Arguição sobre a quantidade de discentes com necessidade especiais no Ensino Superior, no entanto a permanência depara-se com falta de acessibilidade.

Renato Pedrosa Vasconcelos

Resumo

No nosso país há um evidente crescimento de pessoas com necessidade especiais no ensino superior, vindo com o amparo da legislação em vigor, porém podemos observar uma lacuna para este propósito. Apesar da lei em vigor determinar cota de vagas para pessoas com necessidade especiais, em torno de 2 a 5%, tanto nas empresas quanto nas Universidades ainda há muito a ser conquistado. No entanto, levando em conta algumas conquistas importantes, como maior conscientização da necessidade de inclusão social das pessoas com necessidade especiais. Mas ainda temos a impressão de que há muita para ser feito, fazendo-se necessários esforços conjugados no sentido de garantir-lhes que tenham condições de exercer sua cidadania plena. Como foi descrito na pesquisa de saúde (PNS) 2019, existia naquele ano, 17,3 milhões de pessoas com algum tipo de deficiência, pois isso corresponderia em torno de 8,4% da população. E que no ensino superior, segundo o Censo do ensino superior, representam apenas 0,41% dos discentes, ou seja, 66.750, alunos com deficiência(INEP,2019). Estudos mostram que há várias variáveis com relação a baixa adesão de pessoas com algum tipo de deficiência a permanecer nas escolas e consequentemente chegar ao ensino superior.
O presente trabalho tenta aplicar as técnicas de algoritmos de máquina no tema suscitado. Tais técnicas como fundamento teórico, bem como a aplicação do algoritmo de classificação, a Árvore de Decisão, O KNN (K-Nearest Neighbor) é um algoritmo que pode ser usado tanto para classificação como regressão, técnica de validação cruzada, Support Vector Machine (SVM) é um algoritmo supervisionado de aprendizado de máquina que pode ser utilizado para classificação ou regressão e o algoritmo k-means é o mais conhecido quando falamos de tarefas de agrupamento, para que se possa verificar similaridades nas características dos cursos onde há uma maior presenças de pessoas portadoras de deficiência. Objetivando o estudo, há ideia é oferecer uma ferramenta que possa auxiliar na gestão, facilitando a análise na tomada de decisão estratégica,como também a possibilidade de identificar os cursos e as Instituições de Ensino Superior(IES) que estão se destacando na questão do ensino e inclusão de pessoas com deficiência. Dessa forma entregando uma ferramenta para que possam traçar ações e elaborar políticas públicas com o intuito de melhorar o acesso dessa parte da sociedade dentro.

Palavras-chave: mineração de dados; acessibilidade; inclusão.

Abstract

In our country the growth of people with special needs in higher education supported by legislation in force, we can observe a gap for this purpose. Despite the law in force that determines the quota of vacancies for people with special needs is around 2 to 5%, both in companies and universities. Taking into account some important achievements, as a greater

awareness of the need for social inclusion of people with special needs. But we still have the impression that there is much to be done, conjugated efforts to ensure that they are able to exercise their full citizenship. As described in health survey (PNS) 2019, there were 17.3 million people with some type of disability, as this would correspond to around 8.4% of the population. And that in higher education, according to the higher education census, they represent only 0.41% of the students, that is, 66,750, disabled students (INEP, 2019). Studies show that there are several variables regarding low adherence of people with some kind of disability to remain in schools and consequently reach higher education. The present work reports some concepts of technical methodologies and data mining algorithms as theoretical foundation, as well as the application of the classification algorithm, the decision tree, KNN (K-Nearest Neighbor) is an algorithm that can be used For classification like regression, cross-validation technique, Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification or regression and K-Means algorithm is the best known when we speak of grouping tasks, for That can be seen similarities in the characteristics of the courses where there are greater presence of people with disabilities. The objective of this study is to offer a tool that can help in management, facilitating the analysis in strategic decision making, but also the possibility of identifying courses and IES that are standing out in the question of teaching and inclusion of people with disabilities. In this way delivering a tool so that they can draw stocks and elaborate public policies in order to improve access to that part of society within.

Keywords: Data mining; accessibility; Inclusion.

1. Introdução

No presente, devido às mudanças na legislação educacional brasileira, o aumento do número de alunos com deficiência no ensino superior é cada vez mais relevante. No entanto, o pleno acesso e a permanência das pessoas com deficiência no Ensino Superior ainda não é uma realidade.

Mas a isso nós cremos que a representatividade de pessoas com deficiência no ensino superior é uma das consequências da elaboração de políticas públicas que tem o propósito de mediar melhorar a equidade da educação superior.Existem casos de estudos relativos às estratégias de inclusão do alunos com deficiência, tudo isso nos leva a perceber que as questões relativas à democratização de oportunidades, superação das desigualdades, e o reconhecimentos dos diferentes grupos que se fazem cada vez mais presentes. Hoje, a “diferença” tem caído cada vez mais, porque se faz destaque nos debates cotidianos de todo mundo. Até dentro das próprias universidades com apoio a projetos direcionados às pessoas com necessidade especiais com o intuito de colaborar para a permanência do aluno dentro da universidade.

Uma das principais Metas para o desenvolvimento sustentável é a Educação de Qualidade, que tem como objetivo assegurar a educação inclusiva, de qualidade, além da promoção de oportunidades. Isso trará garantias da igualdade de acesso a todos os níveis de educação dos mais vulneráveis, o que inclui as pessoas com deficiência.

Portanto, o Censo do Ensino Superior potencializa a sua importância nos projetos, porque além de ser a sobretudo a fonte de dados para o trabalho em questão é a maior modelo de pesquisa acerca das instituições do ensino superior no Brasil, porque além de dedicar dados estatísticos genuíno, contém inúmeros especificidades a respeito da infraestrutura das instituições de ensino Superior, como por exemplo: os números de vagas ofertadas em cada curso, os candidatos, à matrícula, concluintes e docentes, permitindo realizar análises sobre o contexto estabelecido.

Compreender as necessidades de pessoas com deficiência pode ser apregoado um meio não apenas de inclusão, como uma maneira de expressar o avanço de uma sociedade a respeito dos direitos das pessoas com necessidades especiais. A compreensão sobre tal assunto precisa ir além dos indagação mais básicas, aprofundar os conhecimentos dos impedimento propostos a essas pessoas, com suas necessidades e vontades, que em a maior parte são desconsideradas não apenas através de gestos de ação de públicas exclusivas, mas também de uma resiliência da sociedade em integrá-los como membros da sociedade organizada.

A OMS(Organização Mundial da Saúde) criou um acordo para trabalhar pela diversidade e equidade por meio de iniciativas globais. Essa ação vai ajudar muito a sociedade e a OMS estima que cerca de 5% da população mundial viva com deficiências.

