UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE DO CAMPUS ARARANGUÁ
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

Eduardo Zago

# Análise de clusters de estudantes no Moodle: comparação entre diferentes combinações de atributos de entrada

Araranguá
2024

Eduardo Zago

# Análise de clusters de estudantes no Moodle: comparação entre diferentes combinações de atributos de entrada

Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia de Computação submetido ao Centro de Ciências, Tecnologias e Saúde do Campus Araranguá da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Engenharia de Computação.
Orientador: Prof. Cristian Cechinel, Dr.

Araranguá
2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Zago, Eduardo
Análise de clusters de estudantes no Moodle: comparação entre diferentes combinações de atributos de entrada / Eduardo Zago ; orientador, Cristian Cechinel, 2024.
48 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá, Graduação em Engenharia de Computação, Araranguá, 2024.

Inclui referências.

1. Engenharia de Computação. 2. Dados, Machine Learning, Classificação. I. Cechinel, Cristian . II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Computação. III. Título.

Eduardo Zago

**Análise de clusters de estudantes no Moodle: comparação entre diferentes combinações de atributos de entrada**

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel em Engenharia de Computação e aprovado em sua forma final pelo Curso de Graduação em Engenharia de Computação.

Araranguá, 27 de junho de 2024.


Documento assinado digitalmente
Jim Lau
Data: 04/07/2024 15:19:59-0300
CPF: ***.464.702-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Jim Lau, Dr.
Coordenador do Curso

**Banca Examinadora:**


Documento assinado digitalmente
Cristian Cechinel
Data: 04/07/2024 16:36:08-0300
CPF: ***.628.449-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Cristian Cechinel, Dr.
Orientador


Documento assinado digitalmente
Andrea Sabedra Bordin
Data: 04/07/2024 15:20:32-0300
CPF: ***.193.220-**
Verifique as assinaturas em https://v.ufsc.br

Profa. Andréa Sabedra Bordin, Dra.
Avaliadora
Universidade Federal de Santa Catarina


Documento assinado digitalmente
ELIANE POZZEBON
Data: 04/07/2024 15:55:41-0300
CPF: ***.213.139-**
Verifique as assinaturas em https://v.ufsc.br

Profa. Eliane Pozzebon, Dra.
Avaliadora
Universidade Federal de Santa Catarina

## AGRADECIMENTOS

Gostaria de expressar minha gratidão aos professores que me orientaram, por seus ensinamentos e dedicação ao longo deste trabalho. Agradeço também à minha família pelo apoio incondicional e incentivo contínuo. Sem vocês, esta conquista não seria possível.

Este trabalho foi financiado pelo CNPq por meio da Chamada CNPq/MCTI/SEMPI Nº 56/2022 - Apoio para Estudante Elaborando TCC em Inteligência Artificial, processo número 408101/2022-9

## RESUMO

Através dos ambientes virtuais de aprendizagem, ferramenta amplamente usada pelas instituições de ensino, são geradas grandes quantidades de dados por meio do registro da atividade dos usuários no ambiente. Analisando esses dados é possível descobrir padrões, verificar o desempenho dos alunos e analisar o risco de reprovação ou até mesmo evasão dos estudantes. Sendo assim, este trabalho utiliza registro de interações de alunos gerados no ambiente virtual de aprendizagem Moodle, para elaborar experimentos buscando a classificação de estudantes em situação de risco, conforme o seu comportamento na ferramenta de aprendizagem. Serão consideradas as contagens diárias de interações de cada aluno para classificar através do método de agrupamento, no qual busca separar os estudantes em clusters (grupos) com base na semelhança de suas características. O algoritmo utilizado será o K-means, implementado com a linguagem de programação Python e o auxílio de bibliotecas para o tratamento de dados. Será feita uma análise dos clusters criados e para verificar os resultados do algoritmo, será avaliada a concentração de estudantes em cada cluster em relação a sua nota final no período cursado. Ao final, serão obtidos grupos de estudantes com maior e menor tendência ao risco, que possibilitam a identificação por parte de instrutores e instuições de ensino alunos em situação de dificuldade.



**Palavras-chave**: Risco. Aprendizagem. Classificação. Predição. Agrupamento.

## ABSTRACT

Through virtual learning environments, a tool widely used by educational institutions, large amounts of data are generated by recording user activity in the environment. By analyzing this data, it is possible to discover patterns, check student performance and analyze the risk of students failing or even dropping out. Therefore, this work uses records of student interactions generated in the virtual learning environment Moodle, to develop experiments seeking to classify students at risk, according to their behavior in the learning tool. Each student's daily interaction counts will be considered to classify using the clustering method, which seeks to separate students into clusters (groups) based on the similarity of their characteristics. The algorithm used will be K-means, implemented with the Python programming language and the help of libraries for data processing. An analysis of the clusters created will be carried out and to verify the results of the algorithm, the concentration of students in each cluster will be evaluated in relation to their final grade in the period studied. In the end, groups of students with greater and lesser risk tendencies will be obtained, which will enable instructors and educational institutions to identify students in difficult situations.



**Keywords**: Risk. Learning. Classification. Prediction. Clustering.

## LISTA DE FIGURAS

## LISTA DE QUADROS

## LISTA DE TABELAS

## SUMÁRIO

# 1 INTRODUÇÃO

O avanço da tecnologia e a propagação da internet nos últimos anos possibilitaram o surgimento de recursos nas mais diversas áreas. No âmbito educacional, as principais mudanças refletiram na forma como o conteúdo é ofertado aos estudantes e na comunicação entre instrutor e aluno. Sendo assim, o material de estudo passou a ser fornecido em diferentes mídias oferecendo maior flexibilidade em como e onde o conteúdo é acessado, assim como permitindo uma comunicação síncrona ou assíncrona entre estudante e professor, ou até mesmo entre os alunos (MEANS *et al.*, 2009).

Sendo assim, muitas instituições de ensino superior passaram a fazer o uso dos Ambientes Virtuais de Aprendizagem (AVA) ou como também são conhecidos *Learning Management Systems* (LMS), na língua inglesa. A possibilidade de gerenciar os materiais de estudo e verificar o progresso dos alunos é o que torna os AVAs amplamente utilizado pelas instituições de ensino e instrutores (HOOSHYAR; HUANG; YANG, 2022). Além disso, um diferencial dessas plataformas é a capacidade de registrar as atividades realizadas pelos usuários (TAMADA; GIUSTI; NETTO, 2022). Com esses registros, são gerados crescentes repositórios de dados, nos quais fornecem informações sobre a forma como a aprendizagem ocorre, por meio da exploração e a análise dos mesmos (ROMERO; VENTURA, 2013).

Entre as informações que podem ser extraídas dos dados educacionais, estão a análise de desempenho, o engajamento e a permanência de alunos em determinado curso assim como possíveis intervenções visando adaptar o ambiente educacional para uma melhor experiência do estudante (GASEVIC; DAWSON; PARDO, 2016). Sendo assim, é possível utilizar esses dados para a análise de risco de reprovações ou desistências de alunos, identificando possíveis problemas no processo de aprendizagem ou dificuldades enfrentadas pelos estudantes.

O período universitário pode ser desafiador aos estudantes, visto que a demanda por suas capacidades mentais são elevadas e, para a maioria dos acadêmicos, é o início da vida adulta, onde os mesmo estão lidando com suas aspirações de carreira, começando sua jornada de independência e em muitos casos necessitando conciliar trabalho com seus estudos, além da infinidade de opções oferecidas através das mídias sociais que podem confundir e sobrecarregar o estudante, mais de 75% das doenças mentais começa antes dos 25 anos de idade (RAYASAM, 2020).

Uma pesquisa anual realizado pelo Healthy Minds Study, que desde 2007 examina a saúde mental entre estudantes de graduação e pós-gradução em mais de 530 universidades, mostra que a taxa de alunos que procuraram por terapia aumentou de 13,3% para 29,9% entre 2007 e 2019. Nesse mesmo período, o número de universitários que passaram a fazer o uso de algum medicamento psicotrópico aumentou de 11,8% para 23,9% e a parcela de alunos com alguma doença mental diagnosticada subiu de 21,9% para 35,5% (RAYASAM, 2020).

Por outro lado, o número de estudantes para cada professor tem aumentado, tornando assim difícil o acompanhamento dos instrutores para monitorar o desempenho de cada aluno e identificar possíveis problemas como o risco de reprovação ou desistência, uma solução para isso é tirar proveito dos dados provindos dos registros das plataformas educacionais (BUCOS; DRĂGULESCU, 2020). Obter informações a partir dos dados é a área de atuação do *Data Mining* (DM), ou mineração de dados, em que o processo envolve algortimos de exploração dos dados e desenvolvimento de modelos para descobrir padrões préviamente desconhecidos, podendo ser usado para análise de determinados comportamentos ou predição a partir dos dados(MAIMON; ROKACH, 2010).

A predição baseada em dados ocorre através dos métodos de aprendizagem de máquina, ou *Machine Learning* (ML), nos quais se dividem em duas principais categorias, aprendizagem supervisionada onde o resultado final é conhecido e a aprendizagem não-supervisionada onde o resultado final ainda não é conhecido (SONI, 2018). Uma das técnicas de ML para realizar predições, é a clusterização, ou agrupamento, que busca classificar dados em grupos, ou conjuntos, com características semelhantes, sendo elas amplamente utilizadas na mineração de dados (BENAIMECHE *et al.*, 2022). A predição do risco acadêmico começa pela obtenção dos dados nas plataformas educacionas, em seguida os dados são pré-processados, onde são preparados para a etapa de classificação e análise de resultados (BUCOS; DRĂGULESCU, 2020).

Este trabalho busca explorar técnicas de clusterização de estudantes baseado em suas interações com o ambiente educacional visando a classificação futura de alunos em situação de risco. Serão considerados registros de turmas já cursadas para a elaboração dos experimentos para posteriormente comparar a concentração de estudantes entre os grupos gerados em relação a nota final para verificar a situação de risco em cada grupo gerado. A seguir, no capítulo dois, será apresentado a teoria na qual este trabalho se baseia. Os estudos semelhantes sobre o assunto estão no capítulo três e, na sequência, será detalhada a metodologia empregada neste trabalho. Por fim, serão realizados experimentos e analisados seus resultados.

## 1.1 OBJETIVOS

### 1.1.1 Objetivo Geral

Elaborar experimentos visando a classificação de estudantes em risco através da contagem diária de interações dos alunos com o Ambiente Virtual de Aprendizagem (AVA) Moodle utilizando técnicas de clusterização.

### 1.1.2 Objetivos Específicos

- Utilizar dados referentes ao registro da atividade de alunos extraídos do Moodle com as notas finais.

- Contabilizar as interações de estudantes no Moodle por período de tempo.
- Explorar diferentes combinações das interações contabilizadas.
- Elaborar diferentes experimentos usando métodos de clusterização com as combinações de interações geradas.
- Analisar os grupos gerados verificando a concentração de estudantes em relação as notas finais para a classificação de risco.

# 2 REFERÊNCIAL TEÓRICO

Neste capítulo será apresentada a base teórica empregada no desenvolvimento do trabalho. Será descrito a análise de aprendizagem e a mineração de dados educacionais assim como as principais técnicas de ambas. Por fim, será apresentado a linguagem *Python* e suas bibliotecas mais utilizadas para trabalhar com dados.

## 2.1 LEARNING ANALYTICS

Conforme foram surgindo novas ferramentas de software e o crescente uso da internet na educação, foram sendo gerados crescentes repositórios de dados sobre estudantes fornecendo a possibilidade de exploração dos mesmo, principalmente em relação à como os alunos aprendem (ROMERO; VENTURA, 2013).

Com os Ambientes Virtuais de Aprendizagem (AVA) passaram a ser amplamente utilizados pelas pessoas envolvidas no processo de aprendizagem, possibilitou o registro da atividade dos estudantes dentro do ambiente, surgindo a oportunidade de análisar o seu comportamento e monitorar a sua aprendizagem (SIEMENS, 2013). Assim, um grande desafio das instituições de ensino é lidar com o crescimento exponencial desses dados e como usá-los para melhorar a aprendizagem bem como a qualidade de suas decisões gerenciais (ROMERO; VENTURA, 2013).

