UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Gustavo de Faria Silva

# Aplicação de técnicas de pré-processamento e agrupamento na base de dados do aplicativo michelzinho

**Uberlândia, Brasil**

**2023**

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Gustavo de Faria Silva

# Aplicação de técnicas de pré-processamento e agrupamento na base de dados do aplicativo michelzinho

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Ciência da Computação.

Orientador: Marcelo Zanchetta do Nascimento

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Ciência da Computação

Uberlândia, Brasil

2023

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

S586 Silva, Gustavo de Faria, 1997-
2023 Aplicação de técnicas de pré-processamento e agrupamento na base de dados do aplicativo michelzinho [recurso eletrônico] / Gustavo de Faria Silva. - 2023.

Orientador: Marcelo Zanchetta do Nascimento.
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Uberlândia, Graduação em Ciência da Computação.
Modo de acesso: Internet.
Inclui bibliografia.

1. Computação. I. Nascimento, Marcelo Zanchetta do, 1976-, (Orient.). II. Universidade Federal de Uberlândia. Graduação em Ciência da Computação. III. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

Gustavo de Faria Silva

# Aplicação de técnicas de pré-processamento e agrupamento na base de dados do aplicativo michelzinho

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Ciência da Computação.

Trabalho aprovado. Uberlândia, Brasil, 01 de fevereiro de 2023:

---

**Marcelo Zanchetta do Nascimento**
Orientador

---

**Elaine Ribeiro de Faria**

---

**Adilmar Coelho Dantas**

Uberlândia, Brasil
2023

# Agradecimentos

Agradeço a Deus pelo dom da vida.

Aos meus pais por terem me apoiado de todas as formas que conseguiram durante todo o curso e especialmente nessa etapa final.

À minha namorada por ceder em momentos que foram necessários para a conclusão desse trabalho.

À minha psicóloga que pode me ajudar a entender as tantas dificuldades psicológicas encontradas durante esse trabalho.

Ao meu orientador pela paciência e compreensão.

Por fim, a cada um que de alguma forma me incentivou para que esse curso fosse concluído.

# Resumo

O Transtorno do Espectro Autista (TEA) caracteriza-se pelo comprometimento na comunicação, na interação e na aprendizagem comportamental, apresentando ao paciente, além das dificuldades sociais, comportamento repetitivo e estereotipado. Essas características causam impactos no convívio social desses indivíduos, inclusive na detecção de expressões faciais e reconhecimento das emoções básicas, comprometendo o paciente do ambiente social. Porém, utilizando os jogos sérios no auxílio do ensino e no aprimoramento das habilidades, a psicologia tem alcançado bons resultados no tratamento. Desse modo, o aplicativo MICHELZINHO tem por objetivo, a detecção e o reconhecimento das emoções, culminando no desenvolvimento das habilidades de forma divertida, além de gerar uma robusta base de dados para transmitir aos especialistas em TEA, novos conhecimentos sobre a patologia. Assim, este trabalho foi desenvolvido com o objetivo de analisar a base de dados do aplicativo, utilizando a mineração de dados através de algoritmos computacionais. Para isso, foi desenvolvida uma ferramenta, em python, utilizada no pré-processamento dos dados, a fim de agrupá-los através de três diferentes algoritmos. Os resultados obtidos usando a silhueta simplificada, indicam que a ferramenta proposta contribuiu para o aprimoramento das habilidades de detecção e reconhecimento das emoções, apesar da alta dimensionalidade dos dados após o pré-processamento.

**Palavras-chave**: Autismo, Jogos sérios, Agrupamento, Pré-processamento.

# Abstract

Autistic Spectrum Disorder (ASD) is characterized by severe impairment in communication, interaction and behavioral learning, presenting the patient, in addition to social difficulties, with repetitive and stereotyped behavior. These characteristics impact the social interaction of these individuals, including the detection of facial expressions and recognition of basic emotions, disconnecting the patient from the social environment. However, using serious games to help teach and improve skills, psychology has achieved good results in treatment. In this way, the MICHELZINHO application aims to detect and recognize emotions, culminating in the development of skills in a fun way, in addition to generating a robust database to transmit new knowledge about the pathology to ASD specialists. Thus, this work was developed with the objective of analyzing the application's database, using data mining through computational algorithms. For this, a tool was developed, in python, used in the pre-processing of the data, in order to group them through three different algorithms. The results were validated using the simplified silhouette, indicating that the proposed tool contributed to the improvement of emotion detection and recognition skills, despite the high dimensionality of the data after pre-processing

**Keywords**: Autism; Serious games; Clustering; Preprocessing.

# Lista de ilustrações

Figura 1 – Etapas do processo de *KDD (Knowledge Discovery in Databases)* . . . 20
Figura 2 – Agrupamentos gerados via Kmeans e seus respectivos centroides.) . . . 28
Figura 3 – Conjunto com grupos circulares (conjunto 1), não circulares (conjuntos 2 e 3) e com ruídos (conjunto 3). . . . . 29
Figura 4 – Densidade baseada em centro . . . . . 30
Figura 5 – Agrupamento DBSCAN para 3.000 pontos bidimensionais. . . . . 31
Figura 6 – Estrutura do neurônio biológico (sinapse). . . . . 32
Figura 7 – Estrutura do neurônio artificial (sinapse). . . . . 33
Figura 8 – Rede Neural Artificial de 2 camadas com 4 entradas e 2 saídas. . . . . 33
Figura 9 – O Modelo de Kohonen . . . . . 35
Figura 10 – Etapas do método proposto . . . . . 38

# Lista de tabelas

Tabela 1 – Base de dados do aplicativo MICHELZINHO® . . . . . 40
Tabela 2 – Agrupamento K-Means . . . . . 45
Tabela 3 – Evolução da largura da silhueta versus o número de grupos (k) . . . . 46
Tabela 4 – Bases com mais e menos clusters no algoritmo DBSCAN . . . . . 46
Tabela 5 – Bases com maior e menor percentual agrupado . . . . . 47
Tabela 6 – Correlação da largura da silhueta versus do percentual agrupado (%) . 48
Tabela 7 – Silhueta Relativa Média versus Eps . . . . . 48
Tabela 8 – Silhueta Relativa Média versus MinPts . . . . . 49
Tabela 9 – Silhueta Relativa Média versus MinPts . . . . . 50
Tabela 10 – Bases com maior e menor silhueta no Mapa de Kohonen . . . . . 51
Tabela 11 – Influência da variação do Sigma na silhueta . . . . . 52
Tabela 12 – Influência da variação da taxa de aprendizado na silhueta . . . . . 52
Tabela 13 – Evolução da silhueta em relação ao número de grupos . . . . . 53

# Lista de Algoritmos

1 Kmeans . . . . . 28
2 DBSCAN . . . . . 30
3 SOM . . . . . 34

# Sumário


**1 INTRODUÇÃO . . . . . 12**
**1.1 Objetivos . . . . . 14**
1.1.1 Objetivo geral . . . . . 14
1.1.2 Objetivos específicos . . . . . 14