A pesquisa detalha que 7,8 milhões, ou 3,8% da população acima de dois anos, apresentam deficiência física nos membros inferiores, enquanto 2,7% das pessoas têm nos membros superiores. Já 3,4% dos brasileiros possuem deficiência visual; e 1,1%, deficiência auditiva. Já 1,2% – ou 2,5 milhões de brasileiros – têm deficiência intelectual.

Entre a população com algum tipo de deficiência, 10,5 milhões são mulheres (9,9%), frente a 6,7 milhões de homens (6,9%). Em relação ao local onde moram, 9,7% das pessoas estão em áreas rurais, enquanto 8,2% em zonas urbanas.

O estudo ainda detalha a proporção de pessoas com alguma deficiência entre as etnias: 9,7% eram pretas, 8,5% pardas e 8% brancas.

O levantamento do IBGE aponta que a inclusão da pessoa com deficiência no mercado de trabalho ainda é um obstáculo. Apenas 28,3% delas em idade de trabalhar (14 anos ou mais de idade) se posicionam na força de trabalho brasileira. Entre as pessoas sem deficiência, o índice sobe para 66,3%, segundo o IBGE. Por fim, a ideia dessa pesquisa é apresentar uma pequena contribuição num assunto tão complexo, mas a sociedade não civil organizada não poderá deixar em segundo plano.

2. Objetivos

O presente trabalho busca apresentar conceitos de estatísticas, além de ferramentas e modelos de aprendizagem de máquina para ajudar na manipulação das análises dos

dados do Censo do Ensino Superior do Brasil, na quadra dos anos 2015 a 2019,anos que começaram a registrar discentes com deficiência das condições relativas à acessibilidade disponibilizadas pelos cursos.

2.1 Objetivo principal

Objetivo é criar um panorama para uma análise que faça consonância a inclusão de discentes com deficiência no ensino superior do Brasil encarada em sua amplitude, em face da acessibilidade oferecidas pelos cursos, observando-se padrões para estimar uma previsão de discentes matriculados nos próximos anos, de maneira a auxiliar tomadas de decisões sobre a convenção sobre os direitos das pessoas as próximas gerações dessa parcela de discentes.

2.2 Objetivos secundários

Analisar o número de estudantes com algum tipo de deficiência que concluíram o ensino superior e observar se ocorre uma tendência a continuarem para o ensino superior. Verificar se esses estudantes estão matriculados em cursos que oferecem condições de infraestrutura e acessibilidade, assim como buscar a existência de uma relação entre cursos e universidades que possuem uma maior acessibilidade e a taxa de alunos concluintes.

Verificar se o modelo de regressão linear simples e múltipla, poderá estimar a conjuntura dos discentes com deficiência no ensino superior em datas pré estudadas por órgãos governamentais, para avaliar segundo os Objetivos de Desenvolvimento Sustentável da ONU, se a meta poderá ser cumprida no que se refere a educação de qualidade no quesito de equidade de acesso para pessoas com deficiência ao ensino superior.

Fazer o uso de modelo não supervisionado de clustering, para agrupar as Instituições de Ensino Superior(IES) diante de similaridades entre elas, no que diz respeito a supostos parâmetros ligado à quantidade de alunos com deficiência matriculados no ensino superior e alguns tipos de recursos de acessibilidades ofertados.

Diante do modelo supervisionado de classificação, tentarei descobrir qual a principal importância de uma Instituições de Ensino Superior(IES) visando o conceito da acessibilidade de pessoas com deficiência de maneira a identificar se a quantidade de recursos (físicos, digitais e humanos) na área administrativa (do municipal, estadual, federal) levando em conta atuação dessas esferas quem é mais assertiva em partes maiores na adesão de inclusão dos alunos com deficiência.

3. Trabalhos Correlatos

3.1 Organização do acesso e permanência das pessoas com deficiência no ensino superior a partir da instauração do programa incluir.

O Artigo de Lara e Sebastián-heredero(2020) explica como foi a implementação do programa incluir diante das políticas institucionais e as ações de acessibilidade para pessoas com necessidade especiais no ensino superior nos estados brasileiros. Junto com o apoio do MEC nas Instituições de Ensino Superior(IES) do Brasil, as universidades por

iniciativas próprias criaram os núcleos de acessibilidade nas Instituições de Ensino Superior(IES) e Instituições Estaduais Superior. Assim, o objetivo deste trabalho é analisar como está a situação atual a partir das produções de teses e dissertações, desde sua publicação, para compreender como as Instituições de Ensino Superior estão se organizando para apoiar o ingresso e a permanência de pessoas com deficiência, após esse aporte financeiro e as publicações das políticas públicas de inclusão no Brasil.

3.2 Ingresso e permanência de alunos com deficiência em universidades públicas brasileiras

O artigo de Castro e Almeida (2014) teve o objetivo de identificar ações e iniciativas relacionadas ao acesso de pessoas com deficiência nas universidades brasileiras e apresentou como resultados as condições de permanência dos alunos com deficiência no ensino superior e as barreiras por eles enfrentadas.

3.3. Inclusão de estudantes com deficiência no ensino superior.

Neste trabalho, Pereira et al (2016) destacam alguns marcos importantes relacionados à inclusão na educação e apresenta uma análise acerca da produção científica relacionada ao assunto. Utilizando técnicas de análise e visualização dos dados em grafos e aplicando um Teste de Relevância nas variáveis encontradas, os autores analisam correlações positivas e negativas a fim de apresentar uma revisão sistemática que possa contribuir na construção de conhecimento de alunos com deficiência no ensino superior.

3.4. Os serviço de deficientes visuais em instituições de ensino superior

O artigo visa identificar as dificuldades enfrentadas por deficientes visuais nas Instituições de Ensino Superior no acesso a serviços básicos, como retirada de documentos e obtenção de informações relacionadas ao seu curso quando precisam lidar diretamente com funcionários da instituição. De acordo com Orsini, Coelho e Abreu, essa dificuldade se faz presente antes mesmo do início das aulas, já na realização da matrícula, sendo ampliada nos demais encontros de serviço. A pesquisa identifica como problemas primários, o preconceito e até ações ilegais, como cobranças superiores de mensalidade em instituições privadas, gerando uma percepção desfavorável de uma IES pelos deficientes visuais.