No início de 2011 foi realizado no Canadá a primeira conferência voltada a análise de dados educacionais, *The First International Learning Analytics Conference*, como foi chamada, simbolizou um marco na pesquisa focada na aprendizagem do ponto de vista analítico (JOKSIMOVIĆ; KOVANOVIĆ; DAWSON, 2019). A conferência também simbolizou a criação da *Society for Learning Analytics Research* (SoLAR), ou Sociedade para Perquisa de Análise de Aprendizagem, onde foi definido o termo Análise de Aprendizagem ou *Learning Analytics* (LA).

A *Learning Analytics* (LA) foi definida pela SoLAR como: "Análise de aprendizagem (LA) é a medição, coleta, análise e relatório de dados sobre os alunos e seus contextos, para fins de compreender e otimizar a aprendizagem e os ambientes em que ela ocorre" (SIEMENS; GASEVIC *et al.*, 2011). Em resumo, a Learning Analytics (LA) procura utilizar os dados provenientes das plataformas educacionais procurando melhorar a o sucesso do aluno (SIEMENS; GASEVIC *et al.*, 2011).

As principais aplicações da *Learning Analytics* (LA) são voltadas aos temas de indicadores e preditores, visualizações e intervenções (BROWN, 2012). A primeira, indicadores e preditores, incluem dados obtidos através da mineração de dados e processados através de métodos estatísticos capazes de produzir modelos preditivos, alguns exemplos incluem desempenho acadêmico, engajamento dos alunos e permanência de alunos em um curso (GASEVIC; DAWSON; PARDO, 2016). O segundo tema, as visualizações, são úteis para melhorar o entendimento de um conjunto de dados complexos, dando margem

para explorá-los como também estimular possíveis melhorias, sendo essas visualizações proveitosas por todas as partes envolvidas no processo educacional (GASEVIC; DAWSON; PARDO, 2016). A terceira categoria, está relacionada as interverções possíveis, na qual, através de abordagens analíticas, são exploradas ações para adaptar o ambiente educacional visando melhorar a experiência do estudante (GASEVIC; DAWSON; PARDO, 2016).

## 2.2 MINERAÇÃO DE DADOS EDUCACIONAIS

A mineração de dados iniciou a receber atenção no início da década de 1990 com o uso do termo KDD (*Knowledge Discovery in Databases*), que se traduz em Descoberta de Conhecimento em Bases de Dados, no qual consiste no encontro de informações úteis nos dados enquanto que a Mineração de Dados, ou Data Mining (DM), se refere a um passo especifico nesse processo, sendo esse voltado à aplicação de algoritmos para extrair padrões desses dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

No ambiente educacional, a Mineração de Dados Educacionais (MDE), ou *Educational Data Mining* (EDM) na língua inglesa, como ficou conhecida, começou a ganhar foco através de uma série de workshops realizadas no ano de 2005, em 2008 ocorreu a primeira conferência internacional em mineração de dados educacionais onde passou a acontecer anualmente (YE, 2022) e em 2011 formou-se a Sociedade Internacional de Mineração de Dados Educacionais (SIEMENS; BAKER, 2012).

De acordo com o site da Sociedade Internacional de Mineração de Dados Educacionais (educationaldatamining.org), "A Mineração de Dados Educacionais é uma disciplina emergente, preocupada em desenvolver métodos para explorar os dados únicos e cada vez mais em grande escala provenientes de ambientes educacionais e usar esses métodos para compreender melhor os alunos e os ambientes em que eles aprendem". Para Romero e Ventura (2013), MDE procura extrair padrões de grandes volumes de dados educacionais através do desenvolvimento, pesquisa e aplicações de métodos computadorizados que possibilitam o tratamento diante da enorme quantidade de informações disponíveis.

Mineração de Dados Educacionais (MDE) e Learning Analytics (LA) tem em comum o objetivo de melhorar a educação através de um aprimoramento da compreensão dos problemas existentes na área, assim como uma melhor avaliação dos processos de aprendizagem e um planejamento para intervenções, ambas buscam melhorar a qualidade da análise de grandes volumes de dados educacionais (SIEMENS; BAKER, 2012).

Apesar do mesmo objetivo, MDE e LA diferem principalmente em relação ao foco de sua utilização. MDE possui maior foco na parte tecnológica, procurando novos padrões nos dados com o desenvolvimento de algoritmos e modelos, enquanto de LA tem foco mais voltado à educação, concentrando-se em utilizar os dados integrando a parte tecnológica com as áreas sociais e pedagógicas da aprendizagem (ROMERO; VENTURA, 2020).

NA MDE a prioridade é que a obtenção de informações nos dados seja feita de

forma automatizada, utilizando o julgamento humano como ferramenta para obter esse resultado, enquanto que na LA ocorre o contrário, o julgamento humano é prioridade e a ferramenta para chegar no objetivo é a descoberta automatizada (ROMERO; VENTURA, 2013). Com base nas descobertas de ambos, os sistemas são adaptados e personalizados, para MDE o foco concentra-se na adaptação de maneira automatizada enquanto que para LA baseia-se principalmente em informar alunos e professores (SIEMENS; BAKER, 2012).

Entre as técnicas utilizadas, na MDE as mais usadas são classificação, clusterização, modelagem Bayesiana, mineração de relacionamento, descoberta com modelos (ROMERO; VENTURA, 2013) e visualização (SIEMENS; BAKER, 2012). Na LA as principais técnicas consistem em análise de sentimento, análise de influência, análise de discurso, análise de conceito, modelos de criação de sentido (ROMERO; VENTURA, 2013), a rede social e previsão de sucesso do aluno (SIEMENS; BAKER, 2012).

## 2.3 PRINCIPAIS TÉCNICAS UTILIZADAS NA MDE E LA

Nesta seção serão descritas as principais técnicas de Mineração de Dados Educacionais e *Learning Analytics*. Iniciando pelas duas categorias do *Machine Learning*, associação e classificação, e em seguida será detalhada a clusterização.

### 2.3.1 Associação

O aprendizado de máquina, ou Machine Learning (ML), é a área de estudo dedicada aos sistemas de aprendizagem de máquina e envolve múltiplas disciplinas, entre elas estão estatística, engenharia e ciência da computação além de ciência cognitiva e diversas outras áreas derivadas da ciência e matemática (GHAHRAMANI, 2003). No aprendizado de máquina existem duas principais categorias, que são o aprendizado supervisionado (classificação) e não supervisionado (associação) (SONI, 2018).

Aprendizado de máquina não supervisionado significa o processo no qual a máquina procura extrair padrões em conjuntos de dados sem uma resposta específica como referência, ou seja, a estrutura de dados é explorada visando criar hipóteses ao invés de criar modelos de previsão ou classificação com base em uma resposta ou condições específicas (VALKENBORG *et al.*, 2023).

O aprendizado não supervisionado busca identificar padrões em conjuntos de dados que ainda não possuem rótulos ou uma estrutura definida (NAEEM *et al.*, 2023). Um das técnicas de associação é o agrupamento, ou clusterização. Os algoritmos de clusterização utilizam cálculos para quantificar similaridades ou dissimilaridades entre diferentes pontos de um conjunto de dados onde são agrupados, ou separados, um dos outros conforme a semelhança de suas características (VALKENBORG *et al.*, 2023).

Alguns dos algoritmos de clusterização são: clusterização de particionamento, clusterização baseada em modelo, clusterização hierárquica e clusterização baseada em densidade.

Na clusterização de particionamento é exigido a especificação do número de cluster a serem gerados, sendo o K-means o método mais usado, a clusterização baseada em modelo utiliza a probabilidade de um ponto pertencer a um cluster, a clusterização hierárquica produz conjuntos de clusters entrelaçados em uma estrutura de árvore hierárquica e a clusterização baseada em densidade usa a densidade como critério de agrupamento (LU; UDDIN, 2024).

### 2.3.2 Classificação

O aprendizado supervisionado, diferente do não supervisionado, consiste em compreender a relação entre um conjunto de entradas e uma saída, utilizando esse entendimento para prever saídas para novos dados (CUNNINGHAM; CORD; DELANY, 2008). A aprendizagem supervisionada está presente no contexto da classificação ou regressão. Ambas possuem o mesmo objetivo de encontrar nas entradas relações ou estruturas que convertam nas saídas corretas através dos dados de treinamento, sendo a regressão aplicada para uma saída contínua (SONI, 2018).

Um dos principais métodos de aprendizagem supervisionada são as redes neurais. Inspirada no funcionamento do cérebro humano com sua gigantesca quantidade de neurônios interconectados para processar informações, as redes neurais artificiais possibilitaram extrair inteligência das máquinas (WANG; WANG, 2003). Uma rede neural artificial possui uma estrutura de grafo direcionado, na qual cada nó (vértice) representa um neurônio, onde cada nó realiza alguns cálculos simples e transmite um sinal rotulado por um número de um vértice para outro, chamado de "peso", que determina se o sinal deve ser amplificado ou atenuado em cada conexão (DONGARE; KHARDE; KACHARE *et al.*, 2012).

O aprendizado nas redes neurais artificiais ocorre através da manipulação do peso entre as conexões da rede, sendo eles adaptados em um processo de simulação contínua (treinamento), em que a taxa na qual a rede se adequa é definida como a taxa de aprendizagem e a forma como a adequação é feita define o tipo de aprendizagem da rede (DONGARE; KHARDE; KACHARE *et al.*, 2012).

Uma variação das redes neurais artificiais são as Redes Bayesianas. Também estruturada em forma de grafo direcional, porém, com a adição de probabilidade nas relações entre os vértices, possibilitando lidar com conjuntos de dados incompletos, aprender sobre relações causais, assim como utilizar técnicas estatísticas Bayesianas, que em conjunto com as Redes Bayesianas permitem melhorar a eficiência da rede reduzindo ajustes excessivos nos dados (HECKERMAN, 2008).

### 2.3.3 Clusterização

Algoritmos de clusterização tem como objetivo agrupar dados com características semelhantes, esse processo ocorre através da definição de regras de agrupamento e resulta em partições (clusters) do conjunto de dados conforme os critérios de cluster estabelecidos

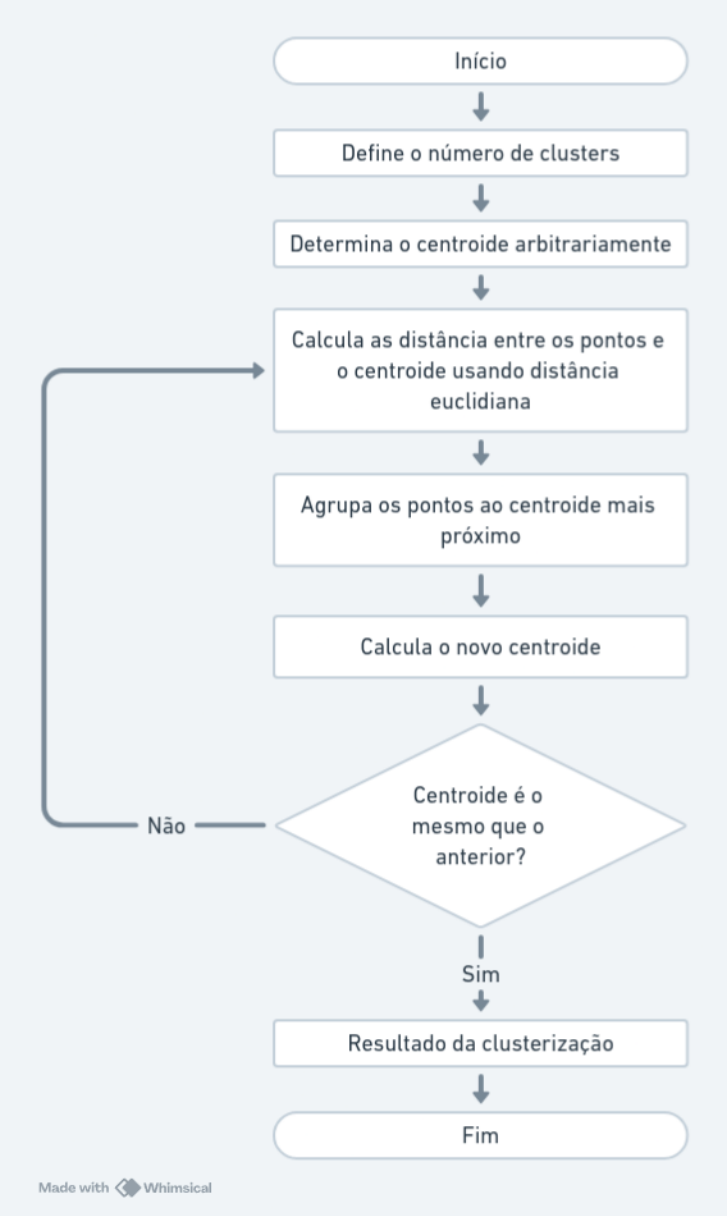
sendo que, em um cenário ideal, as instâncias de um cluster são bem diferentes das instâncias dos outros clusters (AHMED; SERAJ; ISLAM, 2020). De uma maneira geral, o desafio da clusterização é encontrar grupos homogêneos (clusters) de pontos de dados em um conjunto de dados (LIKAS; VLASSIS; VERBEEK, 2003).

