



CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS

Recredenciado pela Portaria Ministerial nº 1.162, de 13/10/16, D.O.U. nº 198, de 14/10/2016
AELBRA EDUCAÇÃO SUPERIOR - GRADUAÇÃO E PÓS-GRADUAÇÃO S.A.

Emanoel Mendes Magalhães

FERRAMENTA PARA TRATAMENTO E VISUALIZAÇÃO DOS DADOS DO ENADE SOBRE
O QUESTIONÁRIO DO ESTUDANTE

Palmas – TO

2021

Emanoel Mendes Magalhães
FERRAMENTA PARA TRATAMENTO E VISUALIZAÇÃO DOS DADOS DO ENADE SOBRE
O QUESTIONÁRIO DO ESTUDANTE

Projeto de Pesquisa elaborado e apresentado como requisito parcial para aprovação na disciplina de Trabalho de Conclusão de Curso II (TCC II) do curso de bacharel em Ciência da Computação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. M.e Fabiano Fagundes.

Palmas – TO

2021

Emanoel Mendes Magalhães
FERRAMENTA PARA TRATAMENTO E VISUALIZAÇÃO DOS DADOS DO ENADE SOBRE
O QUESTIONÁRIO DO ESTUDANTE.

Projeto de Pesquisa elaborado e apresentado como requisito parcial para aprovação na disciplina de Trabalho de Conclusão de Curso II (TCC II) do curso de bacharel em Ciência da Computação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. M.e Fabiano Fagundes.

Aprovado em: ____/____/____

BANCA EXAMINADORA

Prof. M.e Fabiano Fagundes

Orientador

Centro Universitário Luterano de Palmas – CEULP

Prof. M.e Heloise Acco Tives

Instituto Federal do Paraná - Campus Palmas - IFPR

Prof. Bela. Fernanda Pereira Gomes

Centro Universitário Luterano de Palmas – CEULP

Palmas – TO

2021

RESUMO

MAGALHÃES, Emanuel Mendes. Desenvolver uma ferramenta para tratamento e visualização dos dados do Enade sobre o questionário do estudante. 2021. 48f. Trabalho de Conclusão de Curso (Graduação) – Ciência da Computação, Centro Universitário Luterano de Palmas, Palmas/TO, 2021.

O Enade gera todos os anos um grande conjunto de dados abertos que permitem a realização de análises sobre o desempenho de estudantes e instituições de ensino superior brasileiras. A partir da necessidade de entendimento desses dados foi verificado o quão trabalhosa e complexa é a sua utilização, pois estes vêm em um conjunto detalhado de todas as informações que são adquiridas durante a realização do exame, sendo que muitas das vezes não será necessária a utilização de todos os dados para a análise que o usuário necessita. A partir desse contexto, foi idealizada uma ferramenta que irá receber essas informações realizando uma análise das informações obtidas. Neste trabalho foi demonstrado todos os passos realizados para atingir o objetivo proposto, onde as etapas utilizadas foram a coleta das informações, escolha das ferramentas, escrita do referencial, desenvolvimento, escrita dos resultados. Essa ferramenta realizou a análise apenas dos dados dos questionários dos estudantes que consistem em perguntas para formar o perfil dos graduandos. Foi realizado um tratamento desses dados para que possa ser disponibilizado de uma forma que torne o entendimento mais simples e torne este processo mais prático para o utilizador da ferramenta.

Palavras Chave: Dados, Tratamento, Visualização

LISTA DE FIGURAS

Figura 1- Microdados do Enade.	9
Figura 2 - Questionário dos estudantes.	10
Figura 3 - Mineração dos dados.	12
Figura 4 - Cross-validation	13
Figura 5 - Gráfico de linhas.	15
Figura 6 - Diagrama.	16
Figura 7 - Mapa de distribuição de pontos.	17
Figura 8 - Métodos do Trabalho.	20
Figura 9 - Etapas do Desenvolvimento	21
Figura 10 - Dados selecionados	24
Figura 11 - Pré-Processamento	25
Figura 12 - Transformação de string em número.	26
Figura 13 - Pós-Processamento.	26
Figura 14 - Títulos das colunas.	27
Figura 15 - Código de transformação.	27
Figura 16 - Resultado da transformação	28
Figura 17 - Banco de Dados	28
Figura 18 - Modelo de Dados	29
Figura 19 - Dados de treino	30
Figura 20 - Parâmetros	30
Figura 21 - Algoritmo KNN	31
Figura 22 - Erro empírico	31
Figura 23 - k vizinhos	32
Figura 24 - Variáveis finais	32
Figura 25 - Inserir dados	33
Figura 26 - Tratamento gráfico	34
Figura 27 - Gráfico	35
Figura 28 - Seleção das opções	36
Figura 29 - Gráfico de barra	37
Figura 30 - Gráfico de radar	38

Figura 31 - Exemplo Barras

39

Figura 32 - Exemplo Radar

40

LISTA DE ABREVIATURAS E SIGLAS

AIDS	Síndrome da Imunodeficiência Adquirida
Enade	Exame Nacional dos Estudantes
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KNN	K Nearest Neighbor
MD	Mineração de Dados

SUMÁRIO

1 INTRODUÇÃO	7
2 REFERENCIAL TEÓRICO	9
2.1 ENADE	9
2.2 Tratamento e Organização dos Dados	11
2.3 Visualização da Informação	14
3 METODOLOGIA	19
3.1 FERRAMENTAS	19
3.2 MÉTODOS	20
4 RESULTADOS E DISCUSSÃO	23
4.1 SELEÇÃO	23
4.2 PRÉ PROCESSAMENTO	25
4.3 TRANSFORMAÇÃO	27
4.4 BANCO DE DADOS	28
4.5 MINERAÇÃO	29
4.6 AVALIAÇÃO	31
4.7 VISUALIZAÇÃO DOS DADOS	33
5 CONSIDERAÇÕES FINAIS	41
REFERÊNCIAS	43

1 INTRODUÇÃO

“Atualmente, o avanço tecnológico é muito evidente e gera transformações na vida de todos que buscam estar antenados na evolução. A informática, com todos seus recursos, torna-se uma ferramenta de grande poder na formação do ser humano” (LADEIRA; COELHO, 2016, p.1). Esse avanço tecnológico auxilia inúmeras áreas, e a área da educação é uma delas.

No ano de 2018 foi divulgado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP 2020, p.1) que no exame do Enade naquele ano “estavam inscritos 550.842 estudantes concluintes de bacharelados e licenciaturas [...] Desses, 461.845 (83,8%) participaram das provas”. Pode-se assim perceber que o Exame Nacional dos Estudantes (Enade) produz todos os anos um grande número de dados que o possibilita avaliar os alunos graduandos de acordo com o conhecimento adquirido nos estudos além de também permitir a avaliação das instituições.

Neste trabalho serão tratados e analisados dados referentes ao questionário dos estudantes que segundo o (INEP, 2019, p.1) “o questionário tem por objetivo levantar informações que permitam caracterizar o perfil dos estudantes e o contexto de seus processos formativos, relevantes para a compreensão dos resultados do concluintes no Enade” onde é disponibilizado para os alunos que realizam a prova. Este questionário é composto por 68 questões obrigatórias que estão associadas à vida pessoal dos estudantes e sobre o curso em que é concluinte. Estes dados são importantes para poder caracterizar o perfil dos estudantes e como foi o seu processo de formação.

A partir da obtenção dos dados do questionário dos estudantes serão utilizados métodos de mineração de dados que segundo (CÔRTEZ *et al.*, 2002) estão “se tornando cada vez mais populares como uma ferramenta de descoberta de informações, que podem revelar estruturas de conhecimento, que possam guiar decisões em condições de certeza limitadas”, pois com a mineração será possível extrair os dados do questionário dos estudantes disponibilizados pelo INEP.

Este trabalho apresenta o desenvolvimento de uma ferramenta que realiza o tratamento dos dados do questionário dos estudantes referente ao ano de 2017. Durante todo o trabalho será feita referência aos dados de 2017 mas a plataforma permite, a partir de uma seleção manual das informações, aceitar outros anos, assim oferecendo uma forma de visualização que auxilia no entendimento e compreensão

dos dados pelo usuário. Por fim, esses dados são representados de modo que possam auxiliar o entendimento de quem busca essas informações.

Com base nestas informações foi observado que é preciso verificar como implementar uma ferramenta para auxiliar na análise dos dados do questionário dos estudantes do Enade e a partir das técnicas de tratamento de dados e de visualizações será possível desenvolver uma ferramenta que permitirá a análise dos dados do questionário dos estudantes do ENADE.

Assim será possível desenvolver uma ferramenta que realizará o tratamento dos dados do Enade disponibilizando de forma visualmente melhor as informações que os usuários necessitam. Os objetivos que são buscados a alcançar são estudar os métodos para tratamento dos dados, estudar técnicas de visualização da informação, desenvolver uma ferramenta de tratamento dos dados para facilitar a visualização de dados do questionário dos estudantes e desenvolver uma interface para a visualização dos dados tratados e informações.

O Enade é uma avaliação que testa os conhecimentos adquiridos dos concluintes de graduação de acordo com os conteúdos previstos nas diretrizes curriculares dos cursos. Os estudantes também preenchem um questionário que contém perguntas relacionadas a vida do graduando e sobre a instituição de ensino. A partir desta avaliação disponibilizada pelo Inep são liberados dados que possibilitam fazer estudo sobre como está sendo o desempenho dos estudantes e das instituições de ensino. Esses dados são liberados pelo INEP em um conjunto de informações referentes ao questionário dos estudantes, respostas das provas e entre outras informações, assim sendo necessário um maior entendimento e familiaridade com as informações geradas para que seja possível obter o conhecimento desejado.

Existem técnicas computacionais que possibilitam a extração dos dados e tratamento destes para que seja possível visualizar de forma clara e visualmente coerente conjuntos grandes de informações. Como os dados do Inep são um conjunto de diversas informações do exame do Enade, torna-se interessante oferecer uma ferramenta que possa ajudar a todos na análise destes dados e principalmente auxiliar pessoas leigas em informática ou não muito familiarizadas com os dados do Enade para que consigam compreender e tirar conclusões a partir destes dados.

2 REFERENCIAL TEÓRICO

Nesta seção são abordados alguns conceitos importantes para entendimento do trabalho proposto: Enade e seus conceitos, tratamento e organização de dados e visualização das informações.

2.1 ENADE

O Exame Nacional de Desempenho dos Estudantes (Enade) foi criado no ano de 2004 para avaliar o rendimento dos concluintes dos cursos de graduação, a fim de verificar se os conteúdos que constam nas diretrizes dos cursos realmente estão sendo aplicados corretamente e para medir a qualidade das instituições de ensino. A prova é composta por duas partes, sendo a primeira parte de formação geral que é comum aos cursos de todas as áreas contendo 10 questões, e a segunda parte é composta por 30 questões de conhecimentos específicos para cada curso.