**2 FUNDAMENTAÇÃO TEÓRICA . . . . . 15**
**2.1 Introdução da Fundamentação . . . . . 15**
**2.2 As Emoções . . . . . 15**
**2.3 Transtorno do espectro autístico . . . . . 16**
**2.4 Jogos Interativos . . . . . 17**
**2.5 Michelzinho . . . . . 19**
**2.6 Descoberta de Conhecimento em Bases de Dados . . . . . 19**
**2.7 Pré-processamento de Dados . . . . . 21**
2.7.1 Seleção de Dados . . . . . 21
2.7.1.1 Seleção de Atributos . . . . . 21
2.7.1.2 Seleção de Registros . . . . . 21
2.7.2 Limpeza de Dados . . . . . 22
2.7.3 Integração de dados . . . . . 22
2.7.4 Transformação de dados . . . . . 22
**2.8 Mineração de Dados . . . . . 23**
2.8.1 Classificação . . . . . 24
2.8.2 Regressão . . . . . 24
2.8.3 Regras de Associação . . . . . 24
2.8.4 Agrupamento . . . . . 25
2.8.4.1 Métodos Particionais . . . . . 26
2.8.4.2 Métodos baseados em densidade . . . . . 28
2.8.4.3 Métodos baseados em rede neurais artificiais . . . . . 31
**2.9 Pós-processamento . . . . . 35**
2.9.1 Interpretação . . . . . 36
2.9.2 Validação de Dados . . . . . 36
**2.10 Considerações finais . . . . . 37**

**3 DESENVOLVIMENTO . . . . . 38**
**3.1 Introdução . . . . . 38**
**3.2 Coleta e Tratamento de Dados . . . . . 39**
3.2.1 Apresentação da base de dados do MICHELZINHO . . . . . 39

3.2.2 Pré-processamentos Realizados . . . . . 41
3.2.3 Base de dados gerada . . . . . 42
**3.3 Métodos de Agrupamento e Medidas de Validação utilizados . . . . 42**
**3.4 Considerações Finais . . . . . 43**

**4 RESULTADOS . . . . . 44**
**4.1 Introdução . . . . . 44**
**4.2 k-means . . . . . 44**
**4.3 DBSCAN . . . . . 46**
4.3.1 Experimento 1 . . . . . 47
4.3.2 Experimento 2 . . . . . 48
4.3.3 Experimento 3 . . . . . 49
4.3.4 Experimento 4 . . . . . 49
**4.4 Mapas de Kohonen . . . . . 50**
4.4.1 Experimento 5 . . . . . 51
4.4.2 Experimento 6 . . . . . 51
4.4.3 Experimento 7 . . . . . 52
**4.5 Considerações Finais . . . . . 53**

**5 CONCLUSÃO . . . . . 55**
**5.1 Dificuldades Encontradas . . . . . 55**
**5.2 Principais Contribuições . . . . . 56**
**5.3 Trabalhos Futuros . . . . . 56**

**REFERÊNCIAS . . . . . 57**

# 1 Introdução

De acordo com Ekman (2011, p. 37) “as emoções nos preparam para lidar com eventos importantes sem precisarmos pensar no que fazer” alicerçadas no fato dessas serem reações às questões que parecem essenciais para o nosso bem-estar. Assim, Pinto (2001) afirma que a concepção de emoção traduz-se em uma experiência subjetiva que envolve toda a mente e o corpo, sendo uma reação complexa desencadeada por um estímulo que envolve reações orgânicas e sensações pessoais.

As emoções, expressões faciais naturais dos seres humanos, permitem demonstrar os sentimentos, contribuindo para as interações sociais e a comunicação interpessoal, essenciais para a interatividade entre as pessoas (DANTAS e NASCIMENTO, 2022). Os autores afirmam também que as emoções como uma representação universal entre os seres humanos de diferentes grupos sociais e etnias. Baseado em Ekman, Friesen e Hager (2002), uma teoria denominada Sistema de Codificação de Ação Facial (FACS) surgiu para a descrição de todos os movimentos faciais por meio de quarenta e quatro representações denominadas Unidades de Ação (UA). Essas UAs possuem um código numérico que descreve o grupo de músculos da face ativados na formação das diferentes mímicas faciais relativas às emoções, permitindo aprimorar as habilidades emocionais dos seres humanos.

Indivíduos portadores do Transtorno do Espectro Autístico (TEA) possuem dificuldade no processo de iniciação e manutenção de relacionamento social com outras pessoas. Um dos principais fatores que podem comprometer esse processo é a capacidade de reconhecer e expressar as emoções, o que pode afetar de forma negativa as interações sociais (BARON-COHEN et al., 2005). Oliveira e Sertié (2017) definem o TEA como uma limitação neurológica, que causa deficiência no processo de comunicação e socialização, que possui ocorrência precoce e sintomas que variam conforme o nível de intensidade traduzido por comportamentos considerados atípicos.

A Organização Mundial da Saúde (OMS), afirma que uma em cada 160 crianças apresenta características relacionadas ao TEA, no entanto segundo Christensen et al. (2018) a estimativa do Autism and Developmental Disabilities Monitoring (ADDM), nos Estados Unidos, os casos de TEA em crianças de 8 anos aumentou de 1/150 em 2000 para 1/68 em 2012. No Brasil, com base no Censo, realizado no ano de 2010, da população de 18,7 milhões de crianças de 0 a 5 anos (FUNDAÇÃO ABRINQ, 2021), estima-se que aproximadamente 117 mil crianças, nessa faixa etária, podem ter características que estão relacionadas ao TEA. Já quando a referência é de idades superiores, no Brasil, a estimativa atinge o expressivo número de 500 mil autistas, conforme Gomes et al. (2015).

Desse modo, o Ministério da Educação (2011) reforçou o formato educacional para o aluno nessa condição, com a publicação do Decreto 7.611/11, em reforço ao Decreto 6.571/08 (revogado), consolidando a pauta do ensino na garantia dos direitos humanos, no respeito às individualidades e na igualdade de oportunidades, sendo que a ideia principal é minimizar as dificuldades e limitações que o aluno portador de qualquer deficiência possa apresentar durante o aprendizado em sala de aula, através do uso de recursos, serviços e estratégias diferenciadas que possam atender às necessidades de aprendizagem.

Nessa visão, Dantas (2020, p. 19) em seu documento de pesquisa de doutorado afirma que ferramentas computacionais podem facilitar a análise de grandes cargas de trabalho. Afirma ainda que o desenvolvimento de abordagens com essas características permitem ao indivíduo treinar e aperfeiçoar habilidades em diferentes ambientes além do tradicional consultório.

Silva, Ramos e Ribeiro (2019) afirmaram que as tecnologias digitais, de forma revolucionária, têm promovido, nos dias atuais, contribuições para as transformações em relação ao tempo, à cultura e à cognição humana. O desenvolvimento de técnicas computacionais é capaz de contribuir para o ensino das habilidades emocionais em indivíduos com quaisquer tipos de deficiência intelectual (DANTAS, 2022).