O método de agrupamento normalmente mais usado é a clusterização de particionamento, no qual supõe que o conjunto de dados pode ser representado por um número finito de clusters, entre os algoritmos mais populares e antigos usado para essa categoria está o K-means (SINAGA; YANG, 2020). Por se tratar de clusterização de particionamento, o K-means precisa conhecer préviamente o número de clusters. Sendo assim, inicialmente é preciso determinar o número k de clusters, cada cluster terá um ponto central, chamado de centroide, no qual os dados mais próximo de cada centróide serão agrupados a ele formando um cluster (NA; XUMIN; YONG, 2010). O K-means é um algoritmo iterativo, ou seja, após definir um centroide inicial, as distâncias entre os pontos e o centroide são calculadas e agrupadas, é calculado novamente os centroides e as distâncias para agrupar, esse processo ocorre de maneira iterativa até que o centroide determinado não mude significativamente comparado com o anterior (UMARGONO; SUSENO; GUNAWAN, 2019). O fluxograma de execução do K-means é esboçado na figura 1

Existem diversos métodos para o cálculo do número inicial de clusters e para determinar os centróides. Em geral, o centroíde pode ser a média de todos os dados que compõem o cluster ou o dado com mais representatividade no grupo de dados (UMARGONO; SUSENO; GUNAWAN, 2019). A definição do número de clusters, o valor de k, pode ser realizada de formas distintas. Uma forma menos complexa e visual é o método elbow, ou método do "cotovelo", que calcula as distâncias quadráticas de um conjunto de pontos até seu centroide iterativamente para um sequência de valores de k e a soma do quadrado das distâncias representa as inércias dos pontos, indicando menor convergência para valores altos de inércia e maior convergência para baixos valores de inércia (YUAN; YANG, 2019). Em outras palavras, uma maior convergência acontecerá quando o número de clusters tende ao número de pontos do conjunto de dados e o mesmo acontece para quando existe só um cluster onde todos os pontos farão parte do mesmo agrupamento com uma baixa convergência.

O método elbow utiliza o gráfico para avaliar o valor de k (número de clusters). Se colocarmos graficamente, no eixo horizontal os valores de k iniciando em 1 até um determinado número máximo, e no eixo vertical colocarmos as inércias, teremos um gráfico das inércias para cada valor de k, onde será possível observar uma queda acentuada no gráfico em uma determinada região de valores de k, formando um "cotovelo", essa região indica o número ideal de clusters para esse conjunto de dados (YUAN; YANG, 2019). A figura 2 mostra um exemplo de gráfico gerado pelo método elbow, onde é possível perceber como as inércias caem de maneira rápida entre k=1 e k=4, formando o "cotovelo´´, para então estabilizar. O ponto de maior inflexão, está localizado em k=3, onde possui maior

Figura 1 – Fluxograma da execução do K-means

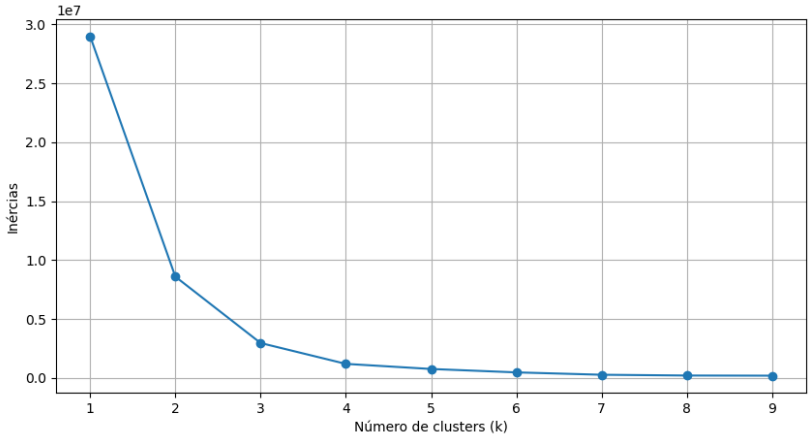



Fonte: Figura adaptada de (UMARGONO; SUSENO; GUNAWAN, 2019)

diferença do número de inércias em relação aos valores maiores de k, indicando que o k=3 é um bom valor para k.

O valor de k nem sempre é obvio e uma das desvantagens do método elbow são os casos onde não está claramente visível o ponto de inflexão ("cotovelo") no gráfico (YUAN; YANG, 2019). Além disso, existem outros métodos para encontrar o valor ideal de k, como o método da silhueta (Silhouette Method) que considera a semelhança de um objeto de dado a um cluster, separando conforme a similaridade com cada cluster, o método utiliza uma escala de -1 até +1 onde valores próximos a +1 indicam alta similaridade e valor próximos à -1 indicam baixa similaridade (YUAN; YANG, 2019).

Figura 2 – Representação gráfica do método elbow




Fonte: Figura do autor (2024)

## 2.4 PYTHON

A linguagem de programação *Python* se popularizou nos últimos anos entre as tarefas envolvendo dados devido ao aperfeiçoamento de suas bibliotecas open-source, que passaram a fornecer uma ampla gama de ferramentas para trabalhar com dados (MCKINNEY, 2022). Entre as prinpicais bibliotecas Python, para o contexto de dados, estão *pandas*, *Matplotlib*, *scikit-learn* e *searborn*.

### 2.4.1 Pandas

O *pandas* é uma biblioteca open-source na qual oferece ferramentas para a manipulação de dados em Python. O *pandas* possui estruturas de dados que permitem carregar datasets em um objeto, chamado de DataFrame, no qual permite manipular de maneira rápida e eficiente um conjunto de dados fornecendo indexação para simplificar o tratamento dos mesmos, além de possibilitar a leitura e gravação de dados na memória em diferentes formatos como CSV, Microsoft Excel, e banco de dados SQL (PANDAS.PYDATA.ORG, 2024).

### 2.4.2 Matplotlib

A biblioteca open-source *Matplotlib* permite criar visualizações estátisticas, interativas e animadas em Python. Através da biblioteca é possível criar uma grande variedade de gráficos customizáveis como gráficos de linhas, gráficos de barras, gráficos de dispersão (scatter plots) e histogramas, além de ser integrada com outras bibliotecas como o *pandas*, facilitando o trabalho com dados (MATPLOTLIB.ORG, 2024).

### 2.4.3 Scikit-learn

A biblioteca open-source *scikit-learn* oferece ferramentas simples para facilitar a análise preditiva de dados, construída sobre as bibliotecas *NumPy*, *SciPy* e *Matplotlib*, a *scikit-learn* fornece ferramentas simplificados para trabalhar com algoritmos de machine learning como classificação, regressão e agrupamento (clusterização) (SCIKIT-LEARN.ORG, 2024b).

### 2.4.4 Searborn

*Seaborn* é uma bilbioteca Python, baseada em *Matplotlib* e integrada às estruturas do *pandas*, para visualização de dados, permitindo a melhor exploração e compreensão dos dados com ferramentas para observar características e padrões através de gráficos estatísticos, além disso, possibilita a integração com *pandas* e *Matplotlib*, simplificando o trabalho com as principais bibliotecas utilizadas no contexto dos dados (SEABORN.PYDATA.ORG, 2024).

# 3 ESTADO DA ARTE

Visando verificar os trabalhos mais recentes relacionados a predição de risco acadêmico utilizando clusterização, foi realizada uma revisão de estado da arte, que tem por objetivo verificar trabalhos recentes, além de poder apontar possíveis áreas de atenção ou fornecer diferentes pontos de vista sobre determinado assunto (DE SOUSA *et al.*, 2018).

O protocolo de pesquisa foi montado a partir das seguintes questões:

1. Quais dados foram utilizados para a predição?
2. Quais foram o algoritmos de clusterização utilizados?

A pesquisa foi realizada utilizando três bases da dados, sendo elas Scopus, Web of Science e IEEE Xplore. O primeiro passo consistiu em definir as palavras chaves, sendo definidas em inglês: *prediction* (predição), *students* (estudantes), *risk* (risco) e *clustering* (clusterização). Assim, o passo seguinte foi montar a *string* de busca que resultou em: prediction AND students AND risk AND clustering.

A partir da pesquisa realizada os resultados foram préviamente filtrados por ano. Visando estudos mais recentes, foram selecionados trabalhos realizados entre os anos de 2020 até 2023. Os resultados são apresentados na tabela 1.

| **Base de dados** | **Resultados** | **Resultados de 2020 à 2023** |
|---|---|---|
| Scopus | 71 | 31 |
| Web of Sciente | 98 | 47 |
| IEEE Xplore | 12 | 7 |

Tabela 1 – Resultados antes e após o filtro por base de dados

Como mostrado na tabela 1, o maior número de resultados foi encontrado na base de dados Web of Science seguido pela Scopus e IEEE Xplore, respectivamente. Ao todo foram encontrados 181 trabalhos nas três bases de dados. Após a filtragem pelo período de tempo restaram 85. Nas bases de dados Web of Science, foram detectados diversos resultados com assuntos sem ligação com a pesquisa, então foi adicionado um novo filtro onde foi refinado pelos tópicos de citação “Education & Educational Research” e “Artificial Intelligence & Machine Learning”. Assim, dos 47 filtrados anteriormente restaram 16, totalizando 54 trabalhos.

A próxima etapa de filtragem consistiu em analisar, de forma prévia, a proximidade de cada trabalho individualmente com o assunto da pesquisa assim como eliminar duplicações encontradas. Nesta etapa também foi priorizada a escolha de trabalhos com acesso aberto, acessíveis através da forma “Open Access”. Os critérios de inclusão e exclusão são mostrados na tabela 2.

Após aplicar os critérios de inclusão e exclusão, restaram 10 trabalhos. Os resultados por base de dados estão expressos na tabela 3.

| **Inclusão** | **Exclusão** |
|---|---|
| Acesso aberto | Acesso fechado |
| Similaridade com o tema | Pouca proximidade com o tema |

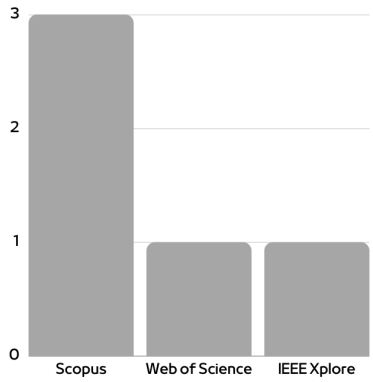
Tabela 2 – Critérios de inclusão e exclusão

| **Base de dados** | **Resultados** | **Selecionados** |
|---|---|---|
| Scopus | 31 | 6 |
| Web of Sciente | 16 | 2 |
| IEEE Xplore | 7 | 2 |

Tabela 3 – Seleção após os primeiros critérios de inclusão e exclusão

A última etapa da filtragem foi realizada a partir de uma análise mais profunda em cada trabalho selecionado. Aqui, o critério de seleção foram os trabalhos que exclusivamente utiliza dados de AVAs combinados com clusterização. O resultado final foi de 5 trabalhos, conforme aprensentados na figura 3.

Figura 3 – Resultados finais por base de dados




Fonte: Figura do autor (2024)

Após a seleção dos trabalhos foram extraídas as informações prévias como o país, o período de tempo e tipo de curso em que foram realizados os experimentos. De acordo com a tabela 4, dos seis trabalhos selecionados dois foram realizados na Estônia, um foi realizado na Romênia, outro no Brasil e um no Reino Unido. O período de tempo avaliado variou entre 1 à 4 anos, sendo apenas um realizado com alunos de curso técnico enquanto os demais foram realizados em universidades.