3.5. Deficientes auditivos e escolaridade: fatores diferenciais que possibilitam o acesso ao ensino superior

Os autores Manente, Rodrigues e Palamin destacam a escassez de informações e estudos sobre a vivência do deficiente auditivo no ensino superior. O trabalho tenta compreender quais são os fatores que podem facilitar ou dificultar o ingresso desses indivíduos em Instituições de Ensino Superior. O estudo pontua os fatores para o ingresso, permanência e também desistência de deficientes auditivos em cursos superiores. Procurando buscar os motivos que levam a desistência antes do vestibular, como o medo do preconceito, dificuldade financeira e falta de acessibilidade para acompanhar as aulas.

4. Referencial Teórico

Como é fundamental o entendimento de um todo da pesquisa, o referencial teórico buscou-se por meio do estudo uma explicação a respeito do Censo do Ensino Superior, quais os recursos que ele apresenta para acessibilidade. Na tentativa de melhora compreensão dos dados, consequentemente o uso da técnica de aprendizado de máquina que venha a ser usada na pesquisa, são relacionados alguns conceitos sobre modelos e algoritmos compreendidos.

4.1 Deficiências e recursos de acessibilidade

Desde que o mundo existe, existe um processo natural de exclusão de pessoas com algum tipo de deficiência, mas todos são iguais perante a lei. Contra essa afirmação não há questionamentos, porém quando o Estado simplesmente não apresenta condições de acessibilidade àqueles que necessitam, instaura-se uma situação. Segundo a Convenção Internacional sobre os Direitos das Pessoas com Deficiência, pessoas com deficiência são aquelas que têm impedimentos de longo prazo de natureza física, mental, intelectual ou sensorial, os quais, em interação com diversas barreiras, podem obstruir sua participação plena e efetiva na sociedade em igualdades de condições com as demais pessoas(BRASIL, 2019).

Tipos de deficiência mapeadas no Censo do Ensino Superior
Descrição de algumas

Deficiência auditiva : Trata-se da perda bilateral, parcial ou total, na percepção normal dos sons. A perda auditiva pode variar de leve, que ocorre quando o indivíduo ouve com dificuldades, a profunda, que é a ausência total da audição

Deficiência física : Considera-se a alteração total ou parcial de um ou mais segmentos do corpo humano e que acarreta comprometimento da função física

Múltipla: Consiste na associação de dois ou mais tipos de deficiência (intelectual/visual/auditiva/física)

Surdez: Não considera apenas a medição de perda auditiva, como também, a adesão à comunidade surda, à experiência da visualidade e ao uso da língua de sinais

Surdocegueira : Trata-se de deficiência única, caracterizada pela deficiência auditiva e visual concomitante.

Baixa Visão : É a perda parcial, com visão reduzida em ambos os olhos.

Cegueira : É a perda total da visão em ambos os olhos.

Superdotação : Aqueles que demonstram potencial elevado em qualquer uma das seguintes áreas, isoladas ou combinadas: intelectual, acadêmica, liderança psicomotricidade e artes, além de apresentar grande criatividade, envolvimento na aprendizagem e realização de tarefas em áreas de seu interesse

TGD - Autismo : Transtorno onde há déficit em três domínios: déficit na sociabilidade; empatia e capacidade de compreensão ou percepção dos sentimentos do outro; déficit na linguagem comunicativa e imaginação; e déficit no comportamento e flexibilidade cognitiva

TGD – Síndrome de Asperger : Síndrome relacionada ao autismo, diferenciando-se deste por apresentar alterações formais na linguagem e na interação social

Cada ser humano possui a sua característica e não é diferente em relação à acessibilidade. O INEP (2016), com o caráter orientador de políticas públicas, não poderia deixar de colocar estratégias que venham propiciar a promoção da acessibilidade como se adaptar ao ambiente físico, fazendo uso de recursos de tecnologias assistiva.

Foram mapeados recursos no Censo do Ensino Superior

Recursos de comunicação de acessibilidade à comunicação
Recursos *de informática recursos de informática acessível*

Intérpretes guia-intérprete *responsável pela comunicação e visão entre a pessoa com surdocegueira e o meio no qual ela está interagindo (Almeida, 2017)*

intérprete de libras *Informa se o curso disponibiliza tradutor e intérprete de língua brasileira de sinais*

Disciplina inclusiva libras *Informa se o curso oferece disciplina de língua brasileira de sinais (Libras)*

Material Digital *Material didático digital acessível*

Materia*l ampliado material em formato impresso em caractere ampliado*

Material *tátil material pedagógico tátil*

Material *impresso material didático em formato impresso acessível*

Material *áudio material em áudio*

Material *braille material em Braille*

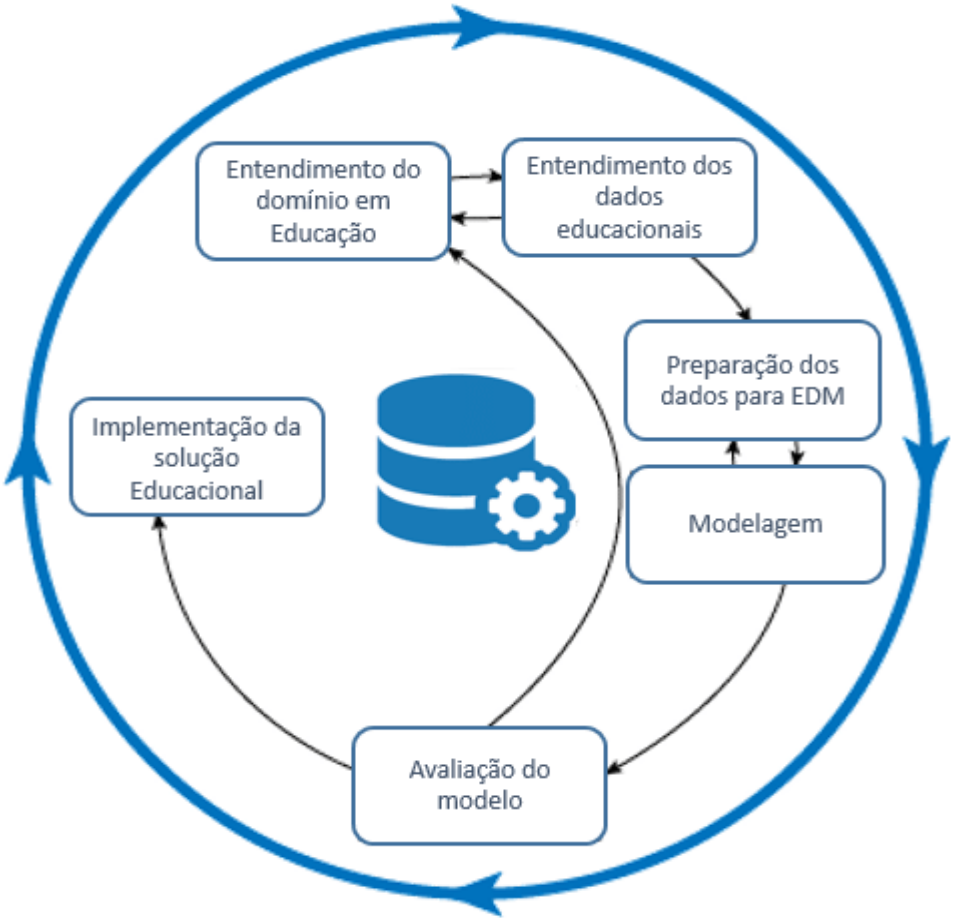
Material l*ibras material didático em língua brasileira de sinais*

O Brasil tem um guia de orientações para pessoas com necessidades especiais, o qual salienta, " pessoa com deficiência tem o direito de ser igual sempre que as diferenças a inferiorizem; e o direito de ser diferente sempre que a igualdade a descaracterize." (Brasil, 2005), Interpretação da frase de Boaventura de Souza Santos.