A fim de gerar essas transformações, de acordo com Ramos e Cruz (2018) o uso de jogos digitais em contextos educativos, denominados jogos sérios, remetem às experiências que podem influenciar os domínios do desenvolvimento humano, aprendizagem, sociabilidade e subjetividade. Conforme Ramos (2020) o uso dos jogos, através de atividades bem delineadas e mediadas, pode gerar contribuições relevantes ao campo do desenvolvimento das emoções em sentidos de intervenção educacional. Nesse contexto, Dantas (2022) desenvolveu o software MICHELZINHO[1], um jogo sério aberto, com características voltadas às habilidades educacionais, para aprimorar a capacidade de reconhecer e expressar emoções em indivíduos com quaisquer tipos de deficiência intelectual, inclusive o TEA. Esse sistema tem contribuído para o aprimoramento das habilidades em pessoas com limitações relacionadas ao reconhecimento e às expressões das emoções.

O software possui módulo de coleta de dados dos usuários para análise quantitativa do aprimoramento no ensino das emoções, o qual gera uma grande quantidade de informações em relação às partidas executadas como o jogo sério. Durante cada partida, as características relacionadas aos músculos da região da face são capturadas e armazenadas em uma base de dados. Atualmente, o número de eventos (músculos ativados) para a representação das emoções não são explorados em relação às informações e tendências do comportamento dos voluntários numa partida do jogo proposto. O uso de técnicas de mineração de dados pode obter informações a partir de um grande conjunto dos mesmos. Entre as técnicas, tem-se que parte dos algoritmos que são empregados na tarefa de agrupamento busca organizar um conjunto de dados em vários grupos, de modo que,

por meio de métricas de similaridade, itens semelhantes pertençam ao mesmo grupo. Essa organização pode mostrar padrões que contribui com o especialista numa tomada de decisão em relação às recomendações das sessões de tratamento de pacientes ou análise do melhoramento em relação à terapia estabelecida.

## 1.1 Objetivos

### 1.1.1 Objetivo geral

O presente trabalho traz uma investigação de diferentes algoritmos de agrupamento com o propósito de caracterizar grupos nos dados relativos às representações das emoções de partidas do jogo sério Michelzinho®.

### 1.1.2 Objetivos específicos

Em um formato mais pontual pretende-se nesta monografia:

- Construir um banco de dados com as informações capturadas da ferramenta de reconhecimento das emoções por meio do aplicativo Michelzinho®;
- Implementar os algoritmos de agrupamento de dados: K-Means, DBSCAN e Mapas de Kohonen;
- Analisar o desempenho dos algoritmos com as métricas de agrupamento; e
- Fornecer grupos de dados relativos às representações que poderão ser empregados em algoritmos para classificação ou seleção de características.

# 2 Fundamentação Teórica

## 2.1 Introdução da Fundamentação

Observando o cenário apresentado no capítulo anterior, neste capítulo são apresentados os conceitos teóricos necessários à construção desta pesquisa, tratando os principais assuntos relacionados, as emoções, os jogos interativos na aprendizagem, o TEA, o aplicativo MICHELZINHO® e o agrupamento de dados, com maior ênfase aos conceitos da mineração de dados utilizando os algoritmos aplicados no desenvolvimento do trabalho.

A seção 2.2 fornece uma visão geral sobre as emoções. A seção 2.3 discorre sobre o TEA – Transtorno do Espectro Autístico. A seção 2.4 apresenta os tipos de benefícios prestados pelos jogos interativos na aprendizagem, enquanto a seção 2.5 delineia o aplicativo de jogo interativo MICHELZINHO® e suas aplicabilidades. A seção 2.6 explica o processo de KDD e cada etapa do mesmo. A seção 2.7 detalha a etapa de pré-processamento dentro do processo de KDD, assim como algumas técnicas comumente usadas. A seção 2.8 traz o conceito de mineração de dados também dentro do processo de KDD, mostrando suas tarefas, em especial a de agrupamento, pois é a tarefa que será utilizada neste estudo. A seção 2.9 apresenta o pós-processamento e suas diferentes medidas para validação de agrupamento. Por último, a seção 2.10 traz as considerações finais.

## 2.2 As Emoções

Segundo Dantas (2020), as emoções podem ser interpretadas como uma reação tipicamente breve e intensa, relacionada a um determinado fato ou evento. Em um formato mais biológico, (LENT, 2013, p. 254) diz que “a emoção pode ser definida como um conjunto de reações químicas e neurais, subjacentes à organização de certas respostas comportamentais básicas e necessárias à sobrevivência dos animais”.

Assim, Melo (2005) diz que a ativação de uma emoção tem como propósito a preparação do organismo para o ato de adaptação, a fim de atingir o bem-estar que o ser humano necessita, gerando como consequências imediatas, alterações no estado corporal e nas estruturas cerebrais. A autora afirma ainda que algumas das emoções são consideradas como sendo emoções básicas, por serem caracterizadas por uma programação inata, onde estamos a falar das emoções tais como a tristeza, alegria, raiva, medo, surpresa e nojo. Dantas (2020) ainda afirmou que “a habilidade de reconhecimento das expressões faciais que demonstrem as emoções básicas é importante nas relações interpessoais de indivíduos em uma sociedade”, sendo que “os indivíduos neurotípicos têm essa habilidade aprimorada

de maneira natural e gradativa durante o ciclo da vida".

Dentre os inúmeros estudos que contribuem em relação ao uso de técnicas para análise das emoções, (NASS; BRAVE, 2007) cita as 4 (quatro) principais estratégias: as respostas neurológicas, as atividades autonômicas, as expressões faciais e a voz. No entanto, em seu trabalho, Dantas (2020) cita que é dada maior ênfase às expressões faciais, sendo essas capazes de proporcionar resultados mais expressivos no processo de reconhecimento das emoções. O autor ainda afirma que existem técnicas de reconhecimento de emoções usando algoritmos computacionaisl, sensores e sinais biológicos sendo que o uso de tecnologias tem contribuído na análise das emoções em determinados casos.

Os autores (Ramos, Silva e Macedo, 2020) afirmam que "as emoções estão relacionadas à afetividade que em contextos educacionais pode influenciar o processo de aprendizagem". Ainda afirmam que "as emoções emergem na interação com os jogos digitais, pois criam experiências e situações em que o jogador precisa tomar decisões e sofre as consequências boas e ruins de suas ações".

## 2.3 Transtorno do espectro autístico

Conforme (ALMEIDA et al., 2018, p. 73), o Transtorno do Espectro Autista (TEA) "é um dos transtornos do neurodesenvolvimento mais prevalentes na infância" sendo caracterizado pelo comprometimento de dois domínios centrais: o déficit na comunicação social e interação social como também por padrões repetitivos e restritos de comportamento, interesses ou atividades.

O termo autismo vem do grego "autos", que significa "de si mesmo". Kanner, em 1943, descreveu um grupo de crianças que apresentavam inabilidade para se relacionarem com outras pessoas, tendência ao isolamento, falha no uso da linguagem para a comunicação e uma necessidade extrema de manter-se na "mesmice" (SAMPAIO; FREITAS, 2011).