Todos os anos o Enade gera microdados que, segundo o QEDU (2020, p.1) “representam a menor fração de um dado e pode estar relacionado a uma pesquisa ou avaliação. A partir da agregação de microdados é construída a informação”. Pode-se ver na Figura 1 a seguir uma demonstração de como são organizados os microdados disponibilizados pelo INEP gerados a partir da realização do Enade.

Figura 1- Microdados do Enade.

NU_ANO	CO_IES	CO_CATECO	CO_ORGA	CO_GRUP	CO_CURS	CO_MODA	CO_MUNI	CO_UF	CLCO_REGIA	NU_IDADE	TP_SEXO	ANO_FIM	ANO_IN	CCO_TURN	TP_INSCRIP	TP_INSCRIN	NU_ITEM	NU_ITEM	NU_ITEM	NU_IT
2018	1356	10003	10020	1	47116	1	3546603	35	3	22	M	2013	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	49	F	1988	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	23	M	2013	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	25	M	2011	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	22	F	2014	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	22	M	2014	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	22	M	2014	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	22	F	2014	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	25	M	2011	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	21	F	2014	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	25	F	2010	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	21	M	2014	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	23	M	2013	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	23	M	2013	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	24	M	2012	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	22	M	2014	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	27	M	2012	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	23	M	2012	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	22	M	2013	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	27	M	2008	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	23	M	2014	2015	4	0	0	8	0	0	0
2018	1356	10003	10020	1	47116	1	3546603	35	3	28	M	2008	2015	4	0	0	8	0	0	0

Como mostra a Figura 1 os microdados que o INEP disponibiliza são bem concisos e diretos para que o conjunto total possa ser o menor possível. Estes microdados armazenam todos os atributos oriundos do exame incluindo também o questionário respondido pelos estudantes que é o foco do trabalho.

O questionário dos estudantes é de caráter obrigatório e necessita que o concluinte preencha todo o questionário para que assim possa ser confirmada a sua inscrição no Enade e a partir do preenchimento é liberado o local de prova para o estudante. Como podemos ver a seguir na Figura 2 algumas questões que são apresentadas aos estudantes no questionário dos estudantes.

Figura 2 - Questionário dos estudantes.

1. Qual o seu estado civil?
A () Solteiro(a).
B () Casado(a).
C () Separado(a) judicialmente/divorciado(a).
D () Viúvo(a).
E () Outro.
2. Qual é a sua cor ou raça?
A () Branca.
B () Preta.
C () Amarela.
D () Parda.
E () Indígena.
F () Não quero declarar.
3. Qual a sua nacionalidade?
A () Brasileira.
B () Brasileira naturalizada.
C () Estrangeira.
4. Até que etapa de escolarização seu pai concluiu?
A () Nenhuma.
B () Ensino Fundamental: 1º ao 5º ano (1ª a 4ª série).
C () Ensino Fundamental: 6º ao 9º ano (5ª a 8ª série).
D () Ensino Médio.
E () Ensino Superior - Graduação.
F () Pós-graduação.

Fonte: Inep (2019, *online*)

Na Figura 2 são apresentados algumas das questões presentes no questionário dos estudantes que é composto por 68 questões divididas em perguntas relacionadas ao curso dos estudantes e sobre a vida pessoal como foi demonstrado na figura anteriormente mencionada.

2.2 TRATAMENTO E ORGANIZAÇÃO DOS DADOS

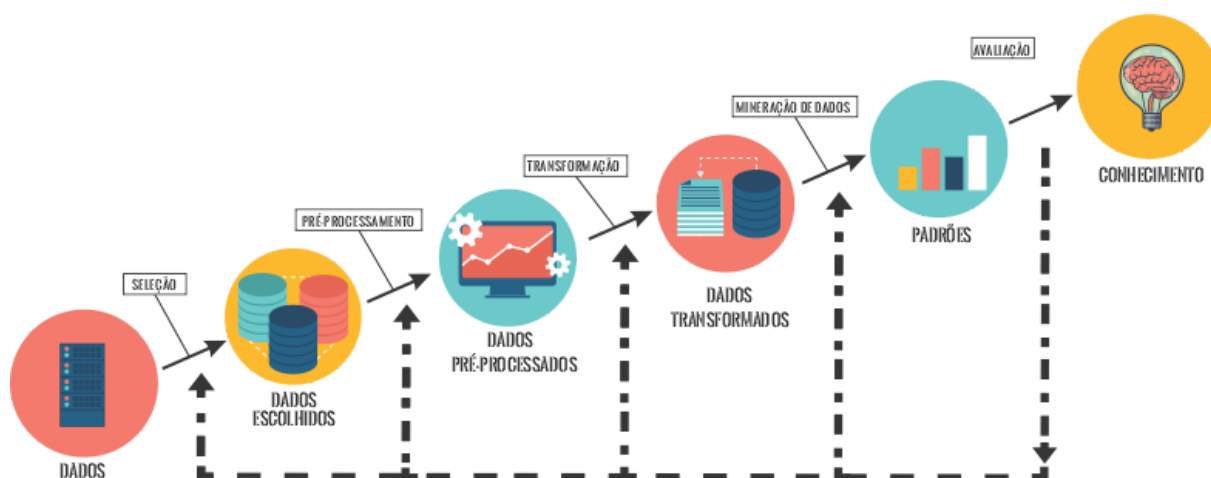
Muitas vezes para que seja possível analisar os dados dos quais o usuário deseja obter mais informações, primeiramente é necessário realizar nesses dados um tratamento para separar o que não tem utilidade, e só então analisar aquilo que esteja de acordo com o objetivo demonstrado pelo usuário.

No tratamento destaca-se a duplicação de registros, normalmente originada por negligência na introdução dos dados, pelo incorreto fornecimento dos mesmos ou por um erro de digitação. É também frequente o aparecimento de dados com valores omissos, para os quais é necessário definir uma estratégia de atuação. Ramos e Santos (2003, p. 5)

Também se faz necessário utilizar técnicas de organização destes dados que segundo Almeida e Bax (2003, p.1) “fazem parte de um corpo de disciplinas que busca melhorias no tratamento de dados, atuando na sua seleção, no seu processamento, na sua recuperação e na sua disseminação”. Assim, a organização dos dados busca proporcionar uma melhor estruturação de forma a facilitar a realização automática da análise e posteriormente auxiliar na disponibilização das informações para melhor entendimento.

Um dos métodos utilizados para que seja possível a obtenção do conhecimento contido no questionário dos estudantes é a Mineração de Dados (MD), pois “trata-se da técnica utilizada para a obtenção de informações a partir de grandes quantidades de dados. Com o uso da MD é possível analisar diferentes tipos de elementos e encontrar diferentes tipos de relações entre eles” (VILARINHO, 2017, p.15). Complementando a definição, Martins (2010, p. 5) diz que “a mineração de dados (MD), é um processo que descobre informações relevantes, como padrões, associações, mudanças, anomalias e estruturas em dados armazenados em banco de dados.” Assim pode ser verificado na Figura 3 a seguir as etapas da mineração dos dados.

Figura 3 - Mineração dos dados.



Fonte: Fayyad et al. (2020, p. 41)

É possível visualizar na Figura 3 as etapas para realizar a mineração dos dados. Primeiramente é preciso selecionar a base de dados que será utilizada e os dados que são desejados para análise, posteriormente é realizado o pré-processamento que consiste na remoção das inconsistências que existirem com os dados para que não possa interferir no algoritmo de mineração, em seguida a transformação que consiste em projetar ou reduzir o tamanho de tais dados.

Após essa preparação começa a etapa da mineração propriamente dita, que consiste em selecionar as técnicas e algoritmos necessários para extrair os dados desejados. Por último acontece a avaliação, responsável por verificar se realmente os métodos aplicados anteriormente estão extraíndo dados úteis ao que foi desejado.

Este processo é intitulado de *Knowledge Discovery in Databases* (Extração de Conhecimento em Bancos de dados – KDD) e durante os anos surgiram diversas definições e etapas, sendo a mais utilizada a apresentada por Cretton (2016, p. 48, apud Fayyad, Piatetsky-Shapiro e Smyth, 1996), “que apresenta o KDD como um processo incomum de exploração e descobrimento de diferentes padrões, ainda desconhecidos, que sejam interessante, corretos e de fácil entendimento.” Assim, esse método vem sendo utilizado para realizar a descoberta de diferentes combinações desconhecidas dentro de grandes quantidades de dados.

Como foi citado anteriormente o processo KDD pode ser apresentado de diversos modelos diferentes em suas etapas mais que segundo Cretton (2016) “ sua

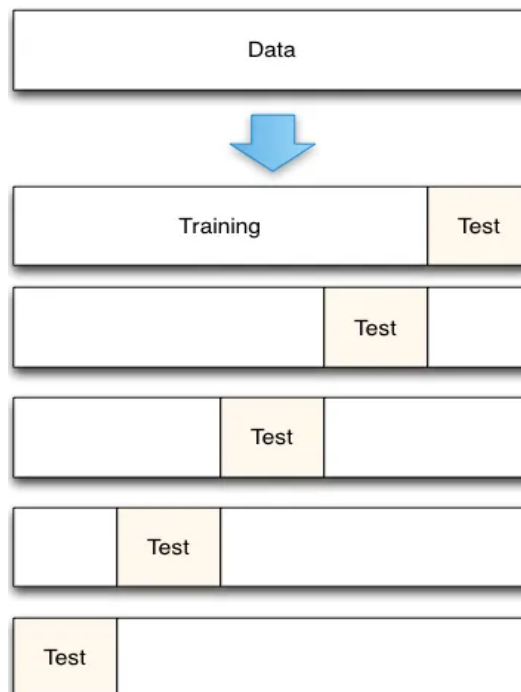
ideia e estrutura básica não mudam, uma vez que estes buscam sempre encontrar novos padrões e podem ter sua estrutura reduzida para três etapas, pré processamento, mineração de dados e pós-processamento. ”

Dentro desta etapa de mineração foi utilizado o Algoritmo KNN (K Nearest Neighbor) que é um dos algoritmos mais utilizados em *Data Mining* e *Machine Learning* e também um dos mais simples, analisando seu processo de cálculo. O KNN é um classificador que aprende se baseando no quão similar um vetor é do outro, assim tornando possível determinar de qual grupo uma determinada amostra faz parte a partir dos seus vizinhos.

Neste processo um método muito utilizado para analisar a capacidade de generalização de um modelo e ajudar a identificar o *overfitting*, que é quando um modelo estático se ajusta ao conjunto de dados anteriormente observado mas que por outro lado se torna ineficaz para prever novos resultados, é intitulado *Cross-validation*.

O *cross-validation* ou validação cruzada é uma técnica que, a partir de uma avaliação, verifica a capacidade de generalização de um modelo, com base em um conjunto de dados. Na Figura 4, a seguir, é possível verificar um exemplo de como funciona este método.

Figura 4 - Cross-validation



Fonte: Rodrigo Santanar (2020)

Como demonstrado na Figura 4 este método funciona da seguinte forma: os dados inicialmente ficam armazenados em conjunto, posteriormente eles são quebrados em pequenas partes de teste e treino, com isso estes dados são embaralhados e divididos em k números de dobras, por fim a cada interação tem-se um conjunto de treino e de teste diferente.