4.2 CRISP-DM

CRISP-DM, significa Cross Industry Standard Process for Data Mining, que poderá ser colocado como Processo Padrão Inter-Indústrias para Mineração de Dados. É um modelo de processo de mineração de dados que representa o método utilizado pelo profissional de mineração de dados para executar o processo MD. O CRIP-DM atende uma maneira estruturada para processos de mineração de dados, sendo bastante utilizada devido à sua característica flexível. Um dos conhecimentos mais relevantes obtidos a partir da visualização de dados é o CRISP-DM, é um ciclo de mineração de dados brutos que poderá ser aplicado em qualquer tipo de negócio. Por motivo da sua generalização, mesmo sendo um método criado pela indústria, a adoção da metodologia é sugerida como boa prática em EDM no trabalho de Sheth e Patel (2010). Este ciclo de vida dos dados é constituído por seis etapas principais: entendimento do domínio em educação, entendimento dos dados educacionais, preparação dos dados para EDM, modelagem, avaliação do modelo e implantação da solução educacional.

Figura 1: Etapas de um processo de CRISP-DM



Fases do CRISP-DM. Fonte: [Shearer, 2000]

A Análise Exploratória de Dados é uma etapa de pré-processamento dos dados, onde podemos visualizar algumas correlações entre dados, assim como comportamentos entre as variáveis que possuímos em nossos dados.

4.3. Técnicas aprendizado de máquina

*A*prendizado em máquina, é uma área de Inteligência Artificial, capaz de identificar padrões, tomar decisões, modificar seu comportamento, de forma autônoma,com base na análise de dados e de suas experiências, com o mínimo de interferência humana e, o desempenho da inteligência artificial está inteiramente condicionado à inteligência humana de quem a treinou, no qual é possível fazer, a partir de uma análise de um conjunto de dados, gerando uma competitividade de um modelo analítico, que consiga reconhecer padrões e realizar predições. A Pesquisa busca justificar indagações relacionadas a análises relativas à representatividade de pessoas com deficiência no ensino superior de forma a permitir que gestores realizem essas análises para tomadas de decisão que decorrem em programas e ações que requerem maior inclusão dessas pessoas na sociedade. Por meio de técnicas de *a*prendizado de máquina, podem ser avaliados padrões e produzir predições para esclarecer problemas como os referenciados em nossas análises. Os modelos de Aprendizado de Máquina podem ser supervisionados, não supervisionados e de aprendizagem. Os modelos supervisionados envolvem modelagem de relações entre formas medidos em um tanto de legendas associado a esses dados, na medida em que este modelo é determinado, ele pode ser usado para aplicar rótulos nos supostos novos dados desconhecidos. Mencionarei alguns tipos básicos de modelos supervisionados que na literatura dizem ser de regressão e classificação, o primeiro fala-se do método de predição utilizado para prever valores contínuos, que podem ser: regressão linear simples, regressão linear múltipla, segundo fala-se do método de predição utilizado para prever valores com duas ou mais categorias discretas.

A regressão linear simples é um modelo mais simples para se prever a relação entre duas ou outras variáveis, simples e fácil de trabalhar, é um dos algoritmos supervisionados de aprendizagem de máquina que aproveita para realizar análises preditivas e será o modelo a ser utilizado no trabalho. A regressão linear simples baseia-se em uma correlação entre duas variáveis, que reflete a que cada valor de uma variável dependente “Y” varia conforme com uma variável independente “X”. Para interpretar a correlação existente usamos Pearson, significa que o coeficiente de correlação e o valor grega rho (ρ) que segundo Frost (2019) é um parâmetro de uma estabelece uma amostra, que mede a força e a direção da relação linear entre duas variáveis contínuas, com valores que podem variar de -1 a 1.
Com o intuito de achar as melhores estimativas para os coeficientes, que minimizam os erros na previsão de “Y” a partir de “X”, é colocada uma equação, que deve ser valorizado ou desconsiderado e por essa razão, é necessário medir o erro padrão sobre uma linha de regressão, por meio de uma análise do coeficiente de determinação “R - quadrado”, que varia de 0 a 1. A ideia é que quanto mais próximo de 1, melhor o ajuste e menor o erro padrão, ou seja, conseguindo assim uma maior probabilidade de precisão das previsões.

A equação utilizada é " $y = b0 + b1 * x + e$”, onde, segundo Bomfim (2018):

- ➔ y: é a variável dependente, ou seja, o valor previsto;
- ➔ x: é a variável independente, ou seja, a variável preditora;
- ➔ b0: é o coeficiente que intercepta ou que corta o eixo y;
- ➔ b1: é o coeficiente que define a inclinação da reta; e
- ➔ e: é o erro padrão

KNN(K Nearest Neighbors) é um dos muitos algoritmos (de aprendizagem supervisionada) usado no campo da mineração de dados e modelos de aprendizado de máquina, ele é um classificador onde o aprendizado é baseado nos vizinhos mais próximos baseiam-se na aprendizagem por analogia, isto é, comparando uma determinada tupla de teste com torres de treinamento semelhantes a ele,comparando uma determinada tupla de teste com torres de treinamento semelhantes a ele. As tuplas de treinamento são descritas por n atributos. Cada tupla representa um ponto em um espaço n-dimensional. Desta forma, todas as tuplas de treinamento são armazenadas em um espaço n-dimensional. Quando dada uma tupla desconhecida, um classificador K-mais próximo do vizinho busca o espaço de padrão para os tubulares de treinamento K mais próximos da tupla desconhecida. Estas torres de treinamento K são os K "vizinhos mais próximos" da tupla desconhecida. "Proximidade" é definido em termos de uma métrica de distância, como a distância euclidiana. A distância euclidiana entre dois pontos ou tuplas, digamos,

$$\boldsymbol{X}_1 = (x_{11}, x_{12}, \ldots, x_{1n}) \text{ and } \boldsymbol{X}_2 = \qquad (x_{21}, x_{22}, \ldots, x_{2n}),$$

e

$$dist(\boldsymbol{X}_1, \boldsymbol{X}_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