Segundo Almeida et al. (2018) o TEA tem como principal característica a falha de comunicação não verbal, variando desde a total falta de expressão facial até a inexistência de integração entre gestos e falas. Indivíduos com TEA, de acordo com Baptista e Bosa (2002), possuem dificuldade no processo de iniciar e manter um relacionamento social com outras pessoas, devendo seu desenvolvimento intelectual ocorrer de forma especial.

Com uma proporção entre meninos e meninas de 4:1, conforme afirma Gomes et al. (2015), o autismo se traduz em condições neurológicas que aparecem precocemente na infância, geralmente antes dos três anos de idade e afetam o desenvolvimento pessoal, social, acadêmico e/ou profissional do indivíduo. De acordo com Benute (2020) os fatores biológicos e genéticos são os responsáveis pelo TEA devido os mecanismos epigenéticos se

mostrarem agentes importantes, porém a autora não descarta as desordens hereditárias como fatores explicativos sobre a ocorrência do TEA.

A observação precoce de problemas ou anormalidades que possam acontecer no desenvolvimento da criança pode ser decisiva em seu futuro, pois quanto mais cedo a ocorrência de diagnóstico do TEA, melhor será a qualidade de vida do autista. Assim, Souza et al. (2004) afirmam que o diagnóstico exige análise e verificação de déficits característicos de comunicação social, comportamentos excessivamente repetitivos, interesses restritos e insistência nas mesmas coisas.

O Manual Diagnóstico e Estatístico de Transtornos Mentais, o DSM 5, elaborado pela "*American Psychiatric Association*" (2014) classifica o TEA em 3 (três) níveis, sendo o nível 1, tido como leve, onde tem-se indivíduos que requerem suporte, o nível 2, como moderado, onde tem-se indivíduos que precisam de suporte substancial e o nível 3, o severo, relacionado a pessoas que requerem suporte muito substancial.

Howes et al. (2017) afirmam que nos dias atuais não há tratamento farmacológico para os sintomas centrais do TEA, sendo que através de estudos em andamento houve apenas a geração de evidências limitadas. Assim, tratamentos alternativos, mesmo diante da complexidade do TEA, são, segundo Soorya et al. (2018), benéficos como a inclusão da criança em equipes de reabilitação interdisciplinar, compostas por especialistas de várias áreas, inclusive psicopedagogos e terapeutas comportamentais e psicológicos.

Dificuldades em uma educação comum, em indivíduos com TEA, se dão, de acordo com Dantas (2020, p.18), devido à "diferença das intensidades cerebrais e das regiões ativadas durante o processo de reconhecimento e expressões das emoções" tornando o ensino um grande desafio, já que há limitação de respostas nas habilidades interativas do indivíduo (Dapogny et al. 2018). Várias ferramentas são propostas para auxiliar os especialistas no ensino dessas emoções aos indivíduos com TEA, como soluções baseadas em jogos sérios com o objetivo de ensinar habilidades.

Segundo os autores Metri,Ghorpade e Butalia (2012), a tecnologia tem proporcionado o surgimento de diversas propostas objetivando melhorar as interações sociais no reconhecimento de emoções. As ferramentas computacionais são desenvolvidas explorando recursos multimídia, figuras planas ou jogos (LACAVA et al. (2007). Dessa forma, Dantas (2020, p. 18) afirma que "há uma quantidade de jogos sérios desenvolvidos para auxiliar no aprimoramento das habilidades emocionais e sociais dos indivíduos com TEA".

## 2.4 Jogos Interativos

Conforme Fernandes (2010, p. 9) "os jogos sempre fizeram parte da vida do ser humano e desde os primeiros anos de vida os jogos e as brincadeiras são mediadores

da criança na sua relação com as coisas do mundo". Ao serem utilizados no ambiente escolar, proporcionam vantagens no processo de ensino, tais como motivação à criança, prazer e realização rumo aos objetivos, mobilização mental estimulando a forma de pensar, ordenação do tempo e espaço, integração da personalidade (afetiva, social, motora e cognitiva) e desenvolvimento de diversas habilidades: coordenação, disciplina, senso de responsabilidade, senso de justiça, iniciativa pessoal e coletiva.

Essas ferramentas quando empregadas no contexto educacional são um recurso didático que contém características que podem trazer benefícios para as práticas de ensino e aprendizagem. No entanto, para serem corretamente utilizadas, devem possuir objetivos bem definidos, a fim de promover o desenvolvimento de estratégias ou habilidades importantes na ampliação da capacidade cognitiva e intelectual do aluno disse Savi e Ulbricht (2008). Os jogos conseguem inserir diversos contextos na mente do aluno, devido à adaptação de várias situações do mundo real, gerando maior estímulo na aprendizagem do que outras técnicas que possam vir a ser empregadas (LI et al , 2012).

Os tempos mudaram, a sociedade evoluiu e as crianças de hoje são muito diferentes, pois nasceram em plena revolução tecnológica, com o computador dentro de casa, mas ainda gostando de brincar (RIVA 2009).

O autor Fernandes, (2010, p. 13) afirma que "um dos interesses das crianças modernas são os jogos digitais, que invadem o nosso dia a dia, e eles são das mais diversas formas e com as mais diferentes finalidades e propostas de entretenimento", que segundo Menezes (2003) possuem desafios a serem vencidos e situações dinâmicas que vão sendo apresentadas ao jogador no decorrer do jogo.

Dentre as vantagens dos jogos digitais no ensino, Prensky (2003) cita a capacidade de proporcionar melhoria na atenção seletiva visual do usuário, aliando a tecnologia ao profissional de educação na execução do processo de aprendizagem. As tecnologias digitais para a construção de jogos interativos possuem uma contribuição relevante no processo de aprendizagem de indivíduos com TEA TANG; HANNEGHAN; RHALIBI, 2009).

O autor Dantas (2020, p. 47) afirma:

> "Segundo dados estatísticos já apresentados anteriormente por órgãos como o Centers for Disease Control and Prevention e a Organização Mundial da Saúde mostram que existe uma grande parcela da população que apresenta sintomas ligados ao TEA. Também há um número cada vez menor de políticas públicas voltadas para o desenvolvimento de ferramentas e tratamentos para esse público. Diversas pesquisas já realizadas demonstraram que os indivíduos com TEA possuem limitações nas habilidades para reconhecimento e expressão das emoções, o que pode comprometer as interações sociais".(DANTAS, 2020, p. 50).

Assim, o autor Dantas (2020, p.47) propõe o desenvolvimento de um jogo da classe sério a fim de auxiliar no ensino de competências emocionais e sociais em pessoas com TEA, o qual foi denominado MICHELZINHO.