Ter uma boa organização desses dados possibilita um melhor aproveitamento do armazenamento das informações evitando aglomeração de informações desnecessárias, e também permite desenvolver métodos eficientes para realizar o tratamento apenas do que realmente foi solicitado para ser analisado.

2.3 VISUALIZAÇÃO DA INFORMAÇÃO

Val (2010, p. 48) descreve que “o uso de ferramentas visuais e interativas disponibilizadas por computadores deu origem à área de estudo da visualização de informações que foca no estudo de representações visuais e interativas com o propósito de ampliar a cognição”. A partir disto, “uma forma efetiva de encontrar informações importantes em grandes massas de dados com várias dimensões é vendo figuras que correspondem a estes números, ou seja, aplicando de técnicas de visualização” (GRÉGIO et al., 2020, p. 2). Estas técnicas buscam trazer clareza para que seja possível otimizar as habilidades visuais dos seres humanos fazendo que a interpretação das informações se torne menos complexa.

Para que seja possível tornar o entendimento do usuário mais claro, pode-se utilizar métodos visuais como por exemplo, o gráfico. Neste caso é necessário sempre se preocupar em saber se os dados estão completamente limpos e buscar tipos de gráficos que possibilitam uma compreensão fácil pois,

A visualização de dados ajuda a contar histórias compilando os dados em um formato mais compreensível, destacando tendências e exceções. Uma boa visualização conta uma história, eliminando elementos irrelevantes dos dados e ressaltando informações úteis (TABLEAU, 2020, p.1).

Assim, “todas estas técnicas procuram representar, em gráficos ou figuras, informações que tentam explorar ao máximo a capacidade de percepção humana, levando à interpretações mais concretas e compreensíveis” (VAZ; CARVALHO, 2004, p. 2). O uso de gráficos é, então, uma forma de apresentar os dados e é bastante utilizada em livros didáticos, jornais e revistas, mas existem outros tipos de técnicas de visualização de dados utilizando gráficos como a utilizada, no monitoramento da bolsa de valores e nos diagramas, pois o diferencial dessa forma de visualizar informações é que ela está sendo atualizada em tempo real, então as informações sempre estão se alterando. Na Figura 5 é demonstrado o gráfico de monitoramento da bolsa de valores.

Figura 5 - Gráfico de linhas.

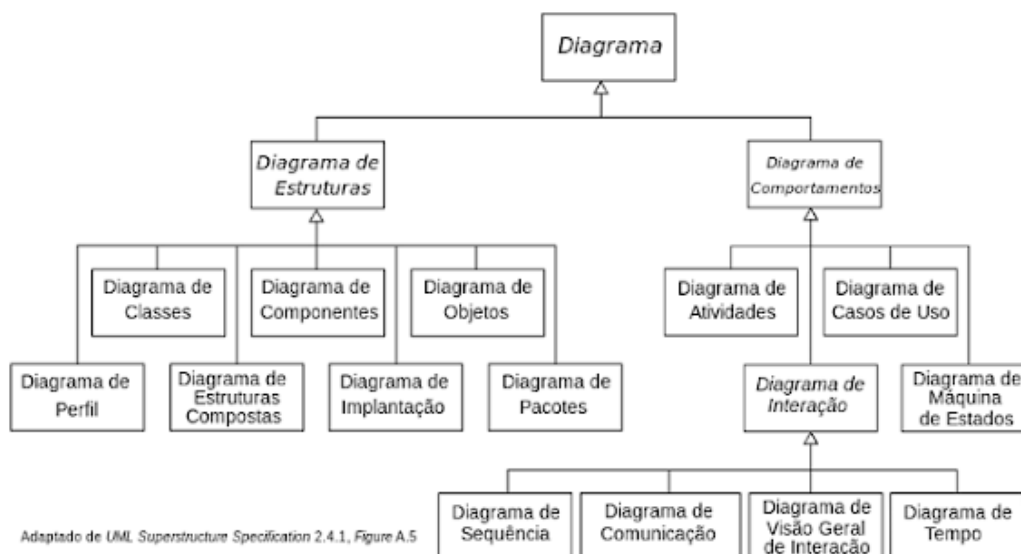


Fonte: Avatrade (2020, *online*).

Como mostra a Figura 5, o gráfico de linhas utilizado no monitoramento da bolsa de valores serve para representar um determinado preço que a bolsa está valendo em uma determinada faixa de tempo. Esse gráfico vai demonstrar as informações dos valores de acordo com o passar do tempo, informando se o valor da bolsa que está sendo analisado vai subir ou descer.

Uma outra forma de visualização de informações é a utilização de diagramas, que possuem elementos que constituem um sistema ou um conjunto. Na Figura 6 tem-se uma demonstração de um diagrama.

Figura 6 - Diagrama.

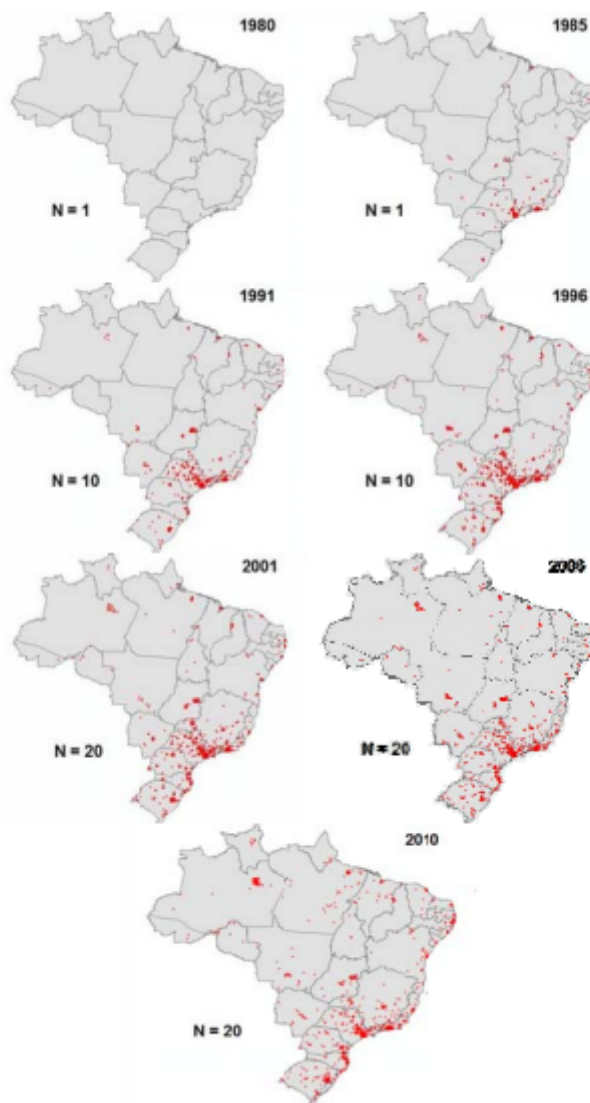


Fonte: Ventura (2016, *online*)

Na Figura 6 é demonstrado o conjunto de diagramas divididos por tipo, muito utilizado em projetos de software pois são uma boa forma de representar sua estrutura e como o sistema vai se comportar. A partir desses diagramas é possível construir um modelo executável do projeto assim tendo uma maior eficiência no desenvolvimento pelo fato de ter os passos construídos no diagrama.

Também é possível utilizar outras formas de visualização de informações, como o mapa de distribuição de pontos apresentado na Figura 7.

Figura 7 - Mapa de distribuição de pontos.



Fonte: Camboim e Sluter (2013, p.17)

Como demonstrado na Figura 7, pode-se observar uma distribuição anual de novos casos de Síndrome da Imunodeficiência Adquirida (AIDS) no Brasil. Para que seja possível observar essas informações foram utilizados mapas de pontos de contagem que são ferramentas utilizadas para análise da densidade de determinados fenômenos, tornando possível localizar os pontos com maior probabilidade de ocorrência.

Estas e outras formas de visualização da informação que venham a se mostrar interessantes ou necessárias para a demonstração dos dados que serão aqui trabalhados foram utilizadas durante o desenvolvimento do projeto como apresenta a próxima seção.

3 METODOLOGIA

Nesta seção serão apresentadas as ferramentas que foram utilizadas para o desenvolvimento do trabalho e os métodos realizados bem como as etapas que foram realizadas no projeto.

3.1 FERRAMENTAS

Para que fosse possível atender às necessidades de tratamento, mineração dos dados e padronização, foi necessária a utilização de algumas ferramentas, são elas:

- Microsoft Excel (2020): é um editor de planilhas que corresponde ao formato dos dados disponibilizados pelo Enade, onde esta ferramenta auxiliou no tratamento dos dados e exclusão das informações desnecessárias;
- Linguagem R e R Studio (2020): esta ferramenta possui diversas bibliotecas com algoritmos que possibilitam atender os objetivos da proposta, assim permitindo um maior entendimento dos dados que foram analisados;
- Linguagem Python (2020): é uma linguagem de alto nível que serviu em conjunto com o Microsoft Excel para o tratamento dos dados assim permitindo lidar com grandes quantidades de dados;
- Django (2020): esta ferramenta foi responsável pelo desenvolvimento do Backend onde está toda a lógica que foi utilizada na ferramenta, onde ocorreu o desenvolvimento dos processos de acesso, processamento e armazenamento das informações;
- Angular (2020): é uma plataforma de aplicação web para desenvolvimento front-end baseado em TypeScript, essa foi a aplicação que responsável pela interação do usuário com a plataforma de análise;
- Sqlite (2020): SQLite é uma biblioteca que implementa um banco de dados SQL utilizando a linguagem C. Programas que utilizam o SQLite têm acesso ao banco de dados sem a necessidade de executar um processo SGBD separado.
- Pandas (2020): Pandas é um pacote Python que fornece estruturas de dados rápidas e flexíveis projetadas para tornar o trabalho com os dados fácil e intuitivo.