Em seguida foram extraídas as informações sobre cada trabalho buscando responder as perguntas elaboradas no início da pesquisa.

Em Bucos e Drăgulescu (2020) foram utilizados os registros de atividades dos alunos dentro da plataforma Moodle e um livro de notas referentes a cada encontro ao

| **Citação** | **País** | **Período de tempo** | **Tipo de curso** |
|---|---|---|---|
| Bucos e Drăgulescu (2020) | Romênia | 2015 à 2019 | Graduação |
| Hooshyar, Huang e Yang (2022) | Estônia | - | Graduação |
| Shafiq *et al.* (2022) | Reino Unido | - | Graduação |
| Tamada, Giusti e Netto (2022) | Brasil | 2016 à 2018 | Curso Técnico |
| Yang *et al.* (2020) | Estônia | - | Graduação |

Tabela 4 – Informações prévias retiradas dos trabalhos selecionados

longo do semestre de determinada disciplina dos cursos de técnologia da Universidade Politécnica de Timișoara, na Romênia. Assim, relacionando a atividade do estudante dentro da plataforma com a sua respectiva performance através do livro de notas onde foram considerados 12 encontros semestrais. O algoritmo de clusterização utilizado foi o K-means, resultando em um número ideal de clusters de 5 e chegando a conclusão de que 86% dos estudantes estão em risco. O trabalho contou com a coleta dos registros entre os anos de 2015 à 2019, o maior período de tempo em relação aos demais trabalhos.

No trabalho de Hooshyar, Huang e Yang (2022), é proposto uma predição de risco baseada nas informações de três disciplinas da Universdade de Tartu, na Estônia, onde foram levados em consideração apenas o registro de interações dentro da plataforma Moodle. O algoritmo utilizado foi o X-Means onde foram separados em apenas dois clusters e analisado o conteúdo acessado, o engajamento e a avaliação do aluno. O modelo proposto se mostrou com um acurácia de mais de 90%.

Em Shafiq *et al.* (2022) é proposto um modelo conceitual para análise de risco utilizando dados da Open University (OU) do Reino Unido, que fornece informações obtidas da plataforma educacional de 32.593 estudantes. Para comparação foram usados os resultados dos estudantes com o algoritmo de clusterização K-means.

No trabalho de Tamada, Giusti e Netto (2022) os dados utilizados foram o registro de interações e informações socioeconômicas dos estudantes de curso técnico em sincronia com ensino médio do Instituto Federal de Roraíma, no Brasil. O método de clusterização empregado foi o K-means com distância Euclidiana. Para o trabalho foram levados em conta o progresso do curso, analisando com 20%, 40% e 60% do curso concluido.

Em Yang *et al.* (2020), assim como Hooshyar, Huang e Yang (2022), foi realizado na Univerdade de Tartu, na Estônia, considerando dados de acesso ao Moodle e avaliando a procastinação dos estudantes em relação as tarefas da plataforma. Com o algoritmo K-means foi realizada a etapa de clusterização obtendo um acurácia de 87% e 84% com o método de classificação L-SVM.

Os resultados das informações necessárias para responder as questões formuladas anteriormente estão expressas na tabela 5.

- Quais dados foram utilizados para a predição?

Para responder a primeira pergunta da pesquisa, de acordo com a tabela 5, os dados analisados foram em sua ampla maioria levando em consideração os registro de interações dentro do AVA onde em alguns casos é feita a análise do resultado para estimar a eficiência do método empregado, enquanto que em outros casos foram avaliadas somente o registro de interações.

- Quais foram o algoritmos de clusterização utilizados?

Conforme a tabela 5, o algoritmo K-means foi o mais empregado, sendo em alguns casos utilizando alguma variação do método com é o caso de Tamada, Giusti e Netto (2022) onde foi empregado o K-means com distância euclidiana. A exceção ao K-means parte do trabalho de Hooshyar, Huang e Yang (2022) que utiliza o X-means.

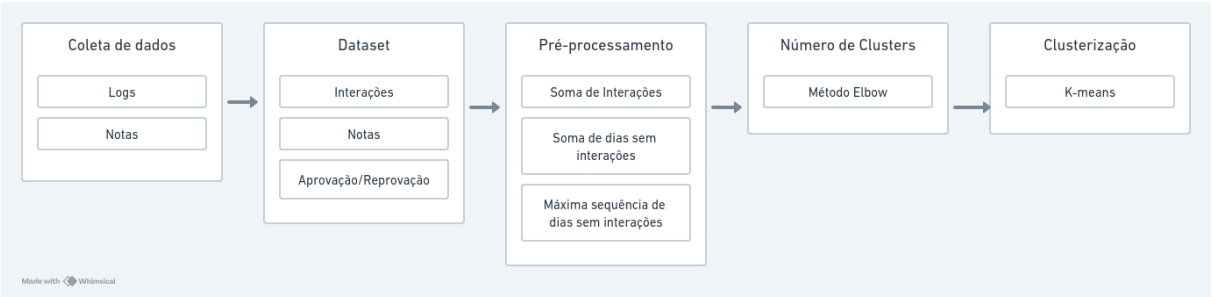
| **Citação** | **Dados utilizados** | **Algoritmo de clusterização** |
|---|---|---|
| Bucos e Drăgulescu (2020) | Registro de interações e resultados | K-means |
| Hooshyar, Huang e Yang (2022) | Registro de interações | X-means |
| Shafiq *et al.* (2022) | Registro de interações e resultados | K-means |
| Tamada, Giusti e Netto (2022) | Registro de interações e informações socioeconômicas | K-means com distância euclidiana |
| Yang *et al.* (2020) | Registro de interações | K-means |

Tabela 5 – Informações prévias retiradas dos trabalhos selecionados

# 4 METODOLOGIA

Nesta seção será demonstrada em detalhes as etapas para a realização dos experimentos de clusterização. O trabalho inicia pela coleta de dados do Ambiente Virtual de Aprendizagem (AVA) Moodle, onde é gerado um dataset no qual possui as interações dos alunos ao longo do tempo assim como as notas e resultado do aluno (aprovado ou reprovado). Em seguida ocorre o pré-processamento do dataset visando adaptá-lo ao experimento que virá em sequência. A próxima etapa consiste em determinar o número ideal de clusters no qual o conjunto de dados pode ser particionado e por fim é aplicada a clusterização como mostrado na figura 4.

Figura 4 – Etapas do trabalho




Fonte: Figura do autor (2024)

## 4.1 COLETA DE DADOS

Os dados utilizados neste trabalho são registros de 12 disciplinas já cursadas ofertadas semestralmente na Universidade Federal de Santa Catarina (UFSC) entre os anos de 2018 até 2023. Ao todo são 12 turmas com um total de 103 estudantes. Serão considerados os registros de atividades dos alunos no Ambiente Virtual de Aprendizagem (AVA) durante 20 semanas, período no qual, em média, representa um semestre letivo. Adicionalmente será considerado o registro de notas finais com o resultado de cada estudante.

### 4.1.1 Atividade dos Usuários

Os dados da atividade dos estudantes dentro do Moodle são extraídos em uma tabela no formato CSV, cada linha represetando uma atividade realizada na plataforma e contém os seguintes parâmetros:

- Data e hora do registro
- Nome completo do estudante
- Usuário afetado (caso a ação tenha afetado algum usuário)

- Disciplina que o evento ocorre
- Componente acessado dentro da disciplina
- Nome do evento
- Descrição do evento
- Origem do evento (plataforma na qual o usuário está acessando a plataforma)
- Endereço IP do usuário

### 4.1.2 Notas Finais

As notas finais dos estudantes são uma tabela obtida separadamente no formato CSV, cada linha representando um estudante e sendo composta pelos seguintes parâmetros:

- Nome
- Sobrenome
- Código do curso
- Nome do curso
- Número de matrícula
- Nota final
- Último download realizado no curso

## 4.2 DATASET

Com os dados coletados na etapa anterior, são selecionadas as informações mais relevantes para gerar um dataset com o auxílio da biblioteca *pandas*. Visando analisar o engajamento do estudante com o ambiente educacional, são retirados dos dados coletados informações referentes as interações do estudante com o AVA, onde serão consideradas a contagem de interações que o aluno realiza por dia, e por fim, o resultado do aluno na disciplina.

### 4.2.1 Contabilização de Interações e Notas

Para a contabilização das interações, são consideradas as contagens de eventos que cada aluno realiza por dia dentro do ambiente educacional. As ações contabilizadas são referentes ao acesso as páginas da disciplina bem como a utilização dos recursos ali presentes. Os recursos oferecidos aos estudantes incluem conteúdos e atividades disponibilizados pelo

professor assim como o registro de presença e notas. Cada evento é reconhecido como uma interação do aluno com a plataforma, o valor é incrementado a cada nova ação realizada pelo estudante.

Assim, o resultado é um dataset com os estudantes organizados em linhas e suas contagens de interações por dia organizadas em colunas. Ao final, são incluidas as colunas de notas e resultados (aprovado ou reprovado). Dessa forma, o dataset resultante possui o formato apresentado na tabela 6, com id_group referente ao ID da disciplina, id_subject contendo o ID do usuário, key com o nome do usuário, datapoints referentes ao número de interações sendo uma coluna referente a cada dia contabilizado, target contendo as notas finais e target_cat referente ao resultado do aluno, aprovado ou reprovado.

| **id_group** | **id_subject** | **key** | **datapoint1** | **...** | **target** | **target_cat** |
|---|---|---|---|---|---|---|
| ID da disciplina | ID do usuário | Nome do usuário | Nº de interações no dia 1 | Nº de interações nos demais dias | Nota final | Aprovado ou reprovado (0 ou 1, respectivamente) |

Tabela 6 – Formato do dataset gerado com os dados coletados

## 4.3 PRÉ-PROCESSAMENTO

Nesta etapa, serão realizadas as filtragens no dataset obtido para realizar a aplicação da clusterização. Visando explorar diferentes combinações entre as interações de cada estudante e, com o auxílio da biblioteca *pandas*, as interações contabilizadas de cada estudante foram organizadas em: somatórios das interações (P1), somatórios dos dias sem interações (P2) e a máxima sequência de dias sem interações. Os pré-processamentos são listados e descritos com mais detalhes na tabela 7.

| **Pré-processamento** | **Descrição** | **Nome da coluna** |
|---|---|---|
| P1 | Soma de todas as interações de cada aluno no período | sum |
| P2 | Soma dos dias sem interações do aluno no período | sum_zeros |
| P3 | Maior sequência de dias sem interações do aluno no período | max_zeros |

Tabela 7 – Tabela de pré-processamentos

Sendo assim, a partir do dataset obtido na etapa anterior, os alunos permanecem organizados em linha e o resultado de cada pré-processamento será uma coluna. Assim, o dataset inicial se reduz ao formato apresentado na tabela 8