## 2.5 Michelzinho

De acordo com Pessini et al. (2014) os jogos sérios têm como objetivo principal a educação através da utilização da diversão e do envolvimento dos jogadores para prover uma experiência educativa. Os jogos sérios têm sido utilizados de forma crescente, explorando diversas práticas pedagógicas a fim de desenvolver o raciocínio e o aprendizado. Com isso, Dantas (2020) propôs o desenvolvimento do aplicativo MICHELZINHO que permite desenvolver as habilidades de reconhecer e expressar as emoções básicas: alegria, tristeza, raiva, desgosto, surpresa e medo, por meio da teoria de sistema de codificação de ação facial e técnicas de processamento digital de imagens.

O jogo possui personagens, cenários e animações modelados em 3D com características próximas à face humana, sendo a arquitetura do aplicativo desenvolvida para utilização em dispositivos móveis, além da computação em nuvem permitir reconhecimento em diversos modelos e configurações de plataformas. Outra importante afirmativa citada por Dantas (2020, p. 81) é:

> "Essa ferramenta ainda fornece um módulo que permite obter relatórios personalizados, construídos a partir de levantamentos com profissionais que atuam com indivíduos com TEA. Esses relatórios possibilitam a exibição de detalhes sobre cada intervenção realizada e a personalização das seções para cada indivíduo, com base nessas informações. Essas informações são essenciais para os profissionais acompanharem a evolução do quadro clínico dos participantes envolvidos".

Assim, com intuito de cumprir os objetivos da pesquisa, a mineração de dados pode avaliar diferentes algoritmos de agrupamentos, buscando a validação dos dados minerados, oriundos dos relatórios do jogo, buscando encontrar informações relevantes sobre os dados em relação ao comportamento dos usuários.

## 2.6 Descoberta de Conhecimento em Bases de Dados

A teoria do refinamento dos dados, com seu início na descoberta de conhecimento em Base de Dados ou *Knowledge Discorvery in Databases (KDD)*, de acordo com Fayyad; Piatetsky-Shapiro e Smith (1996) trata-se de um processo interativo para identificar nos

dados, novos padrões que sejam válidos, potencialmente úteis e interpretáveis. Esse processo ocorre com várias etapas operacionais, conforme a Figura 1.

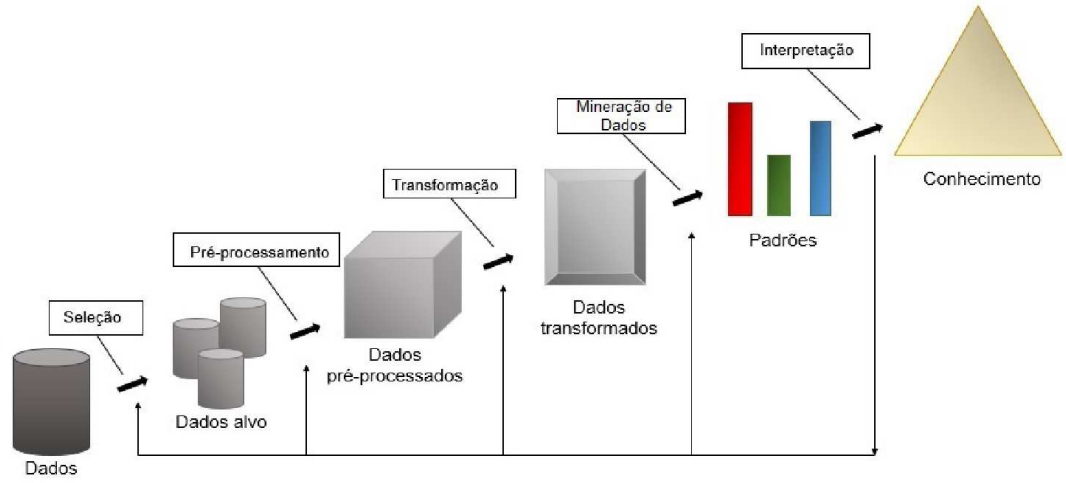



Figura 1 – Etapas do processo de *KDD (Knowledge Discovery in Databases)*. Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Ainda com baseado nos autores (Fayyad; Piatetsky-Shapiro e Smith (1996), enumera-se as seguintes etapas:

- Seleção de Dados: prevê a coleta e seleção dos dados;
- Limpeza: prevê a análise dos dados coletados, verificando a existência de ruídos, tratamento de valores ausentes, entre outras;
- Transformação ou Enriquecimento dos Dados: dedica-se à incorporação/criação de novos dados a partir dos já existentes;
- Mineração de Dados: consiste na aplicação de um algoritmo que, efetivamente, procura por padrões/relações e regularidades, em um determinado conjunto de dados;
- Interpretação e Avaliação: verifica a qualidade do conhecimento (padrões) descoberto, procurando identificar se o mesmo auxilia a resolução do problema original que motivou a realização do processo KDD.

Essas etapas, podem ser agrupadas em três grandes grupos: o pré-processamento de dados, a mineração de dados (do inglês, Data Minning-DM) e o pós-processamento, conforme Chiu e Tavella (2008).

## 2.7 Pré-processamento de Dados

### 2.7.1 Seleção de Dados

De acordo com os autores em Fayyad, Piatetsky-Shapiro e Smyth (1996), a etapa de pré-processamento tem a tarefa de selecionar os dados, definindo-os em subconjuntos ou amostras, para que efetivamente sejam considerados durante a KDD. Já os autores Goldschmidt e Passos (2015) afirmam que existem duas abordagens distintas para essa etapa: a seleção de atributos e a seleção de registros, os quais são discutidos nas próximas seções.

#### 2.7.1.1 Seleção de Atributos

A seleção de atributos permite a ordenação de atributos segundo algum critério de importância, a redução da dimensionalidade do espaço de busca de atributos e a remoção de dados contendo ruídos entre outros (LEE, 2005). Dessa forma, os autores Tan et al. (2006) afirmam que características redundantes ou irrelevantes ao KDD podem ser desconsideradas, representando assim um importante papel na fase de pré-processamento. Assim, um subconjunto, representativo e menor que o original, pode ser selecionado, inclusive em casos nos quais a medição é custosa.

Complementando, o autor Lee (2005) ainda diz que "como resultado da realização da seleção de atributos, a qualidade dos dados pode ser melhorada" tornando os dados mais compreensíveis durante o processo de mineração, além do "possível aperfeiçoamento ou não deterioração dos algoritmos de aprendizado, fazendo com que o processo de mineração de dados seja mais rápido". Como exemplo, o autor Giacomelli, (2020), cita como informações redundantes o raio e o diâmetro de um círculo, acrescentando, que "o número de matrícula e o sexo do aluno são atributos irrelevantes, quando se deseja avaliar o desempenho médio de uma turma em determinada disciplina".

#### 2.7.1.2 Seleção de Registros

Segundo Strodl et al. (2006), essa etapa corresponde à seleção de amostras de registros resultando em conjuntos menores que podem ser melhores utilizados para os algoritmos de *KDD*. A seleção de registro, de acordo com Giacomelli (2020), p. 24 "é empregada quando não é possível trabalhar com todo o conjunto de dados ou quando se deseja otimizar a análise da mineração de dados quanto ao tempo ou custo". Há ainda a contextualização feita por GOLDSCHMIDT e PASSOS (2017), em que é afirmado que em situação de inviabilidade do uso de todo o conjunto de dados por causa da ausência de algum atributo ou por algum outro motivo, então um subconjunto dos registros será empregado para compor o conjunto de dados a ser analisado.