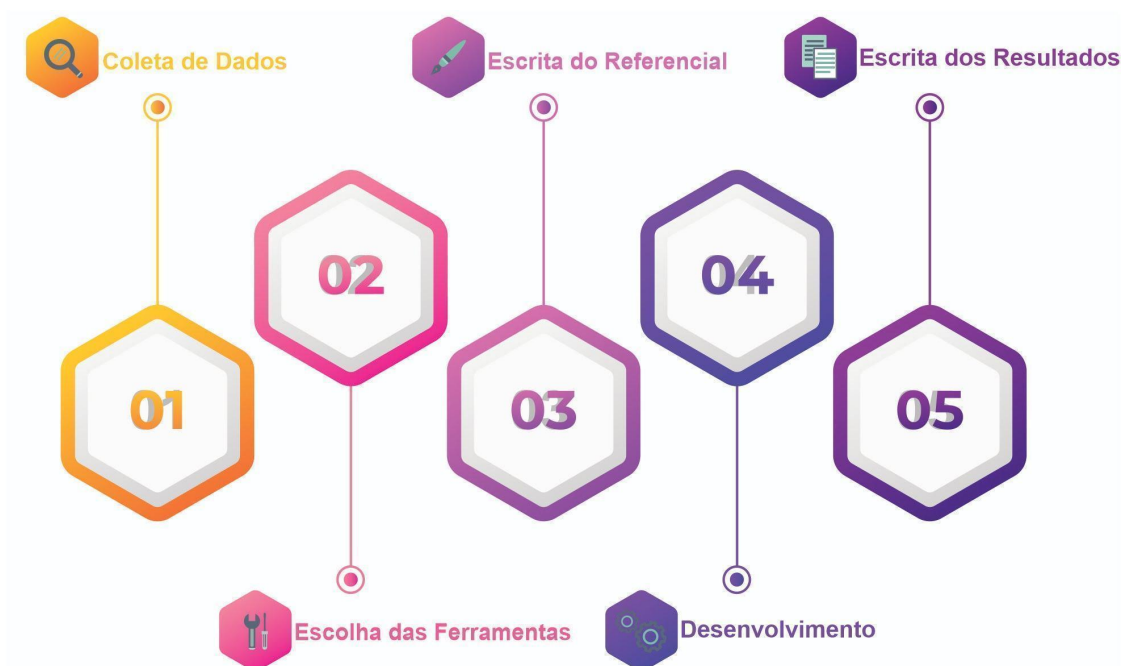
- Scikit-learn (2007): é uma biblioteca de aprendizado de máquina desenvolvido para programação em Python. Ela possui vários algoritmos como classificação, agrupamento e regressão. Esta biblioteca foi útil na realização do pré-processamento de dados e na seleção e avaliação de modelos.
- NumPy(2019): uma biblioteca Python que é usada, principalmente, para realizar cálculos em *Arrays* Multidimensionais. Ela fornece um conjunto de funções e operações que ajudam a executar facilmente cálculos numéricos. Esta biblioteca foi utilizada principalmente para gerar vetores e matrizes dos dados disponibilizados.
- Chart.js(2014): O Chart.js é uma biblioteca JavaScript de código aberto gratuita para visualização de dados em forma de gráficos utilizando apenas HTML, CSS e JS para renderizar os gráficos na tela do usuário.

3.2 MÉTODOS

Este trabalho tem o intuito de apresentar o desenvolvimento de uma ferramenta capaz de auxiliar no processo de análise e visualização do questionário do estudante do Enade. O presente trabalho foi baseado em observações e o artefato que foi desenvolvido serviu como complemento para a base teórica da pesquisa.

O desenvolvimento deste trabalho, como demonstrado na Figura 8, foi realizado seguindo etapas: coleta das informações, escolha das ferramentas, escrita do referencial, desenvolvimento, escrita dos resultados.

Figura 8 - Métodos do Trabalho.



A primeira etapa consistiu em coletar informações referentes ao trabalho verificando as informações sobre os dados do questionário do estudante, os métodos para que fosse possível realizar o tratamento dos dados para disponibilizar uma maior praticidade ao usuário que precisa dessas informações. Essa coleta de informações foi realizada por meio de pesquisas nos repositórios de informações do Enade.

A segunda etapa consistiu em identificar ferramentas que auxiliam na obtenção das informações do questionário dos estudantes que estão sendo buscadas. Estas ferramentas de tratamento foram utilizadas para tratar os dados brutos tornando assim mais prático a obtenção das informações, onde também foi necessário a utilização de uma ferramenta para desenvolver um método de visualização que auxiliará o usuário a entender os dados.

Na terceira etapa foi desenvolvido o referencial teórico utilizando como referência os conceitos e tecnologias descobertos nas etapas anteriores, onde foram transcritas as informações que foram tratadas sobre o tema do trabalho, tornando assim mais claro o intuito do desenvolvimento do trabalho.

A quarta etapa foi a fase de desenvolvimento do projeto, constituída de seis sub-etapas: extração dos dados, entendimento dos dados, aplicação do algoritmo de mineração dos dados, análise e validação da mineração, elaboração da ferramenta

para visualização e disponibilização da ferramenta. É demonstrado esse processo na Figura 9.



No desenvolvimento do trabalho, a quarta etapa do esquema anterior é ilustrada na Figura 9, dividida nas sub-etapas descritas a seguir.

A primeira sub-etapa constituiu na extração dos dados que são disponibilizados pelo Inep que vem em formato de microdados em planilhas .csv, separados por ano de realização do exame do Enade. Para a realização deste trabalho foram selecionados apenas os dados referentes ao questionário dos estudantes de determinados anos, pois periodicamente esses dados sofrem mudanças, assim tornando possível analisar as respostas dos estudantes. Esses dados serão carregados pelo administrador para que a plataforma possa analisar e disponibilizar para visualização do usuário.

Na segunda sub-etapa obteve-se o entendimento dos dados, a partir do dicionário que é disponibilizado com as planilhas foi possível compreender sua estrutura e relacionamentos assim descobrindo como se comportam os dados contidos nas planilhas.

Na terceira sub-etapa ocorreu a aplicação do algoritmo de mineração dos dados que foi realizado após a padronização dos dados como foi citado anteriormente. Com esses dados padronizados foi possível aplicar o algoritmo de mineração permitindo retirar os dados que não são necessários.

Análise e validação da mineração foi a quarta sub-etapa, responsável por verificar se todos os dados que passaram pelo algoritmo realmente estão no formato correto para a análise dos dados do questionário dos estudantes.

A elaboração da ferramenta para visualização corresponde a quinta sub-etapa, quando foram utilizadas técnicas computacionais para o desenvolvimento web. A ferramenta utilizou técnicas de visualização que foram escolhidas, estudadas e estruturadas para apresentar o conteúdo extraído dos dados, tornando o conteúdo de fácil entendimento e agradável.

A sexta sub-etapa consistiu na disponibilização da ferramenta que, após a sua conclusão, ficará disponível ao público para utilização para poder analisar dados do questionário dos estudantes.

Por fim, na quinta etapa da metodologia do trabalho ocorreu a escrita do desenvolvimento e os resultados que foram obtidos para conclusão da pesquisa.

4 RESULTADOS E DISCUSSÃO

Com o objetivo de realizar a análise dos dados do questionário dos estudantes foi utilizada a base de dados do Enade que é disponibilizada todos os anos pelo Inep. A base de dados utilizada inicialmente foi a do ano de 2017, que é composta por 150 variáveis como: idade, sexo, código da categoria administrativa da IES (Pública Federal, Pública Estadual, Pública Municipal, Privada com fins lucrativos, Privada sem fins lucrativos, Especial), entre outras variáveis distintas. A composição destas variáveis possui aproximadamente 500 mil registros.

Dadas as características destes questionários, que são periodicamente alterados, organizou-se o processamento inicialmente para um ano em específico, no caso 2017, para que se verificasse que atributos podem ser trabalhados de forma fixa e como é possível adequar um sistema a eventuais alterações na sua base de entrada de dados.

Para que fosse possível obter esses dados tratados e transformados para serem minerados foi utilizado o processo KDD que tem como objetivo encontrar padrões desconhecidos dentro de uma base de dados. Este processo possui um conjunto de cinco etapas: seleção, pré-processamento, transformação e mineração de dados e avaliação. Para realizar este trabalho foi escolhida a API Pandas pois dispõe de métodos, funções e objetos que auxiliam na manipulação e visualização destes dados do questionário dos estudantes. Seguindo o fluxo do processo KDD cada etapa será demonstrada a seguir.

4.1 SELEÇÃO

Nesta etapa ocorreu o início da manipulação dos dados, que foi realizada de forma manual e um dos primeiros critérios observados foi a delimitação dos dados especificando um conjunto menor para avaliação, voltada a um grupo específico, no caso, o estado do Tocantins. Esta delimitação também ocorre pela necessidade de processamento que o algoritmo exige, pois quanto maior o conjunto mais pesado este processamento se torna e assim exige equipamentos mais potentes para executar todo o processamento. Assim, o total de 537.437 registros foi reduzido

para 5.306 registros que correspondem apenas aos dados relacionados a instituições de Ensino do Tocantins (TO) para que fosse possível iniciar a análise.

Ainda nesta etapa, foi realizada a exclusão de variáveis que não seriam utilizadas na análise, estas variáveis foram excluídas com a ajuda do especialista M.e Fabiano Fagundes assim mantendo o foco nos atributos que realmente interessam. Além dos dados que fornecem informações que auxiliam a avaliação com base na demografia, como idade e sexo, foram mantidos também os atributos de interesse para este trabalho, ou seja, os mais relevantes do questionário do estudante como demonstrados na Figura 10.

Figura 10 - Dados selecionados

CO_IES	CO_ORGA	CO_GRUP	CO_CURS	CO_MODA	CO_MUNI	NU_IDADE	TP_SEXO	ANO_FIM	ANO_IN_	CO_TURN	TP_PRES	QE_I01	QE_I02	QE_I04
453	10020	4006	9440	1	1721000	23	M	2011	2014	4	555 A	A	D	
453	10020	4006	9440	1	1721000	33	M	2003	2012	4	555 B	A	E	
453	10020	4006	9440	1	1721000	22	F	2013	2014	4	555 A	A	C	
453	10020	4006	9440	1	1721000	29	M	2005	2010	4	555 E	B	C	
453	10020	4006	9440	1	1721000	22	M	2013	2014	4	555 A	D	D	
453	10020	4006	9440	1	1721000	21	M	2013	2014	4	555 A	D	F	
453	10020	4006	9440	1	1721000	24	M	2013	2014	4	555 B	A	D	
453	10020	4006	9440	1	1721000	28	M	2005	2007	4	555 A	D	D	
3849	10028	2402	17134	1	1718204	43	M	1996	2010	1	222			
3849	10028	2402	17134	1	1718204	30	F	2009	2013	1	555 E	B	B	
3849	10028	2402	17134	1	1718204	30	F	2005	2011	1	555 A	B	D	
3849	10028	2402	17134	1	1718204	25	F	2010	2011	1	555 A	D	B	
3849	10028	2402	17134	1	1718204	23	F	2011	2013	1	555 A	D	C	
3849	10028	2402	17134	1	1718204	28	F	2009	2013	1	555			
3849	10028	2402	17134	1	1718204	26	M	2009	2012	1	555 A	D	B	
3849	10028	2402	17134	1	1718204	22	F	2012	2013	1	555 A	C	B	
3849	10028	2402	17134	1	1718204	31	M	2003	2013	1	222			
3849	10028	2402	17134	1	1718204	24	F	2010	2011	1	555 A	A	B	

Como o foco desta análise são os dados obtidos com o questionário do estudante, foi realizada uma verificação nas perguntas para selecionar quais seriam as melhores questões para serem utilizadas. Esta seleção foi realizada em conjunto com o especialista do domínio que forneceu informações das questões que realmente seriam relevantes para o trabalho. Assim após essa verificação foram selecionadas algumas perguntas como: “Qual o seu estado civil?”, “Qual é a sua cor ou raça?”; que são perguntas que permitem uma avaliação demográfica. Além de outras perguntas relacionadas à instituição de ensino, sua estrutura e a didática aplicada nos cursos, como: “No curso você teve oportunidade de aprender a trabalhar em equipe?”, “As condições de infraestrutura das salas de aula foram adequadas?”; entre outras questões que irão compor os dados a serem analisados.