Em outras palavras, para cada atributo numérico, assumimos a diferença entre os valores correspondentes desse atributo na tupla X1 e na tupla x2, quadrado dessa diferença e acumulá-la. A raiz quadrada é tirada da contagem total de distância acumulada. Normalmente, normalizamos os valores de cada atributo antes de usar o EQ. (9.22). Isso ajuda a impedir que os atributos inicialmente grandes faixas (por exemplo, renda) de atributos superando com intervalos inicialmente menores (por exemplo, atributos binários). Normalização mínima, por exemplo, pode ser usado para transformar um valor V de um atributo numérico A a V 0 no intervalo [0, 1] por computação. O treinamento é formado por tuplas de várias dimensões. Os algoritmos de classificação na aprendizagem de máquina utilizam dados de entrada para o treinamento de predição da possibilidade que dados subsequentes sejam alocados em categorias predeterminadas. Um exemplo comum desse uso é a classificação de correios eletrônicos entre duas subpopulações; aqueles que são “mensagens eletrônicas” e aqueles que não são. A quantidade de subpopulações não tem um limite, e pode ser dita como número para um conjunto de dados possível. Deduzimos que a classificação nada mais é que um reconhecimento de padrões, e a sua aplicação planeja a busca de similares entre dados quaisquer, como termos amizade, continuação de números dentre outros objetos de interesse. Vários dos algoritmos são usados na classificação, mais árvores de decisão, KNN e SVM (Support Vector Machine) dominam.

Clustering com o algoritmo K-means O algoritmo K-means é um método de cluster que é popular devido à sua velocidade e escalabilidade. K-meios é um processo iterativo de mover os centros dos grupos, ou os centro, para a posição média de seus pontos constituintes, e atribuir instâncias aos seus clusters mais próximos. O K Titular é um hiperparâmetro que especifica o número de clusters que devem ser criados; K-significa atribui automaticamente observações a grupos, mas não pode determinar o número apropriado de grupos. K deve ser um inteiro positivo que seja menor que o número de

casos no conjunto de treinamento. Às vezes, o número de agrupamentos é especificado pelo contexto do problema de agrupar. Por exemplo, uma empresa que fabrica sapatos pode saber que é capaz de suportar a fabricação de três novos modelos. Para entender quais grupos de clientes para atingir cada modelo, pesquisam os clientes e cria três clusters dos resultados. Ou seja, o valor de K foi especificado pelo contexto do problema. Outros problemas podem não exigir um número específico de clusters, e o número ideal de clusters pode ser ambíguo. Vamos discutir uma heurística para estimar o número ideal de clusters chamado o método de cotovelo mais adiante neste capítulo. Os parâmetros de K-meios são as posições dos centróides dos clusters e as observações atribuídas a cada cluster. Como modelos lineares generalizados e árvores de decisão, os valores ideais dos parâmetros K-meios são encontrados minimizando uma função de custo. A função de custo para k-meios é dada pela seguinte equação:

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} \left\| x_i - \mu_k \right\|^2$$

Na equação precedente, “μ K” é o centróide para o agrupamento k. A função de custo soma as distorções dos agrupamentos. A distorção de cada grupo é igual à soma das distâncias quadradas entre seu centróide e suas instâncias constituintes. A distorção é pequena para clusters compactos e grandes para clusters que contêm instâncias dispersas. Os parâmetros que minimizam a função de custo são aprendidos através de um processo iterativo de atribuir observações a clusters e depois movendo os clusters. Primeiro, os centróides dos clusters são inicializados para posições aleatórias. Na prática, estabelecendo as posições iguais às posições de observações selecionadas aleatoriamente produz os melhores resultados. Durante cada iteração, K-meios atribui observações ao agrupamento que eles estão mais próximos e, em seguida, move os centróides para sua visibilidade designada " .

Suponha que K-significa inicializar o centro para o primeiro grupo à quinta instância e o centro para o segundo grupo na décima primeira instância. Para cada instância, vamos calcular sua distância para os centróides e atribuí-lo ao grupo com o centro mais próximo. As atribuições iniciais são mostradas na coluna de grupo da tabela, a saber:

Instance	X0	X1	C1 distance	C2 distance	Last cluster	Cluster	Changed?
1	7	5	3.16228	2	None	C2	Yes
2	5	7	1.41421	2	None	C1	Yes
3	7	7	3.16228	2.82843	None	C2	Yes
4	3	3	3.16228	2.82843	None	C2	Yes
5	4	6	0	1.41421	None	C1	Yes
6	1	4	3.60555	4.12311	None	C1	Yes
7	0	0	7.21110	7.07107	None	C2	Yes
8	2	2	4.47214	4.24264	None	C2	Yes
9	8	7	4.12311	3.60555	None	C2	Yes
10	6	8	2.82843	3.16228	None	C1	Yes
11	5	5	1.41421	0	None	C2	Yes
12	3	7	1.41421	2.82843	None	C1	Yes
C1 centroid	4	6					
C2 centroid	5	5					

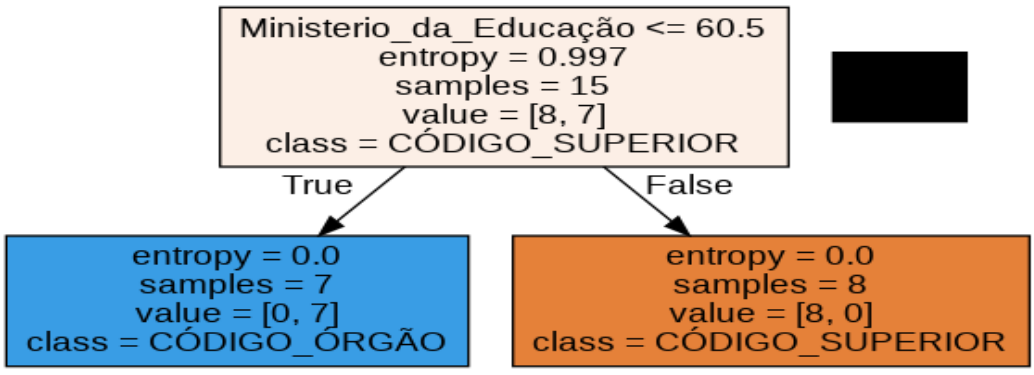
A árvore da decisão é um modelo simples e não linear para tarefas de classificação como o nome já diz, é uma estrutura semelhante a uma árvore que é usada para tomar decisões que são usadas principalmente em declarações de controle condicional. O principal uso do algoritmo se enfatiza em resolver problemas de classificação, utilizando técnicas como Gini e Qui-quadrado para a categorização. Para se decidir como a classificação ocorrerá, deve-se decidir quais variáveis deverão ser analisadas, quanto mais variáveis são colocadas nas regras do algoritmo, mais específica a classificação se torna. Por exemplo, decidir qual tipo de recomendação de um usuário de um produtos nas redes sociais que mais gosta, pode se basear em diversos aspectos que são considerados relevantes para a análise, como a sua idade, seu gênero, o gênero de filme que este usuário mais costuma aceitar ou o histórico de avaliações positivas e negativas em produtos e imagens anteriores. É essencial que as variáveis de uma árvore de decisão sejam independentes, isto é, seja possível analisá-las isoladamente e não apenas em conjunto com outras.