Já Zuege (2018) afirma que a redução dos dados consiste em reduzir o número de objetos ou atributos do conjunto, buscando selecionar os atributos desejáveis, descartando os irrelevantes, comprimindo-os, reduzindo o quantitativo de objetos ou tornando-os discretos com objetivo de gerar maior eficiência no processo de mineração de dados. Para uma boa redução dos dados, o autor (Barros (2011, p. 22) afirma que "é necessário conhecer como esses dados estão organizados, qual a relevância de cada atributo e também qual a relação entre os atributos".

Silva (2014), afirma que as técnicas de redução de dados podem ser aplicadas para que a massa de dados original seja convertida em uma massa de dados menor, sem perder a representatividade dos dados originais. Tsaí (2012) ainda enfatiza que ao se realizar a redução do número de dimensões em um conjunto de dados, espera-se uma melhor visualização desse conjunto, além do encontro de estruturas ocultas que possam auxiliar em uma melhor compreensão analítica.

### 2.7.2 Limpeza de Dados

A limpeza de dados (do inglês, "*Data Cleanning*- DC), também conhecida como "*Scrubbing*", é um conjunto de técnicas utilizadas para detectar e remover anomalias em bases de dados (RAHN & DO, 2000).

A limpeza de dados pode ser abordada de duas formas: a especializada e a genérica. A primeira é utilizada sobre um determinado campo de domínio da base de dados, enquanto a segunda é utilizada de forma sistemática cobrindo um campo mais vasto de problemas, adaptando-se a diferentes domínios, desconhecidos do analista (ALMEIDA et al, 2016). Ambas dependem de técnicas que automatizam, de forma total ou parcial, as tarefas de detecção e correção das (RAHN & DO, 2000).

### 2.7.3 Integração de dados

Conforme Machado (2004), a integração de dados trata-se da forma com que os dados são convertidos, padronizados em um único formato e armazenados na base de dados original ou transacional. Esse processo de integração dos dados é necessário, por serem os mesmos, oriundos de fontes heterogêneas tais como banco de dados, arquivos textos, planilhas, "*data warehouses*"(estoque de dados), vídeos, imagens etc. De acordo com Silva (2014, p.23), é "necessária uma análise aprofundada dos dados observando redundâncias, dependências entre as variáveis e valores conflitantes"

### 2.7.4 Transformação de dados

Hoglin et al. (1983) definem a transformação dos dados como sendo um processo que efetua uma transformação matemática da caraterística original para uma nova ca-

racterística, que pelo menos se aproxime de uma distribuição normal. Conforme descreve Hair et al. (2010), a necessidade de transformação dos dados tem sua justificativa nas razões teóricas e na natureza dos dados ou suas derivações, a fim de corrigir violações das suposições estatísticas das técnicas, melhorar as correlações entre as variáveis e corrigir possíveis erros relacionados à entrada de dados. Silva (2014) ainda afirma que não existe critério único para transformação dos dados, podendo várias técnicas serem utilizadas, de acordo com os objetivos pretendidos. Dentre as técnicas empregadas estão a suavização, o agrupamento, a generalização, a normalização e a criação de novos atributos a partir de outros já preexistentes. No entanto, segundo Hoaglin et al. (1983), nem sempre uma transformação matemática produz os resultados esperados, como por exemplo, a normalização dos dados.

## 2.8 Mineração de Dados

Conforme Carvalho (2005) vários motivos definem a necessidade da mineração de dados, tais como:

- O volume de dados disponível atualmente é enorme;
- Os dados estão sendo organizados;
- Os recursos computacionais estão cada vez mais potentes;
- A competição empresarial demanda técnicas mais modernas de decisão; e
- Programas comerciais de mineração de dados já podem ser adquiridos.

A DM é a principal etapa dentro do processo de KDD, consistindo na aplicação de algoritmos de análise e descoberta de dados, para a extração de padrões (FAYYAD et al., 1996). Hand (2007) afirma que a DM é uma análise de grandes conjuntos de dados, com a finalidade de encontrar relacionamentos inesperados e resumir os dados, tornando-os úteis e compreensíveis. Tan, Steinebach e Kumar (2009) ainda descrevem que as técnicas de mineração de dados agem em grandes bancos de dados objetivando descobrir padrões úteis, não detectados em outros tipos de análises.

Castanheira (2008) afirma que após a escolha da técnica correta, para obtenção de resultados que possam ser analisados nas fases de interpretação e avaliação, o desenvolvimento do algoritmo deve ser adaptado ao problema proposto, a fim de ter a execução plena, com resultados eficazes. A seleção das técnicas de mineração de dados consiste em dois passos (DIAS, 2002):

- Traduzir o problema a ser resolvido em séries de tarefas de mineração de dados;

- Compreender a natureza dos dados disponíveis em termos de conteúdo, tipos de campos de dados e estrutura das relações entre os registros.

A mineração de dados tem função no processo KDD como uma das fases que possui a capacidade de realização de algumas tarefas (LAROSE, 2014), sendo as mais comuns, a classificação, a regressão, as regras de associação e o agrupamento. Enfatiza-se que nesta monografia, maior ênfase será direcionada à tarefa de agrupamento, técnica que será utilizada na mineração de dados do aplicativo Michelzinho pois essa seria a tarefa que poderia descobrir novos padrões de comportamento na realização de expressões faciais.

### 2.8.1 Classificação

Giacomelli (2020) afirma que a tarefa de classificação objetiva a identificação a qual classe um determinado registro pertence. Nesta tarefa, "o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de "aprender" como classificar um novo registro (aprendizado supervisionado)" (SILVA, 2014, p. 21).

Tan, Steinebach e Kumar (2009) afirmam que as técnicas de classificação são mais apropriadas para prever ou descrever conjuntos de dados com categorias nominais, os quais não possuem uma ordem definida, ou binárias, sendo menos efetiva para categorias ordinais, os quais possuem uma ordem definida. Segundo Silva (2018), a classificação pode ser:

- Descritiva: Pode servir como ferramenta explicativa para se distinguir entre os objetos e classes diferentes;
- Preditiva: prevê o rótulo de classe de registros não conhecidos.

### 2.8.2 Regressão

Segundo Larose (2005), a Regressão ou Estimação é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico, podendo assim ser estimado o valor de uma determinada variável analisando-se os valores das demais, assimilando a busca por uma função que organize os registros do banco de dados para valores reais. De acordo com Fayyad et al. (1996), o objetivo dessa tarefa é buscar uma função que faça mapeamento dos dados em variáveis de valor real.