4.2 PRÉ PROCESSAMENTO

Após a seleção dos dados com os quais se trabalhou, foi iniciado o pré processamento. Nesta etapa foi observado que os dados possuem informações que não foram preenchidas pelos alunos, como também dados que estão no formato *string* e dados *float*, informações estas que estão presentes e são solicitadas nas questões do questionário do estudante.

Primeiramente foram identificados os campos nulos utilizando o método `fillna` disponibilizado pela *Api Pandas*. Assim, utilizando este método, todos os campos nulos foram substituídos por *strings* vazias. Este processo foi realizada para tornar possível transformar os campos respondidos com *strings* por números inteiros pois, a princípio, se já fosse substituído os campos nulos pelo inteiro 0, o método utilizado para conversão das strings não aceitaria pois identificaria número inteiro no meio das strings. Por fim, foi feita a substituição de todos os campos vazios em 0 e a conversão de todos os números float em inteiros, assim como demonstrado pela Figura 11.

Figura 11 - Pré-Processamento

```
def PreProcessamento(dados):
    dados.fillna("", inplace=True) #Os campos nulos es

    non_numerical = ['TP_SEXO', 'QE_I01', 'QE_I02', 'QE_I0
    le = preprocessing.LabelEncoder()
    for x in non_numerical:
        le.fit(dados[x])
        # converte string em numero
        dados[x] = le.transform(dados[x].astype(str))

    Dados=dados.replace([""],0)
    Dados=Dados.astype(int)
    print(Dados)
    return Dados
```

Como pode-se observar na Figura 11 o fluxo realizado no pré processamento ocorreu, inicialmente, como citado anteriormente: localizar e substituir os campos vazios. Posteriormente foi realizada a transformação das strings em números como na Figura 12 a seguir.

Figura 12 - Transformação de string em número.

QE_I01	QE_I02	QE_I04	QE_I01	QE_I02	QE_I04
A	A	D	1	1	4
B	A	E	2	1	5
A	A	C	1	1	3
E	B	C	5	2	3
A	D	D	1	4	4
A	D	F	1	4	6
B	A	D	2	1	4
A	D	D	1	4	4

Como demonstrado na Figura 12 as questões foram respondidas com *strings* e a partir do método utilizado do `sklearn` foi possível realizar essa conversão e mantendo a ordem das respostas, onde A se tornou 1, B se tornou 2 e assim por diante de acordo com a quantidade de alternativas existentes em cada questão, esta conversão foi realizada para que fosse possível passar todos os dados pelo algoritmo do KNN. Posteriormente foi realizado o preenchimento dos campos vazios por 0 e, por fim, foi identificado que a API Pandas transforma automaticamente os dados de discretos para contínuos, fazendo com que um número *int* saia como float após o tratamento. Para que não ficasse assim e futuramente interferisse em outro processo foi preciso transformá-los novamente para discreto para que fosse possível continuar com a análise sem alteração dos dados que estão sendo recebidos. Ao final deste tratamento os dados são retornados da seguinte maneira como pode-se observar na Figura 13 a seguir.

Figura 13 - Pós-Processamento.

ANO_FIM_EM	ANO_IN_GRAD	CO_TURNO_GRADUACAO	TP_PRES	QE_I01	QE_I02	QE_I04	Coluna	QE_I07	QE_I08	QE_I09	QE_I11
2011	2014	4	555	1	1	4	4	1	1	4	3
2003	2012	4	555	2	1	5	4	3	3	6	2
2013	2014	4	555	1	1	3	4	2	2	4	5
2005	2010	4	555	5	2	3	2	2	4	5	7
2013	2014	4	555	1	4	4	5	1	2	3	3
2013	2014	4	555	1	4	6	5	3	4	5	8
2013	2014	4	555	2	1	4	4	3	2	1	5
2005	2007	4	555	1	4	4	5	4	3	5	5
1996	2010	1	222	0	0	0	0	0	0	0	0
2009	2013	1	555	5	2	2	4	4	2	5	1
2005	2011	1	555	1	2	4	5	6	3	3	1
2010	2011	1	555	1	4	2	5	4	1	2	1
2011	2013	1	555	1	4	3	4	4	1	3	1
2009	2013	1	555	0	0	0	0	0	0	0	0
2009	2012	1	555	1	4	2	5	1	1	6	1
2012	2013	1	555	1	3	2	4	1	1	4	1
2003	2013	1	222	0	0	0	0	0	0	0	0
2010	2011	1	555	1	1	2	2	5	2	2	1
2011	2013	1	555	1	4	1	2	5	3	3	1
2011	2012	1	555	1	2	2	4	3	2	2	1
2005	2013	1	555	1	1	2	2	5	1	2	1
2007	2012	1	555	2	1	2	4	4	1	2	1

Assim como demonstrado na Figura 13 todas as lacunas que constavam nos dados foram preenchidas e todas as strings substituídas por inteiros, tornando agora possível realizar o próximo passo do KDD que é a transformação.

4.3 TRANSFORMAÇÃO

A etapa de transformação antecede a fase de mineração. Nela os dados devem ser devidamente formatados, com a finalidade de melhorar o entendimento dos dados como podemos visualizar na Figura 14 a seguir.

Figura 14 - Títulos das colunas.

CO_IES	CO_ORGA	CO_GRUPI	CO_CURSO	CO_MODAL	CO_MUNI	NU_IDADE	TP_SEXO	ANO_FIM
453	10020	4006	9440	1	1721000	23	M	2011
453	10020	4006	9440	1	1721000	33	M	2003
453	10020	4006	9440	1	1721000	22	F	2013
453	10020	4006	9440	1	1721000	29	M	2005
453	10020	4006	9440	1	1721000	22	M	2013

Na Figura 14 é possível verificar que todas as colunas possuem um identificador que está sendo apresentado em códigos. Para que seja mais simples o entendimento de cada coluna foi realizada a transformação destes identificadores utilizando a função `rename` que realiza a alteração de todos os códigos identificados das colunas selecionados como podemos verificar no código da Figura 15.

Figura 15 - Código de transformação.

```
def tratamento(dados):
    Dados=dados.rename(columns={'CO_IES':'Codigo instituicao'},
    return Dados
```

Como demonstrado na Figura 15 são recebidos os dados que se deseja alterar passando assim pela função responsável por renomear, onde se informa o identificador da coluna que se deseja alterar e a informação que irá substituir. Assim, como resultado obtemos os resultados como demonstrado na Figura 16.

Figura 16 - Resultado da transformação

Codigo instituicao	Org. academica	Area do curso	Codigo do curso	Modalidade de Ensino
453	10020	4006	9440	
453	10020	4006	9440	
453	10020	4006	9440	
453	10020	4006	9440	
453	10020	4006	9440	
453	10020	4006	9440	
453	10020	4006	9440	

Assim como demonstrado na Figura 16 foram obtidos os novos identificadores das colunas de dados do questionário do Enade. Com a finalização da reformulação dos dados torna-se possível iniciar a etapa de mineração dos dados.

4.4 BANCO DE DADOS

Após a finalização da etapa de transformação foi identificada a necessidade da inclusão destes dados tratados no banco de dados para que, quando o usuário realizar a requisição das informações para gerar um gráfico, estas informações são requisitadas ao banco de dados transformadas para serem recebidas pelos gráficos disponibilizados para o usuários. Pode-se verificar na Figura 17 a organização realizada para o armazenamento dos dados.

Figura 17 - Banco de Dados

```
class Resultado(models.Model):
    kVizinhos = models.CharField(max_length=16, blank=True, null=True, unique=True)
    dobrasF = models.CharField(max_length=16, blank=True, null=True, unique=True)
    erro = models.CharField(max_length=16, blank=True, null=True, unique=True)

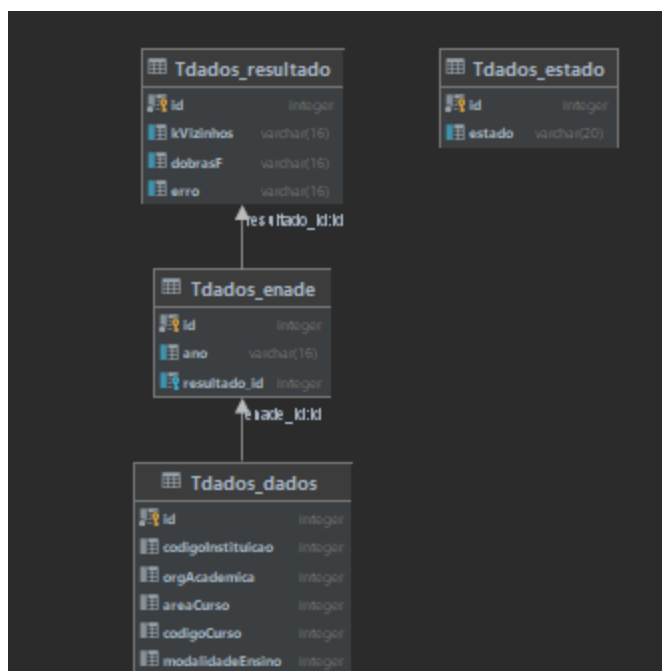
class Enade(models.Model):
    ano = models.CharField(max_length=16, blank=True, null=True, unique=True)
    resultado = models.ForeignKey(Resultado, on_delete=models.CASCADE, blank=True, null=True)

class Estado(models.Model):
    id = models.IntegerField(primary_key=True)
    estado = models.CharField(max_length=20, blank=True, null=True, unique=True)

class Dados(models.Model):
    codigoInstituicao = models.IntegerField(blank=True, null=True)
    orgAcademica = models.IntegerField(blank=True, null=True)
    areaCurso = models.IntegerField(blank=True, null=True)
    codigoCurso = models.IntegerField(blank=True, null=True)
    modalidadeEnsino = models.IntegerField(blank=True, null=True)
    municipioCurso = models.IntegerField(blank=True, null=True)
    idade = models.IntegerField(blank=True, null=True)
```

Nesta Figura 17 é demonstrada toda a estruturação realizada para que cada informação que constasse no arquivo com os dados no questionário inseridos na ferramenta pudesse ser armazenada, além dos dados do questionário dos estudantes que foi armazenado, também foi incluído para cada ano os resultados obtidos no KNN para que posteriormente pudesse ser consultado. É possível visualizar na Figura 18 a seguir a modelagem do banco de dados.

Figura 18 - Modelo de Dados



Na Figura 18 é possível ver a modelagem do banco que explica as características de funcionamento e comportamento do software a partir do qual ele será criado e estruturado.

4.5 MINERAÇÃO

Inicialmente, para conseguir prever a assertividade do modelo utilizado, precisou-se usar parte dos dados do questionário do Enade para treinamento e outra para testes da efetividade do algoritmo de classificação. Primeiramente foi preciso separar os dados em matriz X e um vetor y, como demonstrado na Figura 19 a seguir.