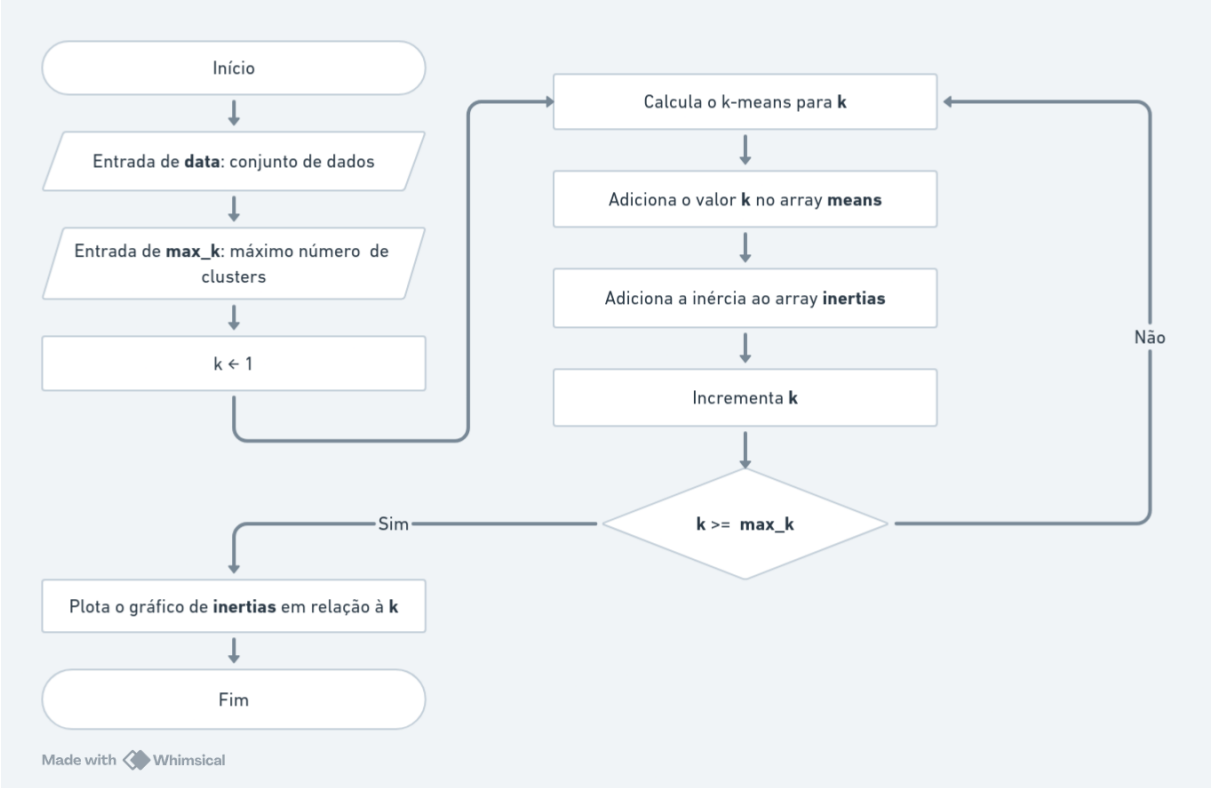
| **index** | **sum** | **sum_zeros** | **max_zeros** | **target** |
|---|---|---|---|---|
| Aluno | P1 | P2 | P3 | Nota final |

Tabela 8 – Formato do dataset resultante do pré-processamento

## 4.4 CÁLCULO DO NÚMERO DE CLUSTERS

Nesta etapa ocorre o cálculo no número ideal de clusters para a execução do algoritmo. Entre as diversas forma de determinar esse valor, será considerado o método elbow, ou método do cotovelo, devido a sua simples complexidade (YUAN; YANG, 2019). A implementação do método ocorre através da biblioteca *scikit-learn*, na qual irá calcula o K-means para k de 1 até um número máximo definido. Para cada valor de k será obtida sua inércia correspondente. Os valores de k são salvos em um array chamado **means** e as inércias são inseridas em um array chamado **inertias** e ao final, com a biblioteca *Matplotlib*, será plotado o gráfico dos valores de inércias em relação aos valores de **k**. A figura 5 apresenta o fluxograma do método elbow.

Figura 5 – Fluxograma do método elbow




Fonte: Figura do autor (2024)

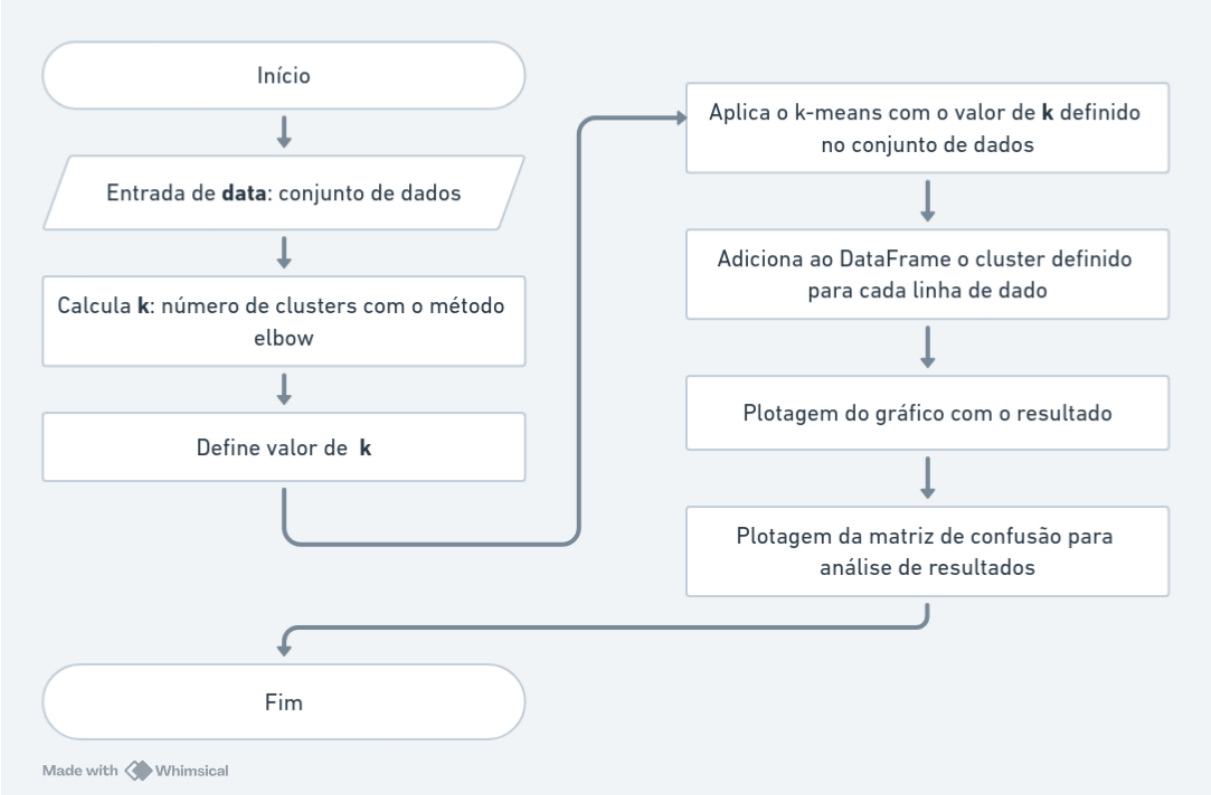
Para esta etapa, considera-se o valor máximo de clusters (max_k) o valor 10, por ser um número suficiente para a grande maioria dos casos.

## 4.5 CLUSTERIZAÇÃO

A etapa de clusterização é feita com o algoritmo K-means, devido a sua ampla popularidade, como mostrado no capítulo 3. Os experimentos acontecem com o uso da biblioteca *scikit-learn* na qual fornece ferramentas para trabalhar com clusterização e, mais especificamenete, o K-means. Por padrão, o K-means implementado pela biblioteca *scikit-learn* utiliza K-means++ para determinar os centroides iniciais, no qual é detalhado em sua documentação: "'k-means++': seleciona centróides iniciais do cluster usando amostragem baseada em uma distribuição de probabilidade empírica da contribuição dos pontos para a inércia geral. Esta técnica acelera a convergência. O algoritmo implementado é 'ganancioso k-means++'. Ele difere do vanilla k-means++ por fazer vários testes em cada etapa de amostragem e escolher o melhor centróide entre eles" (SCIKIT-LEARN.ORG, 2024a).

O experimento acontece conforme o fluxograma mostrado na figura 6. Iniciando pela entrada de dados, é calculado k com o método elbow, aplicação do K-means para o valor de k definido e com o auxilio da biblioteca *Matplotlib* é feita a plotagem do resultado. Adicionalmente, com a biblioteca *seaborn*, é feita uma plotagem de uma matriz de confusão para analisar o agrupamento e verificar a distribuição dos alunos em cada cluster por nota.

Figura 6 – Fluxograma da realização dos experimentos




Fonte: Figura do autor (2024)

Os experimentos são organizados em itens conforme o pré-processamento realizado anteriormente, totalizando 3 experimentos. Cada experimento é realizado baseado em

duas dimensões, na qual serão as duplas de combinações geradas no pré-processamento, conforme listado na tabela 9.

| **Experimento** | **Dados** |
|---|---|
| E1 | P1 e P2 |
| E2 | P2 e P3 |
| E3 | P1 e P3 |

Tabela 9 – Tabela de experimentos

# 5 DESENVOLVIMENTO

Nesta seção serão demonstrados os experimentos de clusterização conforme a metodologia detalhada no capítulo anterior. O roteiro de cada experimento inicia com o cálculo do número de clusters (k), seguido pela aplicação do K-means para esse valor. O resultado é plotado num gráfico de pontos com os dados agrupados e, por fim, será gerada uma matriz de confusão para verificar o experimento através da distribuição dos estudantes por cluster em relação as notas finais.

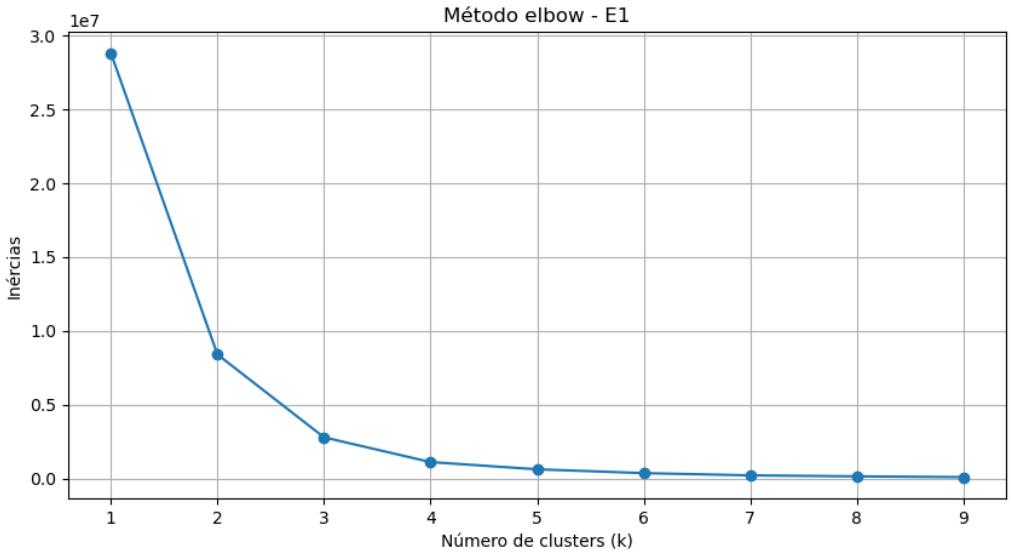
## 5.1 E1: SOMA DE INTERAÇÕES E DIAS SEM INTERAÇÕES

O experimento 1 (E1) irá considerar os somatórios das interações dos alunos (P1) e os somatórios de dias sem interações dos estudantes com o Moodle (P2).

### 5.1.1 Cálculo do Número de Clusters

Os número k de clusters é determinado através do método elbow no qual gera um gráfico indicando o número ideal de k. A figura 7 mostra o resultado da aplicação do método para este experimento.

Figura 7 – Gráfico resultante do método elbow do E1



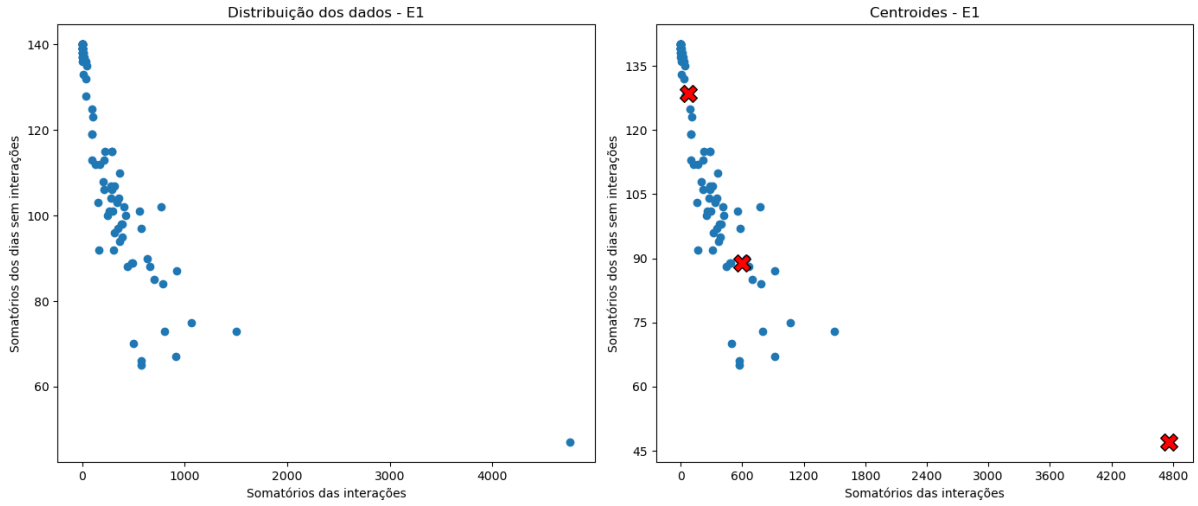

Fonte: Figura do autor (2024)

Conforme a figura 7, a maior queda das inércia está entre 2 e 4. Assim, conforme explicado na seção 4, o valor escolhido será k=3.