### 2.8.3 Regras de Associação

A associação é usada para encontrar características (atributos) altamente associadas dentro dos dados, sendo que tais características podem ser representadas na forma de

regras de implicação (TAN et al., 2006). Conforme Silva (2018), as regras de associação são utilizadas para encontrar padrões que relatem qualidades profundamente associadas nos dados, geralmente exibindo-os na forma de subconjuntos de características ou regras de implicação. Esse tipo de tarefa tem como objetivo a extração de padrões relevantes de forma ágil, causado pelo tamanho exponencial de sua área de busca.

### 2.8.4 Agrupamento

Conforme Larose (2005), o agrupamento, do inglês "*clustering*", é uma coleção de registros similares entre si, porém diferente dos outros registros nos demais agrupamentos. Essa tarefa visa identificar e aproximar os registros similares, sem a pretensão de classificar, estimar ou predizer o valor de uma variável. Os métodos de agrupamento são denominados como aprendizado não supervisionado devido às informações do rótulo de classe não estarem presentes diferindo assim da "Classificação" que necessita de dados previamente categorizados (aprendizado supervisionado) (HAN; PEI e KAMBER, 2011)

Os agrupamentos podem ser classificados em várias categorias, sendo as principais os métodos hierárquicos, os métodos particionais e os métodos baseados em densidade (AMO; ROC, 2003). Cassiano (2014) divide os algoritmos de clusterização de uma forma generalizada, classificando-os em 10 tipos principais:

- Métodos Hierárquicos
- Métodos Particionais
- Métodos Baseados em Densidade
- Métodos Baseados em Grade
- Métodos Baseados em Modelos
- Métodos Baseados em Redes Neurais Artificiais
- Métodos Baseados em Lógica Fuzzy
- Métodos Baseados em Kernel
- Métodos Baseados em Grafos
- Métodos Baseados em Computação Evolucionária

Na investigação desse estudo são detalhados 3 (três) desses métodos, sendo o primeiro um Método Particional, por sua simplicidade e alto uso em diversas áreas, o segundo um Método Baseado em Densidade para a avaliação de diferentes padrões de agrupamento, e por fim, um Método Baseado em Redes Neurais Artificiais, que por ser pouco usado

em trabalhos científicos decidiu-se a investigação do mesmo, gerando a comparação dos resultados obtidos através desses diferentes métodos de mineração de dados.

#### 2.8.4.1 Métodos Particionais

Os métodos particionais procuram obter uma única partição dos dados de entrada em um número fixo de grupos, geralmente por meio da otimização de uma função objetivo (MACÁRIO FILHO, 2015, p.13). Celinski (1998) afirma que os métodos particionais são mais adequados a grandes conjuntos de dados, com maior utilização na segmentação de imagens, em razão da complexidade computacional envolvida.

Quanto aos tipos de agrupamentos dos métodos particionais, define-se que:

> Os métodos particionais podem ser divididos em agrupamento hard e fuzzy. No agrupamento hard, cada objeto é atribuído a apenas um grupo. No agrupamento fuzzy, a função objetivo fornece um aprimoramento conceitual, permitindo que um objeto seja associado a todos os grupos de acordo com um grau de pertinência. Isto pode ser útil quando existe uma sobreposição entre os grupos (MACÁRIO FILHO, 2015, p. 13).

Quanto à função, Silva (2018) afirma que os métodos particionais objetivam agrupar registros em k grupos, consistindo k na quantidade de grupos estabelecidos, sendo que o k (quantidades de grupos) é dado pelo usuário.

No trabalho intitulado de “Análise de Séries Temporais Usando Análise Espectral Singular (SSA) e Clusterização de suas Componentes Baseada em Densidade”, Cassiano (2014, p. 70) cita o "*modus operandi*"do desenvolvimento da mineração de dados utilizando um método particional:

> “Inicialmente, o algoritmo escolhe k objetos como sendo os centros dos k clusters. Os objetos são divididos entre os k clusters de acordo com a medida de similaridade adotada, de modo que cada objeto fique no cluster que forneça o menor valor de distância entre o objeto e o centro do mesmo. Então, o algoritmo utiliza uma estratégia iterativa de controle para determinar que objetos devem mudar de cluster, de forma que a função objetivo usada seja otimizada”.

Em continuidade, a autora expõe a segunda parte da operação:

> Após a divisão inicial, há duas possibilidades na escolha do “elemento” que vai representar o centro do cluster, e que será a referência para o cálculo da

> medida de similaridade. Ou utiliza-se a média dos objetos que pertencem ao cluster em questão, também chamada de centro de gravidade do cluster (esta é a abordagem conhecida como k-means); ou escolhe-se como representante o objeto que se encontra mais próximo ao centro de gravidade do cluster (abordagem é conhecida como k-medoids), sendo o elemento mais próximo ao centro chamado de medoid. (CASSIANO, 2014, p. 70-71).

Desse modo, o método é executado diversas vezes com valores diferentes de k, onde se seleciona resultados que possam expressar uma melhor interpretação dos grupos ou uma melhor representação gráfica (DONI, 2004).

Segundo Cassiano (2014) o K-Means é o mais popular e mais simples algoritmo particional além de ter sido um dos métodos de mineração utilizado em resposta ao questionamento comparativo deste trabalho. A descoberta desse algoritmo ocorreu de forma independente em diversos campos científicos, sendo primeiramente proposto por mais de 50 anos. Já Jain (2009) afirma que o K-Means, em razão de sua facilidade de implementação, simplicidade, eficiência e sucesso empírico, ainda é um dos algoritmos mais utilizados na mineração de dados, além de possuir, várias extensões desenvolvidas em várias formas.

De acordo com Gross (2014), o funcionamento inicia-se com a necessidade da definição de um número k de grupos que se pretende obter, partindo do pressuposto que cada grupo contém um centro, ou centróide, que tem seu cálculo baseado nas características dos dados de cada grupo. Ainda a cada novo dado inserido ao grupo, o centróide do mesmo é recalculado com base na média das características dos dados dos grupos. O algoritmo 1 adaptado de Tan et al., (2019) ilustra o pseudocódigo do Kmeans.

Conforme Tan, Steinbach e Kumar (2009) a estabilização dos grupos acontece até o momento em que não ocorra nenhuma mudança, após repetições do processo de atribuição de elementos a grupos e a atualização dos centróides. O autor Celinski, (1998, p. 16) afirma que uma vantagem do K-Means é a pouca dependência de limiares fornecidos pelo usuário, sendo necessários apenas o número de grupos, que é fixo, e o número de iterações e/ou outras variáveis que possam estabelecer a parada do algoritmo. Como desvantagem, Silva (2018) diz que o método é sensível a ruídos, "*outliers*"e não pode lidar com grupos de densidades diferentes, além de necessitar do fornecimento do número de grupos como parâmetro, e ainda trabalhar, somente com dados numéricos. Na figura 2, abaixo, é possível observar um conjunto de dados gerados aleatoriamente e agrupados pelo algoritmo com seus centróides.