Figura 19 - Dados de treino

```
for i in dados.columns:
    # criar uma matriz X e o vetor y
    x = np.array(dados.iloc[:, 0:12])
    y = np.array(dados[i])
```

Como demonstrado na Figura 19 iniciou-se a criação dos dados de treino e de teste para que assim fosse possível realizar a validação do algoritmo. Foi

realizado um *loop* onde são passadas as colunas que compõem o conjunto de dados que se deseja analisar.

Para realizar o KNN, um dos processos mais importantes é a definição do parâmetro K e a definição de qual será o parâmetro F para o K -Fold. Assim foi utilizado o K -fold como o principal método para realizar esta avaliação. Na Figura 20 a seguir é possível visualizar algumas variáveis que serão utilizadas.

Figura 20 - Parâmetros

```
neighbors = list(range(1,100,2))
cv_list = list(range(10,40))
k_list = []
fold_list = []
cv_scores = []
```

Como é possível visualizar na Figura 20 estão sendo criadas listas para contemplar os intervalos de variação dos parâmetros do K e F e listas para armazenar valores. O *neighbors* irá receber os intervalos de números ímpares de k para utilização do KNN, a *cv_list* está recebendo os intervalos de f para utilização do k -fold. A *k_list*, *fold_list* e a *cv_scores* foram criadas para armazenamento de valores.

Em sequência, será iniciada de fato a execução do algoritmo do KNN. Após o início da execução do algoritmo varia-se o parâmetro k e o *cross-validation* variando também o agrupamento de F de folds, assim populando a acurácia deste modelo para cada K e F . Como demonstrado na Figura 21.

Figura 21 - Algoritmo KNN

```
for k in neighbors:
    for f in cv_list:
        knn = KNeighborsClassifier(n_neighbors=k)
        scores = cross_val_score(knn, x, y, cv=f, scoring='accuracy')
        cv_scores.append(scores.mean()) #popular listas
        k_list.append(k)
        fold_list.append(f)
```

Na Figura 21 é possível ver o Algoritmo do KNN em funcionamento. A principal biblioteca utilizada nesta etapa foi o *sklearn*, onde utilizou-se

KNeighborsClassifier que é o classificador que implementa a votação de k-vizinhos mais próximos. O *cross_val_score* responsável pela pontuação por validação cruzada.

É possível expressar o resultado do algoritmo de diversas maneiras, mas neste caso utilizou-se o erro empírico de classificação MSE (*misclassification error*) como demonstrado na Figura 22 a seguir.

Figura 22 - Erro empírico

```
MSE = [1 - x for x in cv_scores]

df_1 = pd.DataFrame(k_list, columns=['k_list'])
df_2 = pd.DataFrame(fold_list, columns=['fold_list'])
df_3 = pd.DataFrame(MSE, columns=['MSE'])
df_knn = pd.concat([df_1, df_2, df_3], axis=1)

optimal_k = min(df_knn['MSE'])

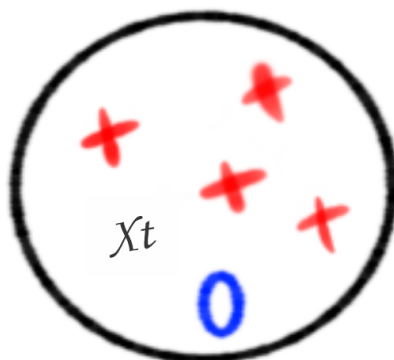
index_opt = df_knn[df_knn['MSE'] == optimal_k].index.item()
```

Como demonstrado na Figura 22 calcula-se o erro com base na acurácia obtida anteriormente. No MSE calcula-se o erro, posteriormente é feita a construção do *dataframe* para localizar os valores de K e F do menor erro, em seguida o *optimal_k* retorna o menor erro obtido e por fim o *index_opt* de disponibilizar os valores de k e f do menor erro obtido.

4.6 AVALIAÇÃO

Após os dados serem selecionados, processados e tratados eles foram submetidos ao algoritmo do KNN para permitir extrair alguma informação que, no caso, foram: o número ótimo de k vizinhos, o número ideal de dobra f e o erro de classificação. No número ótimo de k vizinhos o algoritmo funciona da seguinte maneira: partindo de uma instância X_t o algoritmo encontra os k vizinhos mais próximos de X_t dentro do conjunto de treinamento disponibilizado. Assim a classe de X_t é dada pela classe que ocorre com maior frequência entre os k vizinhos, como demonstrado na Figura 23 a seguir.

Figura 23 - k vizinhos



Na Figura 23 é possível visualizar os 5 vizinhos mais próximos da instância denominada X_t . A partir destes vizinhos foi possível visualizar que X_t possui quatro vizinhos do Grupo vermelho e um vizinho do grupo azul, com isso ao aplicar-se o KNN a instância X_t será classificada como vermelho pois este grupo possui uma maior representatividade na vizinhança de X_t .

O número ideal de dobra K-fold que intitulou-se f para não ficar parecido com o K do KNN, estima o erro do método de aprendizado em observações não utilizadas no treino, assim tornando possível estimar como o modelo construído irá se comportar em novos dados. Este método consiste em dividir a base em f pedaços. Assim, para cada pedaço estima-se o método sem a presença desta parte e verificou-se o erro médio no pedaço não utilizado durante o treino. Por fim, a estimativa do erro de predição de *cross validation* é dada pela média dos erros médios nos k pedaços.

Por fim é realizada a classificação de erro que foi alcançada utilizando o erro empírico de classificação MSE. Desta forma tornou possível calcular a margem de erro com base na acurácia obtida. Assim o algoritmo retorna estas três variáveis como demonstrado na Figura 24.

Figura 24 - Variáveis finais

```
Int64Index([1406], dtype='int64')
O número ótimo de vizinhos k é 93
O número ideal de dobras f é 36
Erro de classificação incorreta é 0.575784
```

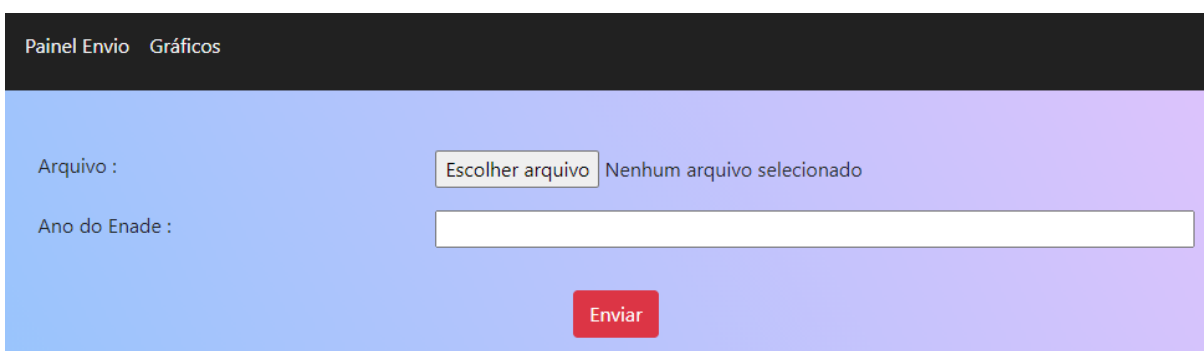
Com a Figura 24 é possível verificar que o número ótimo de k vizinhos foi 93, o número ideal de dobras f é 36 e a margem de erro de classificação foi 0.57 ou

57%. Estes resultados foram obtidos utilizando como base os dados do questionário do Enade específicos do estado do Tocantins e com as colunas pré definidas no processo de seleção.

4.7 VISUALIZAÇÃO DOS DADOS

Inicialmente para que seja possível visualizar os dados em gráficos é necessário inseri-los na ferramenta para que ela possa realizar todo o processo que foi descrito do tópico 4.1 a 4.6 e assim retornar os dados para serem visualizados nos gráficos. Esta primeira etapa de inserção dos dados precisa ser realizada apenas uma única vez para cada ano que se deseja visualizar a partir dos dados tratados na etapa de seleção. Este processo é realizado pelo administrador da ferramenta como podemos visualizar na Figura 25 a seguir.

Figura 25 - Inserir dados



Painel Envio Gráficos

Arquivo : Escolher arquivo Nenhum arquivo selecionado

Ano do Enade :

Enviar

Como pode-se visualizar na Figura 25, para que o administrador possa inserir os dados existem dois campos, que são o campo dos dados desejados e o campo do ano ao qual se refere estes dados. Para este trabalho os dados que estão sendo inseridos já estão sendo tratados inicialmente como descrito no tópico 4.1..

Para visualizar os dados foi realizado um último tratamento para que fosse possível adequá-los ao método de inserção exigido pelo gráficos. Assim, foi realizado o tratamento para cada coluna existente no conjunto como demonstrado na Figura 26 a seguir.

Figura 26 - Tratamento gráfico

```
class processar_api(APIView):
    permission_classes = [permissions.AllowAny, ]

    def post(self, request, format=None):
        enade = Enade.objects.get(ano=request.data["ano"])
        if(enade):
            dados = Dados.objects.filter(enade=enade)

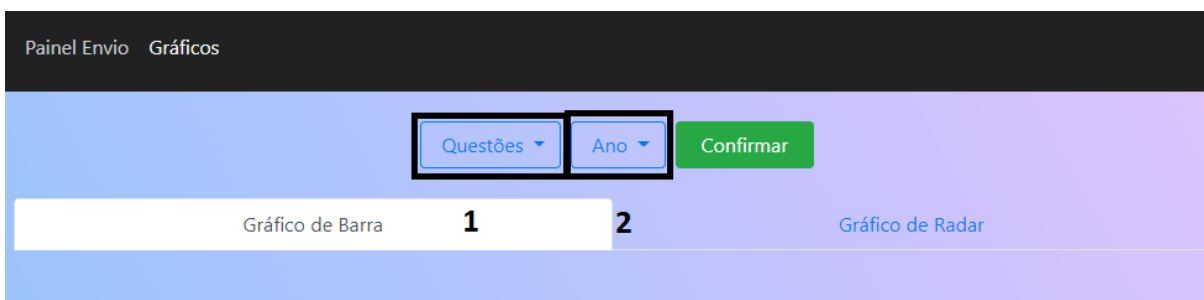
            data_values = {}
            for column in request.data["columns"]:
                for line in dados:
                    if column not in data_values:
                        data_values[column] = {}

                    if column == 'Questao_01':
                        if line.questao01 not in data_values[column]:
                            data_values[column][line.questao01] = 1
                        else:
                            data_values[column][line.questao01] += 1

                    if column == 'Questao_02':
                        if line.questao02 not in data_values[column]:
                            data_values[column][line.questao02] = 1
                        else:
                            data_values[column][line.questao02] += 1
```