### 5.1.2 Clusterização

Com o valor de k definido, os centroides são calculados. A figura 8 apresenta a distribuição dos dados à esquerda e à direita são mostrados os centroides, em vermelho, determinados pelo algoritmo.

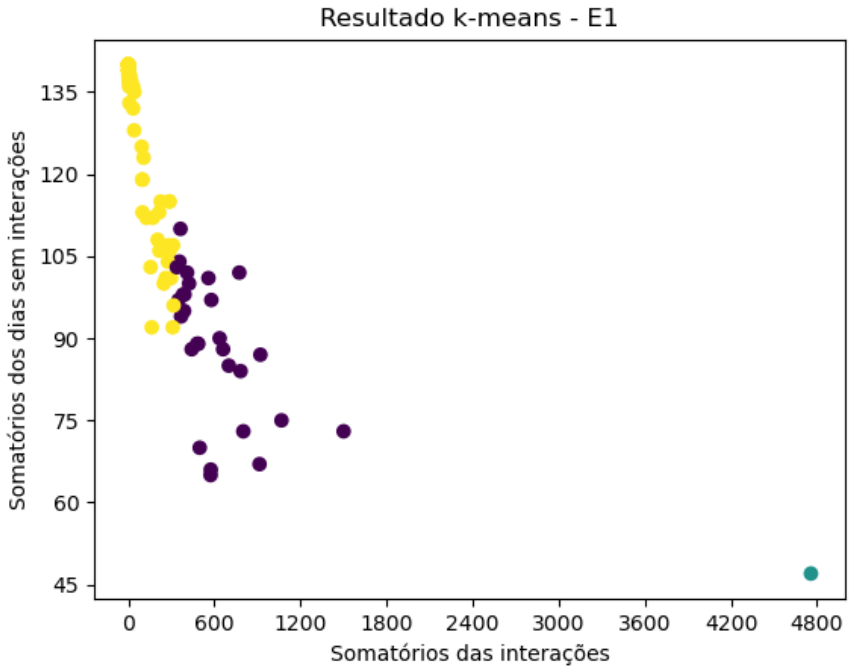
Figura 8 – Gráfico da distribuição dos dados e os centroides do E1




Fonte: Figura do autor (2024)

Sendo assim, é realizada a aplicação do k-means no qual agrupará os dados de acordo com os centroides determinados anteriormente. O resultado é apresentado na figura 9, onde está exposto o resultado do particionamento em 3 clusters.

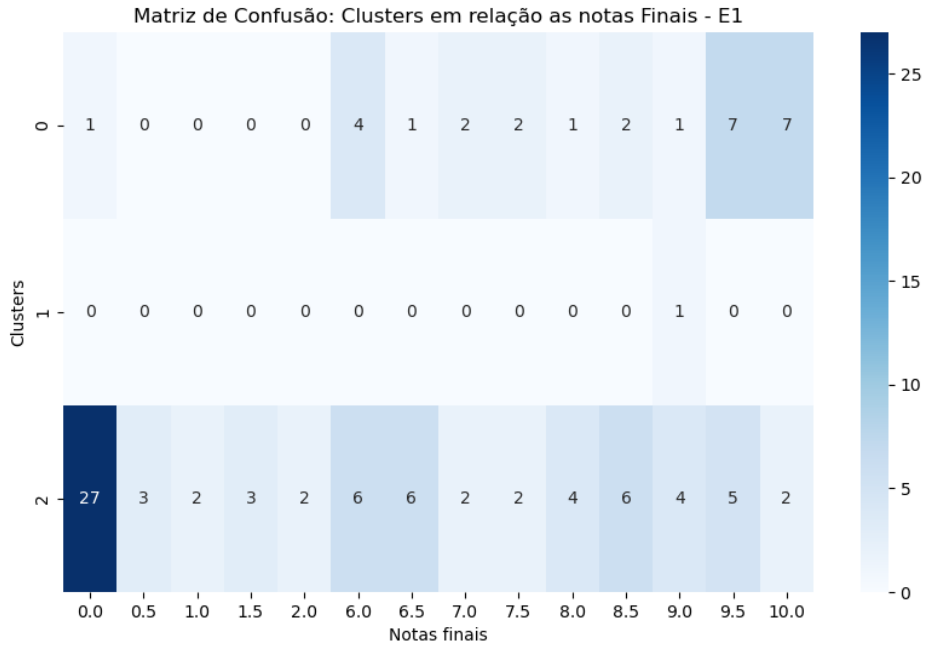
Figura 9 – Resultados do k-means do E1




Fonte: Figura do autor (2024)

Com o objetivo de verificar o resultado, será gerada uma matriz de confusão dos clusters em relação as notas finais dos estudantes, onde mostra a concentração de estudantes em cada cluster com suas respectivas notas, conforme mostrado na figura 10.

Figura 10 – Matriz de confusão do E1




Fonte: Figura do autor (2024)

Através da figura 10 é possível perceber a divisão das notas finais em cada cluster. Os percentuais da distribuição de notas por cluster são listados na tabela 10, na qual separa o percentual de notas menores e notas maiores que 5.

| **Cluster** | **Notas: 0 - 5 (%)** | **Notas: 5 - 10 (%)** |
|---|---|---|
| Cluster 0 | 3,57 | 96,43 |
| Cluster 1 | 0 | 100 |
| Cluster 2 | 50,00 | 50,00 |

Tabela 10 – Tabela percentual da distribuição de notas por cluster do E1

### 5.1.3 Análise

O experimento 1 (E1) divide o conjunto de dados em 3 cluster, no qual, a composição do Cluster 0 possui cerca de 96% de alunos com resultado final acima de 5, indicando que os estudantes desse cluster possuem, em sua maioria, um menor risco. O Cluster 1, possui apenas um estudante com nota acima de 5, no qual é uma excessão, devido ao estudante possuir a soma de suas interações significativamente maior que os demais alunos presentes no dataset. Por fim, o Cluster 2 possui 50% de alunos com notas maiores e menores que 5, indicando equilíbrio, porém, com maior tendência ao risco, visto que possui grande concentração de alunos com nota final igual a zero.

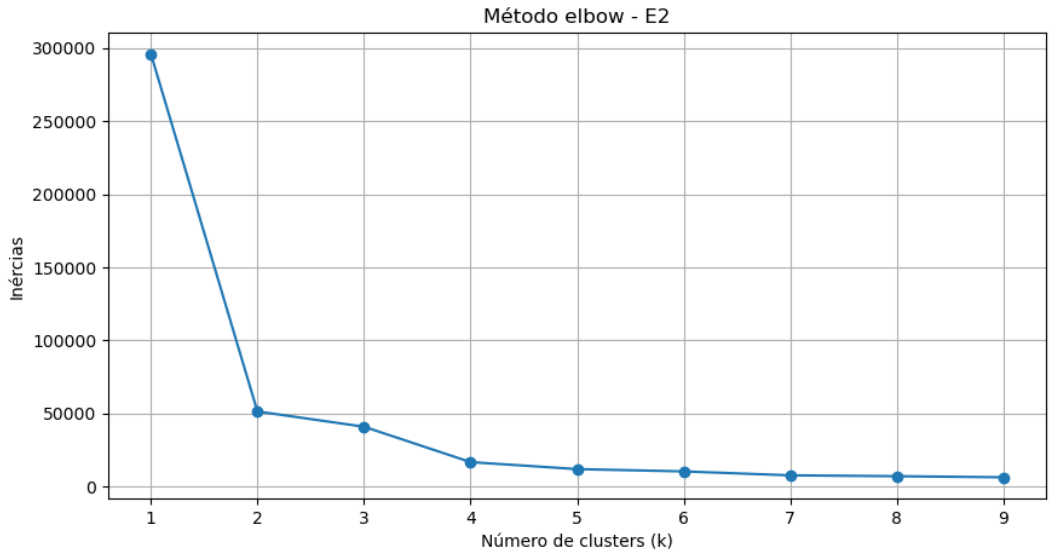
## 5.2 E2: DIAS SEM INTERAÇÕES E MÁXIMA SEQUÊNCIA SEM INTERAÇÕES

O experimento 2 (E2) considera os somatórios dos dias sem interações (P2) e as máximas sequências de dias sem interações dos alunos com o Moodle (P3).

### 5.2.1 Cálculo do Número de Clusters

Os número k de clusters é determinado através da figura 11, que mostra o resultado da aplicação do método elbow.

Figura 11 – Gráfico resultante do método elbow do E2



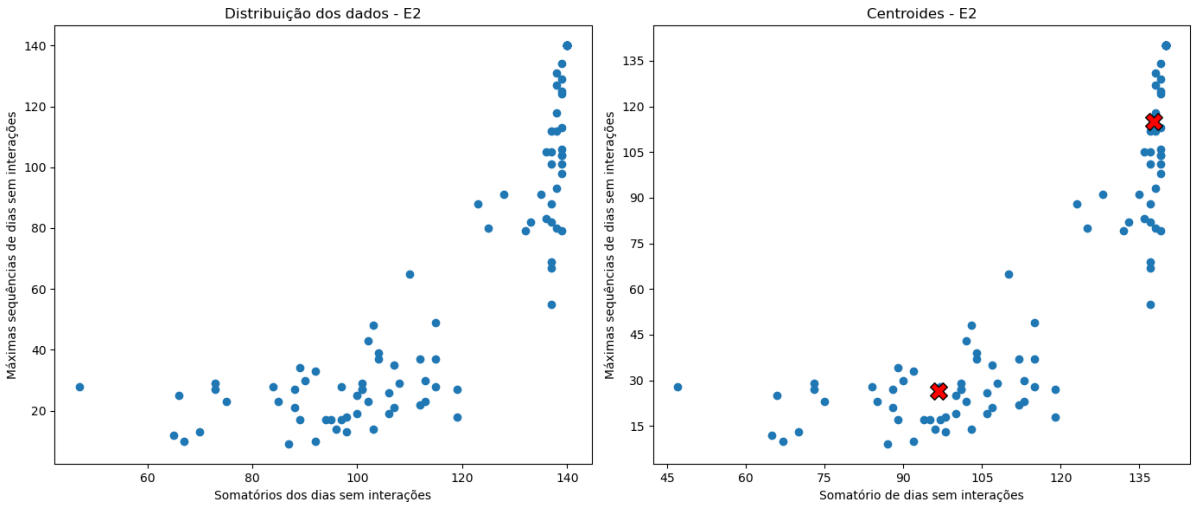

Fonte: Figura do autor (2024)

O resultado na figura 11 mostra irregularidade na queda do gráfico, porém, como a maior queda das inércia ocorre em k=2, será escolhido o valor 2 para k.

### 5.2.2 Clusterização

Na figura 12 são mostrados os centroides calculados pelo algoritmo e a distribuição dos dados.

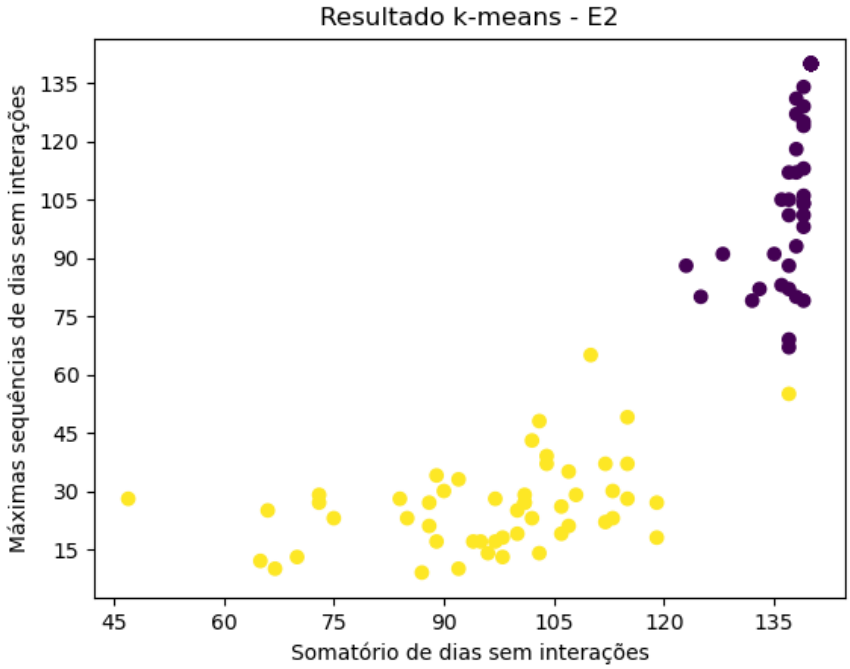
Figura 12 – Gráfico da distribuição dos dados e os centroides do E2




Fonte: Figura do autor (2024)

Com os centroides definidos, o K-means é aplicado gerando o resultado exposto na figura 13.