No trabalho, tentarei utilizar a técnica de classificação a fim de buscar similaridades nas Instituições de Ensino Superior(IES) públicas em relação a acessibilidade existente para pessoas com deficiência. O propósito principal é classificar as instituições de acordo com as receitas existentes, além de buscar quais delas possuem mais discentes com deficiência, e mais discentes desse pedaço que concluíram seus referentes cursos.

Segundo VanderPlas (2017), as técnicas de modelagem não supervisionados envolve a modelagem de recursos de um conjunto de dados, em que não há referência a rótulos,“o conjunto de dados falam por si mesmo”. Esses modelos incluem tarefas como redução de dimensionalidade e agrupamentos. Algoritmos de agrupamento identificam grupos distintos de dados, enquanto algoritmos de redução de dimensionalidade buscam representações mais sucintas dos dados

Os algoritmos mais utilizados no processo de modelagem dos grupos são o k-means, o DBSCAN e o Mean Shift. Esses são os três algoritmos mais comuns usados como agrupamento, mas possuem facilidades e algumas limitações, sendo necessário fazer uma análise antes de decidir sobre qual usar.

```
dot_data = tree.export_graphviz(clf, feature_names=X.columns, class_names=[
graph = pydotplus.graph_from_dot_data(dot_data)
Image(graph.create_png())
```



O k-Means baseia-se na realização de médias para definição dos centros de n grupos e nele devemos usar parâmetros de entrada. No algoritmo k-means especificamos um número de cluster que estimamos ter, atribuímos os pontos de dados a um cluster aleatoriamente, determinamos os centróides do cluster, em seguida, atribuímos novamente cada ponto ao centróide do cluster mais próximo e reinicializamos os centro do grupo. Repetimos as últimas duas etapas até que não haja melhorias possíveis. Nesse momento devemos alcançar os pontos ótimos de agrupamento.

O Density-Based Spatial Clustering of Applications with Noise (DBSCAN), ou Agrupamento espacial baseado em densidade para aplicações com ruído, é um algoritmo que baseia-se em regiões de alta densidade. Os parâmetros de entrada são o “epsilon”, distância máxima entre dois pontos para que eles possam ser considerados vizinhos (ou pertencentes ao mesmo grupo) e o “min_samples”, número mínimo de pontos para que eu possa formar um cluster. Além disso, é necessário passar uma métrica de distância, como forma de avaliar a distância entre os pontos. No DBSCAN não é necessário inserir o número de cluster, o algoritmo automaticamente define quantos agrupamentos são necessários de acordo com os parâmetros de entrada.

O k-means é um algoritmo poderoso de agrupamento comumente usado em técnicas de agrupamento e, é competente para processar conjuntos de dados grandes, suscetível para outliers e ,é também normalmente adequado para conjunto de dados numéricos.Mas, seu método deixa a desejar não funciona bem em relação a dados que contêm variáveis categóricas. Essa falta de eficiência decorre quando a função de custo em K-means é calculada diante da distância euclidiana que só é adequada para dados numéricos.

Para clustering baseado na existência de análise com dados categóricos, Huang (1998) propôs um algoritmo chamado k-Mode que é criado para lidar com algoritmos de agrupamento com este tipo de dado. O algoritmo de k-mode estende as médias de k usando uma dissimilaridade de correspondência em função de dados categóricos, um método baseado em frequência para atualizar o processo de agrupamento que reduz a função de custo.

Diante dos problemas e das dificuldades em trabalhar o clustering com uma base de dados mista, Huang propôs ainda um algoritmo denominado k-Prototype que é criado para lidar com algoritmos de agrupamento com os tipos de dados mistos e um método de agrupamento baseado em particionamento. Seu algoritmo é uma melhoria do algoritmo de clustering k-Means e k-Mode para lidar com o clustering com os tipos de dados mistos.

Por visualização, **k-prototypes fornece clusters mais distinguíveis**. K-Means é geralmente referenciado por clusters 4-5, enquanto os clusters K-Prototypes são mais igualmente distribuídos com limites claros. Desse modo, K-Prototypes é bem menos complexo e o k-prototypes, é um dos algoritmos indicados para clustering que possuem base de dados mistos (categóricos e numéricos).

Há possibilidade também de fazer o uso de mais de um deles e analisar quão bom é resultado, por meio de uma avaliação do coeficiente de silhueta, por exemplo.

Além dos modelos supervisionados e não supervisionados, o Machine Learning aborda a aprendizagem por reforço.

“ O Aprendizado por Reforço (ou Reinforcement Learning – RL), conhecido como modelo de aprendizado semi-supervisionado em Aprendizado de máquina, é uma técnica para permitir que um agente tome ações e interaja com um ambiente, a fim de maximizar as recompensas totais. Aprendizado por Reforço é geralmente modelado como um Processo de Decisão de Markov (MDP)”(Data Science Academy. Deep Learning Book, 2022.)

Assim como os outros modelos, o aprendizado por reforço também é baseado em feedbacks fornecidos pelo ambiente, no entanto, neste caso, essa informação é mais qualitativa, recebendo “recompensa” ou “penalidade” pelas ações executadas nos algoritmos, para entendimento se uma determinada ação em um estado foi positiva ou negativa, respectivamente.