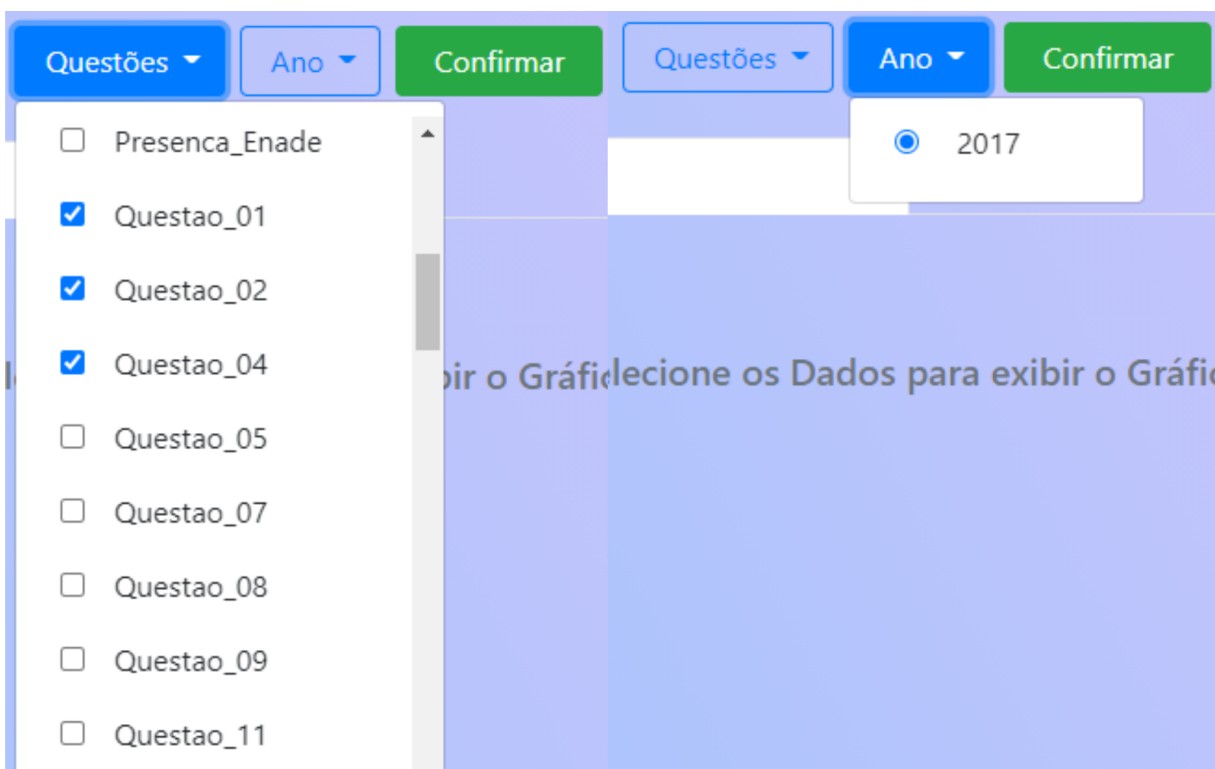
Como demonstrado na Figura 26 seleciona cada coluna e é identificada cada opção existente dentre o conjunto de dados de cada coluna como, por exemplo, sobre a Questão 1: ela possui cinco alternativas onde cada participante do Enade escolheu a que melhor se adequa, com isso tem-se um grande conjunto de respostas que estão gravadas de forma aleatória pelo Inep. Esta etapa está identificando estas alternativas inicialmente, posteriormente somando a quantidade de ocorrências para cada alternativa, retornando para o usuário a quantidade de vezes que aquela alternativa foi selecionada. Na Figura 27 a seguir pode-se visualizar como a solicitação das informações é realizada pelo usuário.

Figura 27 - Gráfico



Para que fosse possível visualizar o gráfico dos dados disponibilizados pela ferramenta, inicialmente selecionou-se na opção 1, demonstrada na Figura 27, as colunas que se deseja comparar. A ferramenta permite que sejam selecionadas quantas colunas desejar visualizar. Na opção 2 são disponibilizados os anos dos conjuntos de dados cadastrados na ferramenta, assim o usuário irá selecionar o ano que deseja visualizar como demonstrado na Figura 28.

Figura 28 - Seleção das opções



Após a seleção demonstrada da Figura 28 das colunas desejadas e do ano o usuário irá confirmar e assim serão demonstrados os gráficos com a comparação

das respectivas colunas selecionadas do determinado ano que foi cadastrado, que neste caso é 2017. Assim, para cada visualização o usuário realiza este processo.

A plataforma disponibiliza a visualização dos dados em dois tipos de gráficos, que são os gráficos de Barra e de Radar, estes tipos de gráficos foram selecionados pois possuem uma disposição que facilita muito na compreensão dos dados ali representados, e permite que os valores sejam mistos, tornando possível que os valores de cada questão fiquem no mesmo elemento gráfico como demonstrado na Figura 29 e Figura 30.

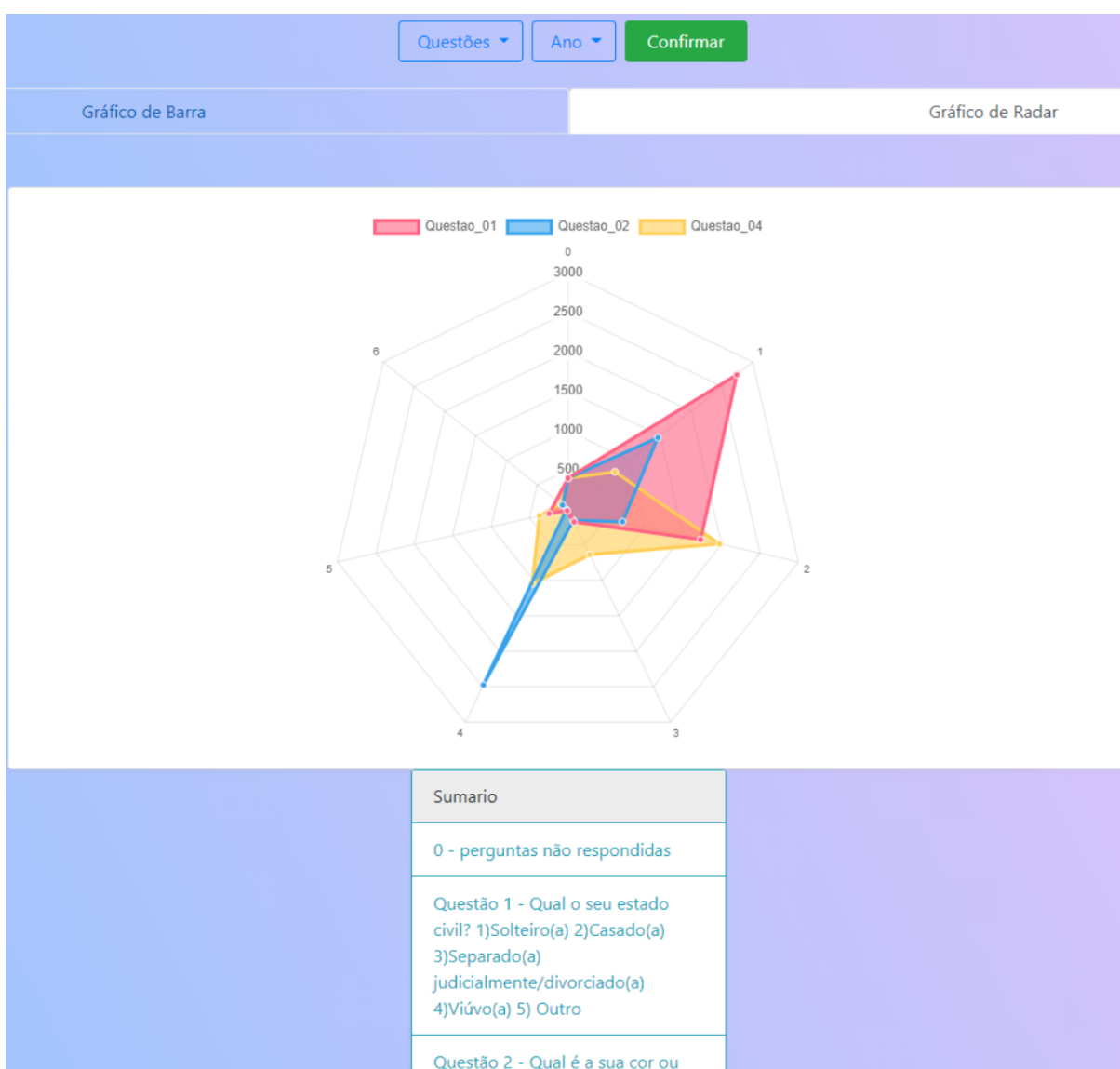
Figura 29 - Gráfico de barra



Na Figura 29 pode-se observar o gráfico de barra que está estruturado da seguinte maneira. Tem-se demonstrado neste gráfico as questões de número 1, 2 e 4, que estão estruturadas pelas alternativas que as compõem. Assim pode-se

visualizar no conjunto de colunas 1 a quantidade de vezes que a alternativa 1 foi assinalada em cada questão, no conjunto de colunas 2 a quantidade de vezes que a alternativa 2 foi assinalada e assim sucessivamente. Ou seja, pode-se ver que para a questão um, representada pela cor rosa a, alternativa 1 foi assinalada 2741 vezes, para a questão dois esta alternativa foi marcada 1461 vezes e a questão 4, 763 vezes. E assim segue no gráfico o demonstrativo das outras outras alternativas das questões seleccionadas para visualização no gráfico. Lembrando que a coluna 0 demonstra a ocorrência de pessoas que não responderam às respectivas questões.

Figura 30 - Gráfico de radar



Na Figura 30 são visualizadas as mesmas questões que foram demonstradas na Figura 29 estruturadas no gráficos de radar para que possa ser possível ter uma perspectiva diferente pois é um método gráfico de apresentar dados multivariáveis na forma de um gráfico bidimensional representadas em eixos que partem de um mesmo ponto. A seguir nas Figura 31 e Figura 32 demonstra mais um exemplo de relacionamentos das informações utilizando as questões 57 e 62.

Figura 31 - Exemplo Barras



Pode-se visualizar mais um gráfico demonstrado na Figura 31 com o relacionamento das questões 57 e 62 onde vê-se inicialmente o quantitativo de pessoas que não responderam às questões na coluna zero. Estes dados estão

sendo visualizados em um gráfico que é composto por barras retangulares com comprimento proporcional aos valores apresentados.

Figura 32 - Exemplo Radar



Na Figura 32 apresenta-se a comparação das questões 57 e 62 utilizando o gráfico de radar onde foi possível verificar a ocorrência de de alternativas selecionadas pelos participantes e visualizar as alternativas que mais foram selecionadas.

5 CONSIDERAÇÕES FINAIS

Este trabalho apresentou o desenvolvimento de uma ferramenta que realiza o tratamento dos dados do Questionário dos Estudantes que realizam a prova do Enade, onde a partir destes dados tratados e manipulados por um algoritmo de mineração disponibilizam em formas de gráficos estas informações. Durante este processo foram realizados estudos para que fosse possível identificar maneiras de como tratar estes dados de forma a deixá-los preparados para passarem pelo método do KNN, método de classificação não paramétrico que é utilizado para classificação ou regressão.

