99	1.13	16.42

Verifica-se que as questões que mais aparecem nas relações são a “Q.14. Os planos de ensino contribuíram para as atividades acadêmicas e para seus estudos” (em 3 de 4 regras), “Q. 29. As avaliações de aprendizagem realizadas foram compatíveis com os conteúdos ou temas trabalhados” (em 3 de 4 regras), “Q.31. Os professores demonstraram domínio dos conteúdos abordados” (em 4 de 4 regras) e “Q.32. Os professores utilizaram tecnologias da informação e comunicação (TICs) como estratégia de ensino” (em 3 de 4 regras), assim, os resultados obtidos para quais itens tem maior relevância para os graduandos parecem estar convergindo para as mesmas questões.

6. Discussão

Dos itens melhor ranqueados nos dois métodos utilizados de seleção, verificamos que itens acadêmicos estão entre os mais relacionados à percepção de melhor formação integral do aluno. Apesar do ranqueamento com o agrupamento de respostas e sem agrupamento tenha itens em colocação diferente, vimos que os 9 itens melhor ranqueados são os mesmos.

Dentre todos os itens melhor ranqueados, aqueles relacionados a atuação do professor tiveram maior associação a percepção de formação integral dos graduandos. Isso aponta para o importante papel do professor, tanto com seu conhecimento quanto na preparação e apresentação das aulas, com conteúdo sempre atualizado e voltado para a atuação profissional, bem como o uso de TIC´s como estratégia de ensino.

Essas descobertas parecem estar de acordo com o estudo de [Carrano, Albergaria, Infante et al 2019], onde identificou-se que os indicadores acadêmicos são relevantes para a evasão escolar, assim, em sentido inverso, bons indicadores devem evitar a evasão e tendem a tornar o aluno mais satisfeito com sua formação pessoal e profissional.

E finalmente, pode-se associar a formação e conhecimento do professor com a satisfação dos alunos, assim, um importante item de investimento das instituições de ensino deve ser na contratação, retenção e formação contínua constante dos professores.

Uma questão que pode ser colocada como ameaça ao estudo é que o conjunto de questões do ENADE utilizadas não costuma ser objeto de estudo, por isso ainda não há validação se os estudantes realmente se empenham na resposta a estes itens ou apenas colocam uma alternativa sem prestar a devida atenção. Acreditamos que caso essa parte do questionário seja mais utilizada no futuro, os alunos vejam a importância na atenção a este trecho da prova.

Uma limitação possível a este trabalho foi o descarte de diversas questões existentes no conjunto de dados inicial, talvez a inclusão de mais dados possa alterar a ordenação dos itens de relevância na formação do aluno. Além disso, houve descarte de respostas nulas, que também podem influenciar na ordenação de itens.

Mais um fator de limitação foi a opção quanto ao não uso de itens socioeconômicos, talvez a inclusão destes dados pode levar a associações diversas.

7. Conclusão

Os dados públicos sobre a educação são bastante estudados, sob diversas técnicas de mineração de dados, mas normalmente se buscam respostas sobre como evitar evasão escolar ou sobre critérios socioeconômicos dos alunos.

Acreditamos que um fator importante a ser estudado é o nível de satisfação dos graduandos quanto a formação recebida no curso feito, se o estudante acredita que o curso o preparou bem para a atuação profissional e sua vida pessoal. Estas respostas podem ser buscadas através do estudo dos dados do ENADE.

Foi essa a intenção deste estudo, apresentar uma relação entre a percepção da satisfação do aluno formado, em associação com itens disponibilizados na instituição de ensino durante o curso.

Esses dados podem ser utilizados pelos gestores da educação, na intenção de melhorar suas instituições e com isso atrair mais estudantes. Na mesma linha de uso, os gestores podem definir opções de investimento mais adequadas, como por exemplo na aquisição e retenção de talentos, uma vez que a atuação do professor foi associada fortemente à satisfação dos alunos.

Os estudantes podem se basear nos dados obtidos este estudo para comparar as condições de várias instituições de ensino, e utilizar estes dados como um auxiliar na escolha de seu futuro local de estudo.