O aprendizagem por reforço é particularmente eficiente quando o ambiente não está completamente determinístico, muitas vezes quando é muito dinâmico e quando é impossível medir um erro de modo preciso.Segundo Bonaccorso (2017)

Na discretização *de dados transformamos dados numéricos mapeando valores em rótulos de conceito de intervalo. Tais métodos podem ser usados para gerar automaticamente hierarquias de conceito para os dados, o que permite a formação em múltiplos níveis de granularidade. As técnicas de discretização incluem binning, análise de histograma, análise de cluster, análise de árvores de decisão e análise de correlação. Para hierarquizadas nominais de conceito de dados podem ser geradas com base em definições de esquema, bem como no número de valores distintos por atributo. Podemos pensar na discretização como a redução do número de valores para determinados atributos contínuos.*

Outliers A ideia em primeiro lugar, conseguimos ter que entender o outlier, um outlier é uma análise nos dados ou instância que está distante ou longe das outras observações de dados ou você pode dizer uma observação de dados que não representa o comportamento ou padrões semelhantes em relação ao resto do conjunto de dados que é chamado de outliers.
Há diferentes tipos de outliers como:
Outliers univariados: Há um valor extremo em uma única variável ou atributo entre os diferentes atributos ou variáveis.
Outliers multivariados: Existem múltiplos valores extremos ou incomuns em pelo menos dois ou mais dois atributos ou variáveis.
Coleta de dados e outliers:
Como sabemos, o primeiro passo de qualquer projeto de aprendizado de máquina ou data science é a coleta dos dados e este é o ponto onde os outliers se incorporam em nosso conjunto de dados. Vamos discutir como os outliers se incorporam ao conjunto de dados.

O **PCA** reduz as dimensões de um conjunto de dados projetando os dados em um subespaço dimensional. Por exemplo, um conjunto de dados dois dimensionais pode ser reduzido projetando os pontos em uma linha; Cada instância no conjunto de dados seria então representada por um único valor em vez de um par de valores. Um conjunto de dados tridimensional pode ser reduzido a duas dimensões projetando as variáveis em um plano.

Em geral, um conjunto de dados N-dimensional pode ser reduzido projetando o conjunto de dados em um subespaço k-dimensional, onde K é menor que n. Mais formalmente, o PCA pode ser usado para encontrar um conjunto de vetores que abrangem um subespaço, o que minimiza a soma dos erros quadrados dos dados projetados. Esta projeção reterá a maior proporção da variância do conjunto de dados original.

Realizando análise de componentes principais Existem vários termos que devemos definir antes de discutir como funciona a análise de componentes principais. Variância, covariância e covariância Matrizes.

Variância é calculada como a média das diferenças quadradas dos valores e média dos valores, conforme a seguinte equação.

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$$

A covariância é uma medida de quanto duas variáveis mudam juntas; É uma medida da força da correlação entre dois conjuntos de variáveis. Se a covariância de duas variáveis for zero, as variáveis estão não correlacionadas. Note que as variáveis não correlacionadas não são necessariamente independentes, já que a correlação é apenas uma medida de dependência linear. A covariância de duas variáveis é calculada usando a seguinte equação:

$$\operatorname{cov}\left(X,Y\right) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{y}\right)}{n-1}$$

Se a covariância é diferente de zero, o sinal indica se as variáveis são positivas ou negativamente correlacionadas. Quando duas variáveis estão positivamente correlacionadas, uma aumenta à medida que os outros aumentam. Quando as variáveis são negativamente correlacionadas, uma variável diminui em relação à sua média à medida que a outra variável aumenta em relação à sua média. Uma matriz de covariância descreve os valores da covariância entre cada par de dimensões em um conjunto de dados. O elemento (I, J) indica a covariância das dimensões IH e JTH dos dados. Por exemplo, uma matriz de covariância para um dado tridimensional é dada pela seguinte matriz:

$$C = \begin{bmatrix} \operatorname{cov}\left(x_1,x_1\right) & \operatorname{cov}\left(x_1,x_2\right) & \operatorname{cov}\left(x_1,x_3\right) \\ \operatorname{cov}\left(x_2,x_1\right) & \operatorname{cov}\left(x_2,x_2\right) & \operatorname{cov}\left(x_2,x_3\right) \\ \operatorname{cov}\left(x_3,x_1\right) & \operatorname{cov}\left(x_3,x_2\right) & \operatorname{cov}\left(x_3,x_3\right) \end{bmatrix}$$

Calcular a matriz de covariância para o seguinte conjunto de dados:

```
[2, 0, -1.4]
[2.2, 0.2,-1.5]
[2.4, 0.1, -1]
[1.9, 0, -1.2]
```

As médias das variáveis são 2,125, 0,075 e -1,275. Podemos então calcular as covariâncias de cada par de variáveis para produzir a seguinte matriz de covariância:

```
import numpy as np
X = [[2, 0, -1.4],[2.2, 0.2, -1.5],[2.4, 0.1, -1],[1.9, 0, -1.2]]
print(np.cov(np.array(X).T))
```

```
[[ 0.04916667  0.01416667  0.01916667]
 [ 0.01416667  0.00916667 -0.00583333]
 [ 0.01916667 -0.00583333  0.04916667]]
```

5. Ferramentas e Métodos

Utilizarei várias ferramentas para o processo de análise e visualização dos dados, descritas logo abaixo:

- Linguagem: Farei uso da linguagem de programação Python como base para as bibliotecas de análise e visualização dos dados, além de utilizar para as técnicas de aprendizado de máquina aplicado ao trabalho.
- Colab: Farei uso da plataforma Google Collaboratory para criar e executar os códigos.
- Bibliotecas: Farei uso da biblioteca pandas para análise e modelagem dos dados, a biblioteca seaborn para plotagem e visualização dos gráficos, a biblioteca numpy para cálculos matemáticos, a biblioteca sklearn para implementar os modelos de aprendizado de máquina e k modes para algoritmo de agrupamento k-Prototype.

Depois do acerto do tema a ser desenvolvido no presente trabalho e do dataset com os dados do Censo do Ensino Superior, tentarei buscar informações complementares, para dar mais alicerçar as análises. Usando o processo de CRISP-DM do trabalho consta das seguintes fases:

5.1 Pré-processamento

Com o intuito de resolver problemas referentes a duplicação, o preparo dos dados é fundamental,como erros de digitação, valores ausentes não encontrados. Sendo assim criar um data frame limpo para análise e aplicação de modelos de aprendizado de máquina.

No planejamento para o entendimento dos dados limpamos e ajustamos os seguintes tópicos:

- Enriquecimento nos dados para constar apenas os discentes com deficiência, visto que, os arquivos contendo os dados do censo do ensino superior eram altamente pesados e precisavam ser otimizados para o foco do trabalho.

- Seleções referentes aos dados de recursos de acessibilidade disponibilizados nos cursos, com intuito de realizar uma análise sobre a quantidade de recursos oferecidos pelos cursos em relação ao número de estudantes com deficiência.