Posteriormente foram realizados estudos para verificar os melhores métodos para demonstrar este conjunto de dados para os usuários, com isso foi estudada a variedade de gráficos que se tem disponíveis para utilização, pois os gráficos são um meio de expressar visualmente dados ou valores numéricos de maneira diferentes assim permitindo uma maior compreensão dos dados.

Após a pesquisa realizada sobre métodos de tratamento e visualização foi realizado o tratamento nos dados e processamento assim retornando as métricas calculadas pelo KNN. Todo o conjunto de dados tratado foi armazenado para tornar possível a visualização pelos usuários assim podendo analisá-los de uma maneira visualmente mais agradável, em formato de gráficos. Esta visualização foi elaborada para que o usuário consiga, dentre o grande conjunto de dados, formar combinações de colunas e compará-las.

Um dos passos que não foram desenvolvidos durante o processo de inserção foi o tratamento automático dos dados, que iria tornar esta etapa de inserção do conjunto de dados pelo administrador mais simples. Este tratamento seria realizado automaticamente pela ferramenta porém não foi possível ser finalizado por alguma incompatibilidade do método desenvolvido com as outras bibliotecas utilizadas. Com isso foi optado a não finalização desta etapa para que fosse possível realizar a conclusão de outras etapas que ainda não estavam totalmente completas e que seriam fundamentais para a ferramenta.

Para projetos futuros seria interessante finalizar o processo de inserção dos dados, dando assim a possibilidade de na própria ferramenta o administrador possa selecionar as colunas definidas para a análise, tornando assim o processo de inserção dos dados ainda mais simples. E tornar possível que o usuário possa

realizar comparações de um ano para o outro, assim podendo selecionar colunas de dois anos e analisá-las através dos gráficos disponibilizados pela plataforma.

REFERÊNCIAS

ALMEIDA, Mauricio B.; BAX, Marcello P.; **Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção**, Brasília, Volume. 32, Número. 3, p. 7-20, dez./2003. Disponível em: <https://www.scielo.br/pdf/ci/v32n3/19019>. Acesso em: 8 jun. 2020. ANGULAR. **angular**. Disponível em: <https://angular.io/features>. Acesso em: 19 mai. 2020.

AVATRADE (ed.). **Como interpretar gráficos Bolsa**. Disponível em: <https://www.avatradeportuguese.com/education/trading-for-beginners/how-to-read-a-t-rading-chart>. Acesso em: 29 abr. 2020.

BERTOLINI, Rogério. **Estudo de caso sobre visualização de dados na área da saúde**. 2009. 85 f. TCC (Graduação) - Curso de Curso de Bacharelado em Ciência da Computação, Universidade de Caxias do Sul, Caxias do Sul, 2009. página. 14. Disponível em: <https://repositorio.uces.br/xmlui/bitstream/handle/11338/1269/TCC%20Rogério%20Bertolini.pdf?sequence=1&isAllowed=y>. Acesso em: 09 abr. 2020.

BRASIL. INEP. (org.). **Questionário do Estudante**. 2019. Disponível em: <http://portal.inep.gov.br/web/guest/questionario-do-estudante>. Acesso em: 28 mar. 2020.

CARTJS. **Chartjs**. Disponível em: <https://www.chartjs.org/> Acesso em: 19 jun. 2021.

CÔRTEZ, Sérgio da Costa *et al.* **Mineração de dados - funcionalidades, técnicas e abordagens**. Rio de Janeiro: Puc, 2002. 35 p. Disponível em: ftp://139.82.16.194/pub/docs/techreports/02_10_cortes.pdf. Acesso em: 16 jul. 2020.

CAMBOIM, Silvana Philippi; SLUTER, Cláudia Robbi. **Estudo sobre um algoritmo para a construção de mapas de pontos de contagem**. Boletim de Ciências Geodésicas, [s. l.], v. 19, ed. 1, p. 65-83, 2013. Disponível em: <https://www.scielo.br/pdf/bcg/v19n1/a05v19n1.pdf>. Acesso em: 18 jun. 2020.

CRETTON, Nícollas Nogueira; GOMES, Geórgia Regina Rodrigues. **Aplicação de técnicas de mineração de dados na base de dados do enade com enfoque nos cursos de medicina**. Acta Biomédica Brasiliensia, [S.l.], v. 7, ed. 1, p. 74-89, 20 jun. 2016. DOI <https://doi.org/10.18571/acbm.100>. Disponível em: <https://www.actabiomedica.com.br/index.php/acta/article/view/130>. Acesso em: 14 set. 2020.

DJANGO. **Django**. Disponível em: <https://www.djangoproject.com/start/overview/>. Acesso em: 19 mai. 2020.

FAYYAD, Usama *et al.* From Data Mining to Knowledge Discovery in Databases. **Ai Magazine**, Providence, v. 17, n. 3, p. 24-26, 18 jul. 2020.

GRÉGIO, André Ricardo Abed; FILHO, Benício Pereira de Carvalho; MONTES, Antônio; SANTOS, Rafael. **Técnicas de Visualização de Dados aplicadas à Segurança da Informação**. In: IX SIMPÓSIO BRASILEIRO EM SEGURANÇA DA INFORMAÇÃO E DE SISTEMAS COMPUTACIONAIS, 2009, Campinas - SP. Anais

[...]. Campinas - SP: [s. n.], 2009. p. 1-42. Disponível em: <http://www.lac.inpe.br/~rafael.santos/Docs/SBSEG/2009/sbseg2009.pdf>. Acesso em: 26 abr. 2020.

HARRIS, Charles R.; MILLMAN, K. Jarrod; WALT, Stéfan J. van Der; GOMMERS, Ralf; VIRTANEN, Pauli; COURNAPEAU, David; WIESER, Eric; TAYLOR, Julian; BERG, Sebastian; SMITH, Nathaniel J.. Array programming with NumPy. **Nature**, [S.L.], v. 585, n. 7825, p. 357-362, 16 set. 2020. Springer Science and Business Media LLC. <http://dx.doi.org/10.1038/s41586-020-2649-2>. Disponível em: <https://numpy.org/>. Acesso em: 04 mai. 2021.

INEP (comp.). **Questionário do Estudante**. 2019. Disponível em: <http://portal.inep.gov.br/web/guest/questionario-do-estudante>. Acesso em: 09 jul. 2020.

INFOGRAM (ed.). **O que é visualização de dados?** Disponível em: <https://infogram.com/pt/pagina/visualizacao-de-dados>. Acesso em: 01 mai. 2020.

LADEIRA, Gustavo de Almeida; COELHO, Frederico de Miranda. **Desenvolvimento de um Software Educacional de Apoio à Alfabetização de Crianças Especiais**. 2016. 13 f. Curso de Ciência da Computação, Departamento de Ciência da Computação, Universidade Presidente Antônio Carlos (unipac), Barbacena, 2016. página. 1. Disponível em: <https://www.unipac.br/site/bb/tcc/tcc-1c82bb878813f170519057581b7381d1.pdf>. Acesso em: 10 abr. 2020.

MARTINS, Idemara Marcell. **A mineração de dados para descoberta de conhecimento e uma oferta adequada no canal de televisão aberta**. 2010. 70 f. TCC (Graduação) - Curso de Universidade Federal do Paraná, Universidade Federal do Paraná, Curitiba, 2010. Disponível em: <https://acervodigital.ufpr.br/bitstream/handle/1884/48109/TCC%20%20Idemara%20Marcelli%20Martins.pdf?sequence=1>. Acesso em: 28 mar. 2020.

MICROSOFT EXCEL. **Excel**. Disponível em: <https://www.microsoft.com/pt-br/microsoft-365/excel>. Acesso em: 17 mai. 2020.

PYTHON. **O Tutorial Python**. Disponível em: <https://docs.python.org/pt-br/3/tutorial/index.html>. Acesso em: 18 mai. 2020.

PANDAS (comp.). **Visão geral**. 2020. Disponível em: https://pandas.pydata.org/docs/getting_started/overview.html. Acesso em: 18 set. 2020.

PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent. Scikit-learn: Machine Learning in Python. **Journal Of Machine Learning Research**. Mit Press e Microtome Publishing (Estados Unidos), p. 2825-2830. out. 11. Disponível em: <https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>. Acesso em: 04 mai. 2021.

QEDU (ed.). **O que são microdados?** Disponível em: <https://academia.qedu.org.br/glossario/o-que-sao-microdados/>. Acesso em: 28 abr. 2020.

R. **O Projeto R para Computação Estatística.** Disponível em: <https://www.r-project.org/>. Acesso em: 18 mai. 2020.

RAMOS, Isabel; SANTOS, Maribel Yasmina. **Data Mining no suporte à construção de Conhecimento Organizacional.** In: CONFERÊNCIA DA ASSOCIAÇÃO PORTUGUESA DE SISTEMAS DE INFORMAÇÃO, 4., 2003, Porto. CAPSI : actas da 4.^a conferência [...]. Porto: Associação Portuguesa de Sistemas de Informação (APSI), 2003. p. 1-15. Disponível em: https://repositorium.sdum.uminho.pt/bitstream/1822/2302/1/CAPSI2003_IMR_MYS.pdf. Acesso em: 8 jun. 2020.

RSTUDIO. **Recursos do RStudio IDE.** Disponível em: <https://rstudio.com/products/rstudio/features/>. Acesso em: 17 mai. 2020.

SQLite. **sqlite.** Disponível em: <https://www.sqlite.org/index.html>. Acesso em: 19 mai. 2020.

SANTANA, Rodrigo. **Validação Cruzada: Aprenda de forma simples como usar essa técnica.** 2020. Disponível em: <https://minerandodados.com.br/validacao-cruzada-aprenda-de-forma-simples-como-usar-essa-tecnica/>. Acesso em: 28 mai. 2021.

TABLEAU (ed.). **Guia prático da visualização de dados: definição, exemplos e recursos de aprendizado.** Disponível em: <https://www.tableau.com/pt-br/learn/articles/data-visualization>. Acesso em: 25 mar. 2020.

VAL, Ronaldo Borges do. **Visualização de dados aplicados em educação à distância no processo de avaliação ao aluno.** 2010. 94 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Universidade Federal de Pernambuco, Recife, 2010. Disponível em: https://repositorio.ufpe.br/bitstream/123456789/2449/1/arquivo3455_1.pdf. Acesso em: 08 jun. 2020.

VENTURA, Plínio. **Entendendo o Diagrama de Atividades da UML.** 2016. Disponível em: <https://www.ateomomento.com.br/uml-diagrama-de-atividades/>. Acesso em: 29 abr. 2020.

VILARINHO, Renato Avilez. **Uso de Técnicas de Mineração de Dados para Classificação das Ocorrências de Casos de Dengue nos Municípios Brasileiros.** 2017. 44 f. TCC (Graduação) - Curso de Sistemas de Informação, Universidade Federal de Ouro Preto, João Monlevade, 2017. Disponível em: https://www.monografias.ufop.br/bitstream/35400000/326/1/MONOGRRAFIA_UsoTecnicasMinera%C3%A7%C3%A3o.pdf. Acesso em: 28 mar. 2020.