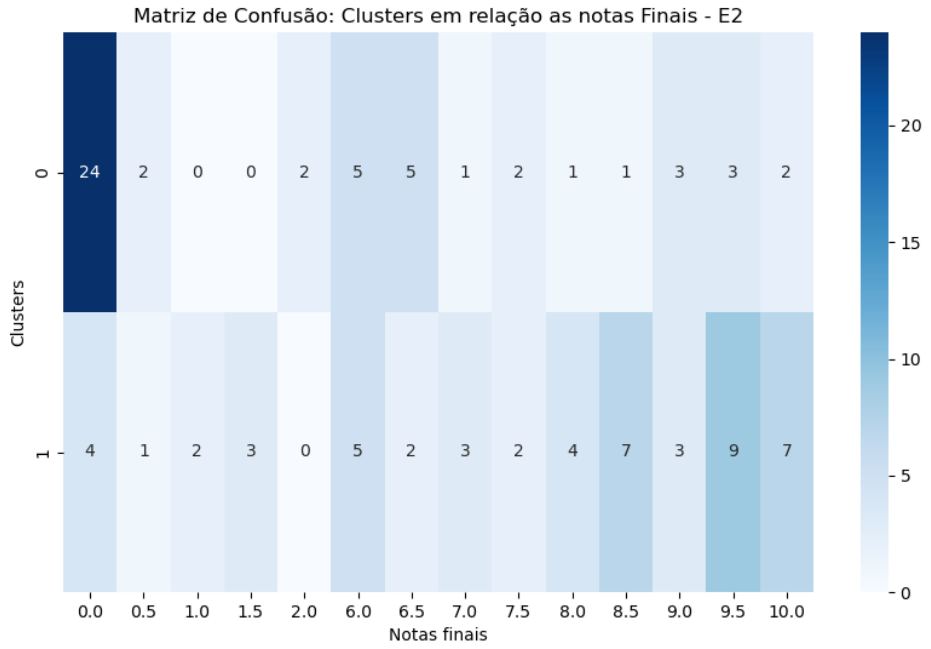
Figura 13 – Resultados do k-means do E2




Fonte: Figura do autor (2024)

Assim, com o auxílio da matriz de confusão, são analisados os resultados através da concentração de estudantes em cada cluster em relação as suas notas, conforme a figura 14.

Figura 14 – Matriz de confusão do E2




Fonte: Figura do autor (2024)

Dessa forma, são calculados os percentuais de notas maiores e menores que 5, conforme a tabela 11.

| **Cluster** | **Notas: 0 - 5 (%)** | **Notas: 5 - 10 (%)** |
|---|---|---|
| Cluster 0 | 54,90 | 45,10 |
| Cluster 1 | 19,23 | 80,77 |

Tabela 11 – Tabela percentual da distribuição de notas por cluster do E2

### 5.2.3 Análise

O experimento 2 (E2) divide o conjunto de dados em dois clusters. O Cluster 0, apesar de não agrupar uma parcela significativa de estudantes com notas inferiores a 5, cerca de 54%, o Cluster 0 indica maior tendência ao risco dos estudantes desse grupo, principalmente devido ao fato de concentrar grande número de estudantes com nota final igual a zero. O Cluster 1, possui uma divisão mais clara dos estudantes em relação ao seu resultado final, com aproximadamente 80% dos estudantes desse grupo com notas superiores a 5, apontando menor risco aos estudantes desse cluster.

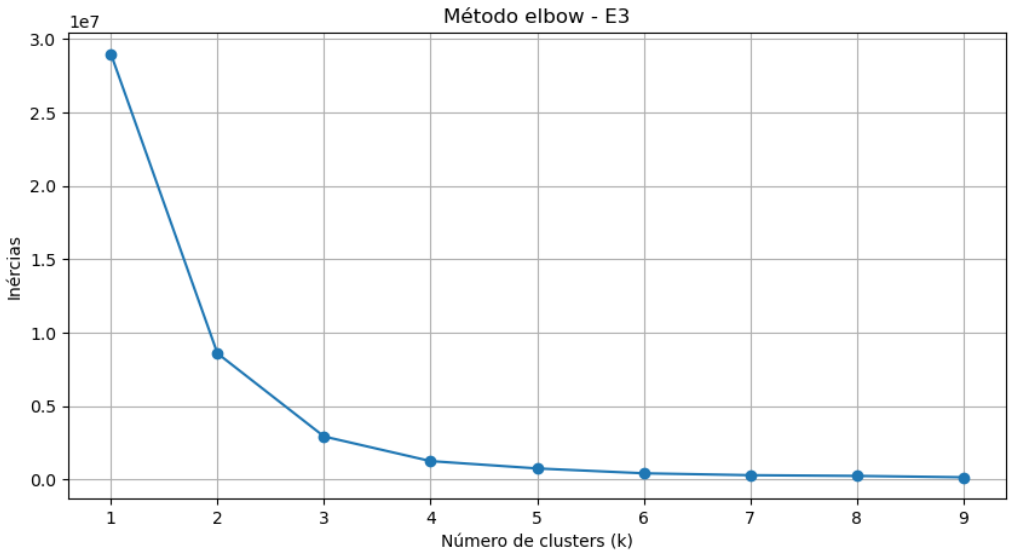
## 5.3 E3: SOMA DE INTERAÇÕES E MÁXIMA SEQUÊNCIA SEM INTERAÇÕES

O experimento 3 (E1) considera os somatórios das interações semanais (P1) e as máximas sequências de dias sem interações dos alunos com o Moodle (P3).

### 5.3.1 Cálculo do Número de Clusters

Conforme os experimentos anteriores, o número k de clusters é determinado pelo método elbow, apresentado na figura 15.

Figura 15 – Gráfico resultante do método elbow do E3



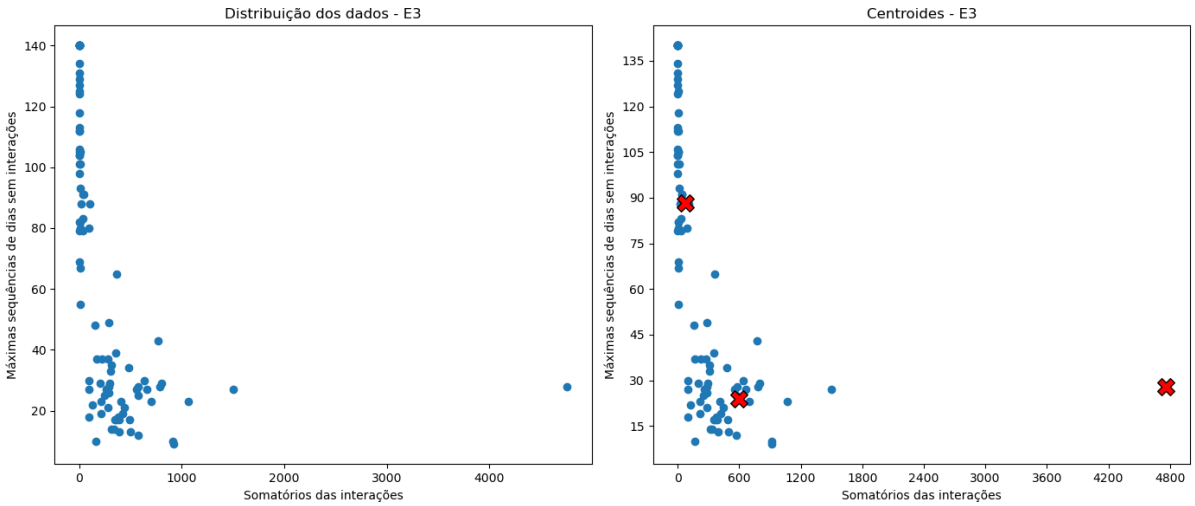

Fonte: Figura do autor (2024)

Assim, conforme o gráfico da figura 15, foi selecionado o valor de k=3.

### 5.3.2 Clusterização

Dessa forma, são definidos os centroides pelo algoritmo, mostrados na figura 16, em vermelho.

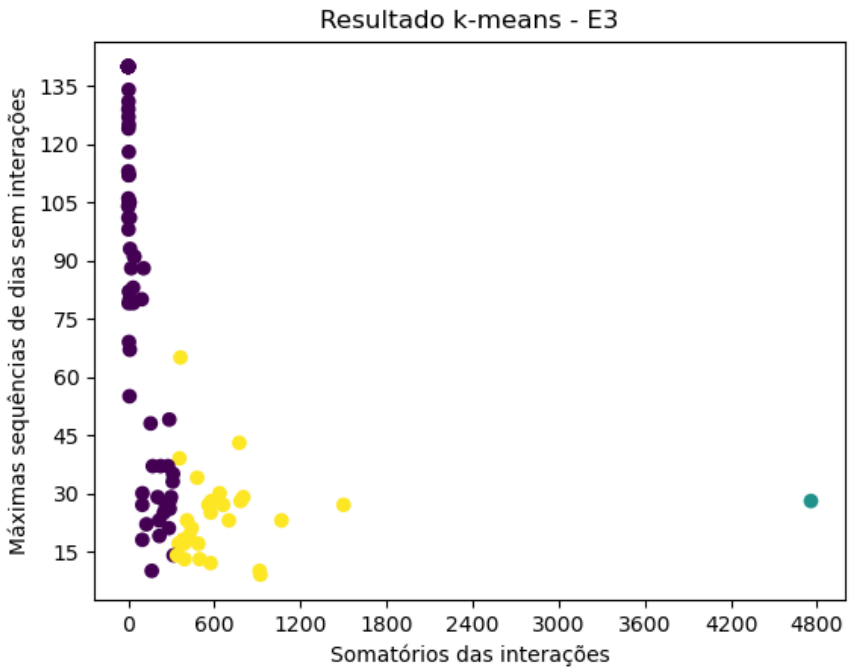
Figura 16 – Gráfico da distribuição dos dados e os centroides do E3




Fonte: Figura do autor (2024)

Com base nos centroides, a clusterização é gerada, conforme mostra a figura 17, onde os dados são divididos em 3 clusters.

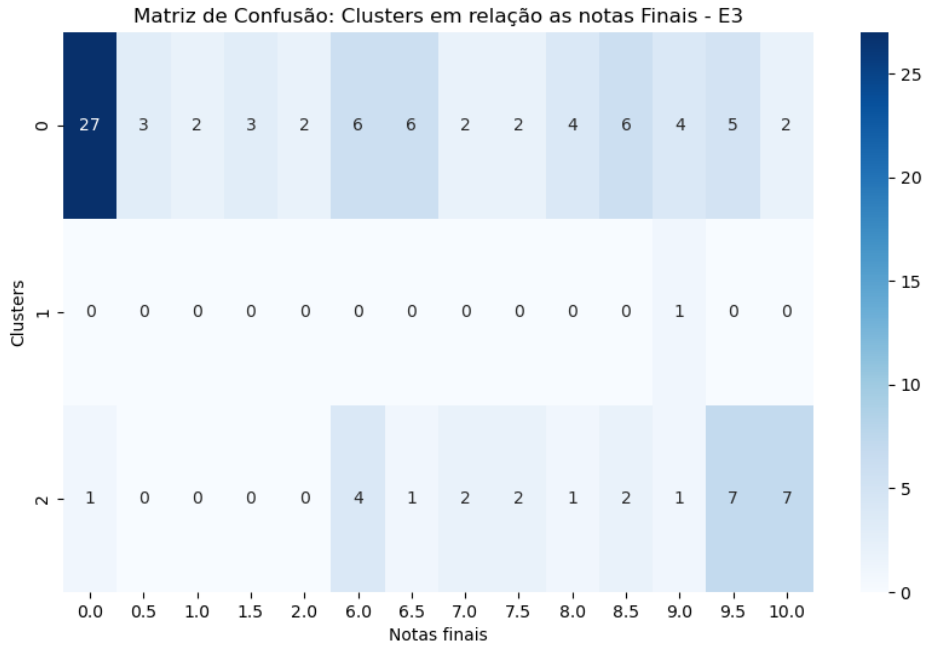
Figura 17 – Resultados do k-means do E3




Fonte: Figura do autor (2024)

Para verificar os resultados, utiliza-se a matriz de confusão da figura 18, que mostra a concentração de notas por cluster.