- Codificação dos Dados – nessa etapa os dados devem ser codificados de forma Numérica – Categórica, transformando os valores reais em categoria ou intervalos;

ou Categórica – Numérica, que representa numericamente valores de atributos categóricos, para que possam ser utilizados como entrada para os algoritmos de Mineração de Dados.

- Seleção de Dados – essa função envolve descobrir quais cursos funcionam(tipo da situação de funcionamento do curso), em resumo, a identificação dos dados existentes que devem, de fato, ser considerado no processo de CRISP-DM

5.2 Seleção de dados

Foram escolhidos os dados do Censo do Ensino Superior com base na obtenção dos dados acerca dos discentes com deficiência e também a respeito das Instituições de Ensino Superior, contudo, também irei pesquisar dados da população com o censo do IBGE para relacionar se às Instituições de Ensino Superior(IES) tentarei encontrar na região se há uma maior taxa de alunos com deficiência, observado na localidade da região se porventura existem grande número de pessoas com deficiência tentando entender as ações da política pública aplicada estão beneficiando aquela parcela da população. Partindo desse pressuposto, temos o intuito de usar dados de instrumentos de avaliação da educação superior, para evidenciar indicadores de qualidade para fazermos uma comparação com os dados referentes aos discentes com deficiência e aos recursos de acessibilidade das Instituições de Ensino Superior(IES). Sendo assim, os dados selecionados e utilizados no presente trabalho, foram:

● *Censo Ensino Superior: Dados de alunos, IES, cursos (2015 a 2019)*

● *Projeção da população com deficiência em 2019, da Pesquisa Nacional de Saúde(2019)*

● *Indicadores de Qualidade da Educação Superior (2019)*

● *Censo com os dados da população do IBGE (2000 a 2010)*

● *Dados da projeção da população do IBGE (2015 a 2019)*

5.3 Transformação dos dados

Para entender melhor foi realizado uma redução e transformação dos dados com o intuito de realizar uma Análise Exploratória, foram feitas algumas perguntas para dar um norte a esta etapa, como:

● Os alunos estão matriculados nos cursos que oferecem melhores condições de infraestrutura e acessibilidade?

● As universidades que oferecem maior acessibilidade são também aquelas que têm mais alunos que terminam os cursos? Para responder essas perguntas, fizemos uma redução e transformação dos dados para o tipo categórico das Instituições de Ensino Superior(IES)

ser apenas instituições públicas, pois o intuito do trabalho é analisar apenas a esfera pública da educação.

Foi feito tratamento dos dados referentes aos alunos dos anos 2015 a 2019, para que acontesse possível agruparmos todas as informações disponíveis sobre os discentes com deficiência desde o início do registro desses dados, em um processo conhecido como “staging” ou “staging area”, que corresponde a uma área de dados usadas para um agrupamento de modo consistente, preservando as características originais dos dados.

Com análise dos dados, tinha-se observado que os datasets dos discentes disponibilizados pelo INEP,tinham dados diferentes para anos distintos.

6. Análise Exploratória

Inicialmente, analisamos o comportamento da evolução dos alunos com deficiência no ensino

7. Aprendizado de máquina em fase de construção

8. Resultados

Os resultados dos modelos de aprendizado de máquina utilizados...

Referências

https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/notas_estatisticas_censo_da_educacao_superior_2020.pdf

https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior/resultados

https://agenciabrasil.ebc.com.br/saude/noticia/2021-08/pessoas-com-deficiencia-em-2019-eram-173-milhoes

https://www.redalyc.org/journal/6377/637766276013/html/

https://www.cnnbrasil.com.br/noticias/brasil-tem-mais-de-17-milhoes-de-pessoas-com-deficiencia-segundo-ibge/

https://pesquisa.bvsalud.org/brasil/resource/pt/biblio-1291532

https://www.tibco.com/pt-br/reference-center/what-is-supervised-learning
https://www.voitto.com.br/blog/artigo/modelo-preditivo
Estatística básica Interpretando Coeficientes de Correlação
https://statisticsbyjim.com/basics/correlations/

Bonfim, C.A. (2018). Como funciona uma Regressão Linear? Uma introdução sobre Regressão Linear. Disponível em: https://medium.com/data-hackers/como-funciona-uma-regress%C3%A3o-linear-f7208fa6c662

Morettin, P. A., Bussab, W. O. (2010). Estatística Básica. Editora Saraiva. 6ª edição.
VanderPlas, J.(2017) Python Data Science Handbook. Essential Tools for Working with Data. O'Reilly.
Chamberlin, E. (2020). Machine Learning with Python. ISBN: 1801098247

Data Science Academy. Deep Learning Book, 2022. Disponível em: <https://www.deeplearningbook.com.br/>. Acesso em: 10 Janeiro. 2022. https://www.deeplearningbook.com.br/aplicacoes-da-aprendizagem-por-reforco-no-mundo-real/
https://antonsruberts.github.io/kproto-audience/
SHETH, J.; PATEL, B. Best practices for adaptation of Data mining techniques in Education Sector. National Journal of System and Information Technology, v. 3, n. 2, p. 186, 2010. ISSN 0974-3308.

HACKELING,GAVIN. scikit-learn : Machine Learning Simplified: Implement scikit-learn into every step of the data science pipeline (English Edition)

A INCLUSÃO DO SURDO NO ENSINO SUPERIOR NO BRASIL
ournal of Research in Special Educational NeedsVolume 16Number s12016 85–88doi:

10.1111/1471-3802.12128 https://nasenjournals.onlinelibrary.wiley.com/doi/10.1111/1471-3802.12128

Manente, M. V., Rodrigues, O. M. P. R, de Abreu, N. R. Deficientes auditivos e escolaridade: fatores diferenciais que possibilitam o acesso ao ensino superior

https://www.researchgate.net/journal/Revista-Brasileira-de-Educacao-Especial-1413-6538

Rossetto, E. (2009). “Sujeitos com deficiência no ensino superior: vozes e significados”. Tese apresentada ao programa de Pós-Graduação em Educação da Universidade Federal do Rio Grande do Sul, como requisito parcial para obtenção do título de Doutor em educação. 2009. http://repositorio.minedu.gob.pe/handle/20.500.12799/580.

Castro, S. F., Almeida, M. A. (2014). “Ingresso e permanência de alunos com deficiência em universidades públicas brasileiras”. In: Revista Brasileira de Educação Especial [online]. 2014, v. 20, n. 2. https://doi.org/10.1590/S1413-65382014000200003

Guia de orientações básicas para a inclusão de pessoas com deficiência

https://www2.senado.leg.br/bdsf/bitstream/handle/id/42/742398.pdf?sequence=3

DBSCAN a partir de matriz vetorial ou matriz de distância
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html