Assim, a percepção geral após este trabalho é que a qualidade dos professores é um grande indicativo de satisfação dos alunos formados nas instituições de ensino, pois um bom professor consegue produzir aulas com conteúdo mais atualizado, mais próximas a necessidade profissional dos jovens e com isso torna-los mais preparados para o futuro.

Futuros trabalhos podem se debruçar sobre mais questões respondidas pelos alunos que fazem o Enem, para apurar se a inclusão de outros itens altera a seleção dos dados, ou ainda aplicar o método utilizado em outras áreas do conhecimento ou mesmo em estados diferentes da federação.

Referências

Arruda, D. Z. M (2019). Evasão escolar no ensino técnico: um estudo de caso numa escola técnica do Centro Paula Souza. Tese (mestrado).

Basgalupp, P. (2010). LEGAL - Tree: Um algoritmo genético multi-objetivo lexicográfico para indução de árvores de decisão. Tese (doutorado).

Calixto, K., Segundo, C. and Gusmão, R. P. De (2017). Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. *Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017)*, v. 1, n. Cbie, p. 1447.

Campos Neto, C. de M. (2016). Análise inteligente de dados em um banco de dados de procedimentos em cardiologia intervencionista. USP/IDPC/Biblioteca/64/16.

Carrano, D., Albergaria, E. T. De, Infante, C. and Rocha, L. (2019). Combinando Técnicas de Mineração de Dados para Melhorar a Detecção de Indicadores de Evasão Universitária. n. Cbie, p. 1321.

Destefani, L. A., Motyczka, L. B., Sausen, P. and Sausen, A. (2017). Busca De Padrões Utilizando O Algoritmo Apriori Para Mineração De Dados Nas Subestações Subterrâneas Da Ceee . 1 Pattern Search Using the Apriori Algorithm for Data Mining in Ceee Underground Substations. n. Md.

Elias, B., Isotani, S. and Penteado, B. (2017). Dados abertos educacionais: que informações temos disponíveis?.

Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (1996). Advances in Knowledge Discovery & Data Mining. 1 ed. American Association for Artificial Intelligence, Menlo Park, Califórnia, 611 folhas

Inep, Portal de Microdados, disponível em: http://inep.gov.br/microdados

Lima, P. da S. N., Ambrósio, A. P. L., Ferreira, D. J. and Brancher, J. D. (2019). Análise de dados do Enade e Enem: uma revisão sistemática da literatura TT - Data analysis of Enade and Enem: a systematic review of literature. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, v. 24, n. 1, p. 89–107.

Lopes, K. M. de O. (2013). Modelos Baseados em Data Mining para Classificação Multitemporal de Culturas no Mato Grosso Utilizando Dados de NDVI/MODIS. p. 126.

Lourenço, V., Formigoni1, M., Icaro, ;, et al. (2018). Artigo Original Open Access Mineração De Dados Na Base De Dados Aberta Da Câmara Legislativa Federal Brasileira: Ênfase Na Análise Dos Dados Da Legislatura 54 (2011-2013) Data Mining in the Open Database of the Brazilian Federal Legislative Chamber: Emphasis in the Analysis of Data in Legislature 54 (2011-2013). Brazilian Journal of Production Engineering, v. 4, p. 156–170.

Neves Junior, R., Nascimento, R. L. S. Do, Fagundes, R. A. de A. and Mattos Neto, P. S. G. De (2019). Estimação de Índices de Aprovação e Reprovação Escolar do Ensino Médio. n. Cbie, p. 339.

Rezende, S. (2003). Sistemas Inteligentes: fundamentos e aplicações. 1ª Ed. Barueri. São Paulo. Ed Manole Ltda.

Silva, L. G. F., Rocha, M. E. and Fagundes, R. A. (2017). ENADE: Math and Science Students Performance Analysis. *IEEE Latin America Transactions*, v. 15, n. 9, p. 1742–1746.

WEKA, disponível para *download* em https://waikato.github.io/weka-wiki/downloading_weka/.