Figura 18 – Matriz de confusão do E3




Fonte: Figura do autor (2024)

Assim, analisando o percentual de notas acima e abaixo de 5, são obtidos os resultados da tabela 12.

| **Cluster** | **Notas: 0 - 5 (%)** | **Notas: 5 - 10 (%)** |
|---|---|---|
| Cluster 0 | 50,00 | 50,00 |
| Cluster 1 | 0 | 100 |
| Cluster 2 | 3,57 | 96,43 |

Tabela 12 – Tabela percentual da distribuição de notas por cluster do E3

### 5.3.3 Análise

No experimento 3 (E3), os alunos são agrupados em 3 clusters. Sendo assim, o Cluster 0 possui equilibrio entre a composição de estudantes com notas superiores e inferiores a 5, com 50% cada, porém, concentrando uma quantidade maior de estudantes com notas iguais a zero, manifestando um grupo com maior tendência ao risco. O Cluster 1, assim como no experimento 1 (E1), é composto por apenas um estudante, devido ao mesmo apresentar um somatório de interações significativamente superior aos demais, sugerindo baixo risco. Por fim, o Cluster 2 concentra majoritariamente os estudantes com notas superiores a 5, com aproximadamente 96%, apontando um baixo risco aos estudante desse cluster.

# 6 CONCLUSÃO

Através dos experimentos, foi possível observar que a contagem das interações com o AVA sugerem uma relação com o risco do estudante durante a disciplina cursada, porém, a classificação de risco nem sempre ocorre de maneira precisa. Parte da causa dessa imprecisão, está relacionada ao conjunto de dados utilizado ser relativamente pequeno para uma análise mais aprofundada sobre os padrões de comportamento que influenciam no risco do estudante, fator limitante na busca de melhores resultados.

Para trabalhos futuros, novas combinações entre interações podem ser consideradas, bem como levar em consideração os detalhes de cada interação. Além disso, classificar por período de tempo, como mensal e semanal, permitirá verificar as mudanças de comportamento dos alunos ao longo do período cursado. Experimentos com variações dos algortimos de agrupamento poderão fornecer diferentes resultados para uma validação mais robusta do método.

Assim, os experimentos mostram, apesar de suas limitações, que podem contribuir na identificação de estudantes em situação de risco, permitindo que instrutores e instituições de ensino verifiquem o progresso e também possam identificar dificuldades dos alunos, intervindo caso necessário, evitando reprovações ou até mesmo evasão.

# REFERÊNCIAS

AHMED, Mohiuddin; SERAJ, Raihan; ISLAM, Syed Mohammed Shamsul. The k-means algorithm: A comprehensive survey and performance evaluation. **Electronics**, MDPI, v. 9, n. 8, p. 1295, 2020.

BENAIMECHE, Mohamed Amine *et al.* A k-means clustering machine learning-based multiscale method for anelastic heterogeneous structures with internal variables. **International Journal for Numerical Methods in Engineering**, Wiley Online Library, v. 123, n. 9, p. 2012–2041, 2022.

BROWN, M. Learning Analytics: Moving from Concept to Practice (EDUCAUSE Learning Initiative Briefs). **Louisville (CO), EDUCAUSE**, 2012.

BUCOS, Marian; DRĂGULESCU, Bogdan. Student cluster analysis based on Moodle data and academic performance indicators. *In:* IEEE. 2020 International Symposium on Electronics and Telecommunications (ISETC). [*S.l.*: *s.n.*], 2020. P. 1–4.

CUNNINGHAM, Pádraig; CORD, Matthieu; DELANY, Sarah Jane. Supervised learning. *In:* MACHINE learning techniques for multimedia: case studies on organization and retrieval. [*S.l.*]: Springer, 2008. P. 21–49.

DE SOUSA, Luís Manuel Mota *et al.* Revisões da literatura científica: tipos, métodos e aplicações em enfermagem. **Revista portuguesa de enfermagem de reabilitação**, v. 1, n. 1, p. 45–54, 2018.

DONGARE, AD; KHARDE, RR; KACHARE, Amit D *et al.* Introduction to artificial neural network. **International Journal of Engineering and Innovative Technology (IJEIT)**, Citeseer, v. 2, n. 1, p. 189–194, 2012.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996.

GASEVIC, Dragan; DAWSON, Shane; PARDO, Abelardo. How do we start? Sate and directions of learning analyics adoption. International Council for Open e Distance Education, 2016.

GHAHRAMANI, Zoubin. Unsupervised learning. *In:* SUMMER school on machine learning. [*S.l.*]: Springer, 2003. P. 72–112.

HECKERMAN, David. A tutorial on learning with Bayesian networks. **Innovations in Bayesian networks: Theory and applications**, Springer, p. 33–82, 2008.

HOOSHYAR, Danial; HUANG, Yueh-Min; YANG, Yeongwook. A Three-Layered Student Learning Model for Prediction of Failure Risk in Online Learning. **Human-centric**

**Computing and Information Sciences**, KOREA INFORMATION PROCESSING SOC 1002HO YONGSUNGBIZTEL 314-1 2GA HANKANGRO ..., v. 12, 2022.

JOKSIMOVIĆ, Srećko; KOVANOVIĆ, Vitomir; DAWSON, Shane. The journey of learning analytics. **HERDSA Review of Higher Education**, v. 6, p. 27–63, 2019.

LIKAS, Aristidis; VLASSIS, Nikos; VERBEEK, Jakob J. The global k-means clustering algorithm. **Pattern recognition**, Elsevier, v. 36, n. 2, p. 451–461, 2003.

LU, Haohui; UDDIN, Shahadat. Unsupervised machine learning for disease prediction: a comparative performance analysis using multiple datasets. **Health and Technology**, Springer, v. 14, n. 1, p. 141–154, 2024.

MAIMON, Oded; ROKACH, Lior. Introduction to knowledge discovery and data mining. *In:* DATA mining and knowledge discovery handbook. [*S.l.*]: Springer, 2010. P. 1–15.

MATPLOTLIB.ORG. **Matplotlib – Visualization with Python**. 2024. Disponível em: `https://matplotlib.org/`.

MCKINNEY, Wes. **Python for data analysis**. [*S.l.*]: "O'Reilly Media, Inc.", 2022.

MEANS, Barbara *et al.* Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. Centre for Learning Technology, 2009.

NA, Shi; XUMIN, Liu; YONG, Guan. Research on k-means clustering algorithm: An improved k-means clustering algorithm. *In:* IEEE. 2010 Third International Symposium on intelligent information technology and security informatics. [*S.l.*: *s.n.*], 2010. P. 63–67.

NAEEM, Samreen *et al.* An unsupervised machine learning algorithms: Comprehensive review. **International Journal of Computing and Digital Systems**, University of Bahrain, 2023.

PANDAS.PYDATA.ORG. **pandas - Python Data Analysis Library**. 2024. Disponível em: `https://pandas.pydata.org/about/index.html`.

RAYASAM, Ajay Siva. **Predicting at-risk students from disparate sources of institutional data**. 2020. Tese (Doutorado) – Massachusetts Institute of Technology.

ROMERO, Cristobal; VENTURA, Sebastian. Data mining in education. **Wiley Interdisciplinary Reviews: Data mining and knowledge discovery**, Wiley Online Library, v. 3, n. 1, p. 12–27, 2013.

ROMERO, Cristobal; VENTURA, Sebastian. Educational data mining and learning analytics: An updated survey. **Wiley interdisciplinary reviews: Data mining and knowledge discovery**, Wiley Online Library, v. 10, n. 3, e1355, 2020.

SCIKIT-LEARN.ORG. **KMeans – scikit-learn 1.5.0 documentation**. 2024. Disponível em: `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans`.

_____. **scikit-learn: machine learning in Python**. 2024. Disponível em: `https://scikit-learn.org/`.

SEABORN.PYDATA.ORG. **seaborn: statistical data visualization**. 2024. Disponível em: `https://seaborn.pydata.org/`.

SHAFIQ, Dalia Abdulkareem *et al.* A Conceptual Predictive Analytics Model for the Identification of at-risk students in VLE using Machine Learning Techniques. *In:* IEEE. 2022 14th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS). [*S.l.*: *s.n.*], 2022. P. 1–8.

SIEMENS, George. Learning analytics: The emergence of a discipline. **American Behavioral Scientist**, Sage Publications Sage CA: Los Angeles, CA, v. 57, n. 10, p. 1380–1400, 2013.

SIEMENS, George; BAKER, Ryan SJ d. Learning analytics and educational data mining: towards communication and collaboration. *In:* PROCEEDINGS of the 2nd international conference on learning analytics and knowledge. [*S.l.*: *s.n.*], 2012. P. 252–254.

SIEMENS, George; GASEVIC, Dragan *et al.* **Open Learning Analytics: an integrated & modularized platform**. 2011. Tese (Doutorado) – Open University Press Maidenhead.

SINAGA, Kristina P; YANG, Miin-Shen. Unsupervised K-means clustering algorithm. **IEEE access**, IEEE, v. 8, p. 80716–80727, 2020.

SONI, Devin. **Supervised vs. Unsupervised Learning**. [*S.l.*: *s.n.*], 2018. urlhttps://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d. Acessado: 20 jun. 2023.

TAMADA, Mariela Mizota; GIUSTI, Rafael; NETTO, José Francisco de Magalhães. Predicting students at risk of dropout in technical course using LMS logs. **Electronics**, MDPI, v. 11, n. 3, p. 468, 2022.

UMARGONO, Edy; SUSENO, Jadmiko Endro; GUNAWAN, S. K-Means clustering optimization using the elbow method and early centroid determination based-on mean and median. *In:* SCITEPRESS—SCIENCE e TECHNOLOGY PUBLICATIONS

SETUBAL, PORTUGAL. PROCEEDINGS of the International Conferences on Information System and Technology. [*S.l.*: *s.n.*], 2019. P. 234–240.

VALKENBORG, Dirk *et al.* Unsupervised learning. **American Journal of Orthodontics and Dentofacial Orthopedics**, v. 163, n. 6, p. 877–882, 2023. ISSN 0889-5406. DOI: https://doi.org/10.1016/j.ajodo.2023.04.001. Disponível em: https://www.sciencedirect.com/science/article/pii/S0889540623001932.

WANG, Sun-Chong; WANG, Sun-Chong. Artificial neural network. **Interdisciplinary computing in java programming**, Springer, p. 81–100, 2003.

YANG, Yeongwook *et al.* Predicting course achievement of university students based on their procrastination behaviour on Moodle. **Soft Computing**, Springer, v. 24, p. 18777–18793, 2020.

YE, Dan. The history and development of learning analytics in learning, design, & technology field. **TechTrends**, Springer, v. 66, n. 4, p. 607–615, 2022.

YUAN, Chunhui; YANG, Haitao. Research on K-value selection method of K-means clustering algorithm. **J**, MDPI, v. 2, n. 2, p. 226–235, 2019.

# APÊNDICE A – TABELA DE ACRÔNIMOS

Quadro 1 – Modelo A.

| **Acrônimo** | **Significado** |
|---|---|
| AVA | Ambiente Virtual de Aprendizagem |
| DM | Data Mining |
| EDM | Educational Data Mining |
| KDD | Knowledge Discovery in Databases |
| LA | Learning Analytics |
| LMS | Learning Management Systems |
| MDE | Mineração de Dados Educacionais |
| ML | Machine Learning |
| UFSC | Universidade Federal de Santa Catarina |

Fonte: Elaborada pelo autor (2024)