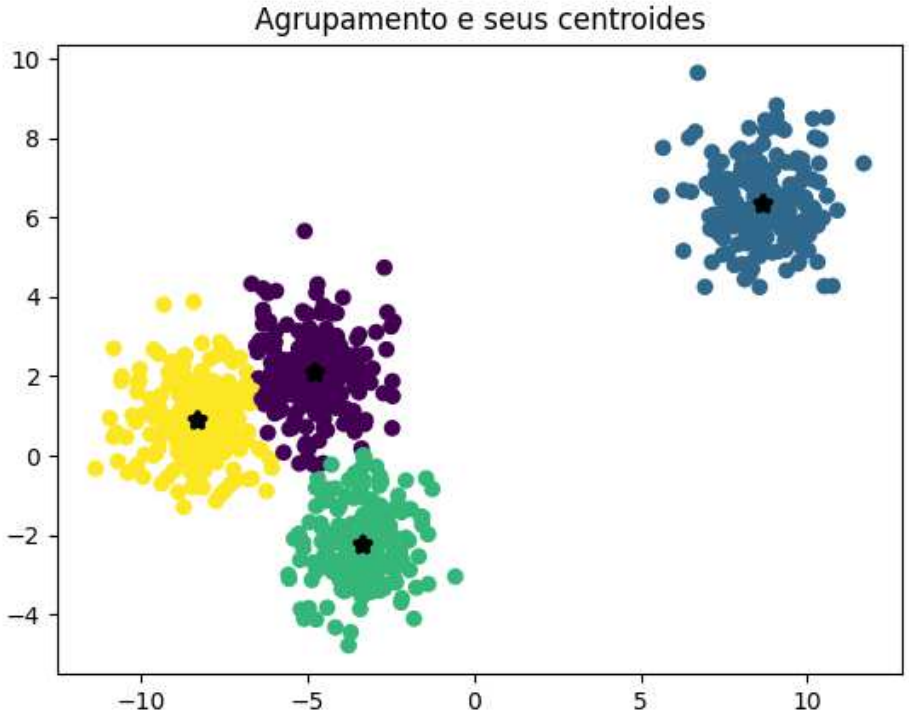



Figura 2 – Agrupamentos gerados via Kmeans e seus respectivos centroides. Fonte: Elaborado pelo autor.

**Algoritmo 1:** Kmeans

**Entrada**: o número K de grupos
**início**
  Inicialização;
  **repita**
    Forme K grupos designando cada ponto ao centroide mais próximo.;
    Recalcule o centroide de cada grupo;
  **até** *Centroides não mudarem;*
  ;
**fim**

#### 2.8.4.2 Métodos baseados em densidade

Os métodos baseados em densidade permitem descobrir grupos de formatos arbitrários. Estes métodos consideram grupos como sendo regiões densas de registros no espaço de dados que são separados por regiões de baixa densidade, que geralmente representam ruídos (HAN & KAMBER. 2006). Cassiano, Souza e Pessanha (2014) afirmam que nos métodos baseados em densidade, os clusters são definidos como regiões densas, separadas por regiões menos densas que representam ruídos.

De acordo com Cassiano, Souza e Pessanha (2014), os métodos baseados em densidade são capazes de identificar grupos de formato irregular ou arbitrário, diferentemente dos métodos particionais que definem apenas grupos de formato circular ou esférico, sendo inclusive, eficientes, para encontrar ruídos. Na Figura 3 é possível observar a diferença entre os métodos, sendo que métodos particionais conseguem fazer uma identificação eficiente de grupos apenas no conjunto 1 (grupos circulares ou esféricos), enquanto métodos baseados em densidade identificam grupos, com facilidade, em quaisquer dos 3 (três) conjuntos.

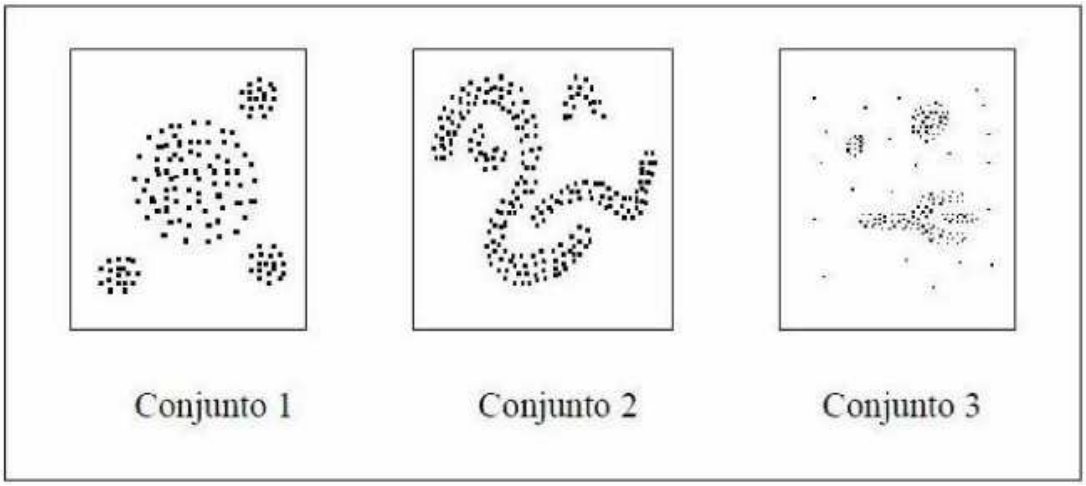



Figura 3 – Conjunto com grupos circulares (conjunto 1), não circulares (conjuntos 2 e 3) e com ruídos (conjunto 3).Fonte: Adaptado de Ester et al., (1996).

De acordo com Silva (2018) uma maior densidade específica de pontos (em quantidade considerável) dentro de cada grupo é o principal motivo pelo qual o reconhecimento é mais eficiente. Ester et al. (1996) afirmaram que a densidade dentro das áreas de ruído, é menor do que a densidade em qualquer um dos grupos. Segundo Cassiano (2014), isso possibilita que o cérebro humano reconheça os grupos e ruídos automaticamente quando utilizado o conceito de grupos formados por densidade. Desse modo, neste trabalho, será investigado o algoritmo DBSCAN, detentor de uma estrutura de método baseado em densidade.

Segundo Mendes (2017, p. 36), o "DBSCAN é um algoritmo de agrupamento baseado em densidade simples e eficaz que ilustra uma quantidade de conceitos que são importantes para qualquer abordagem de agrupamento baseado em densidade". O nome DBSCAN é a abreviação do termo *'Density Based Spatial Clustering of Application with Noise'* (Clusterização Espacial Baseada em Densidade de Aplicações com Ruído) (CASIANO, 2014).

Ester et. al (1996) ainda define que o DBSCAN é um algoritmo não paramétrico, com o objetivo de descobrir diferentes formas de agrupamento (formato arbitrário e de diferentes tamanhos), e ainda, indicar a presença de ruídos nas bases de dados, detec-

tando clusters 'naturais' e arranjos nos dados, sem informação preliminar. No algoritmo 2 adaptado de Tan et al. (2019) é possível observar o pseudocódigo do DBSCAN.

**Algoritmo 2:** DBSCAN

**Entrada**: eps, min_pts
**início**
Inicialização;
Nomeie todos os pontos como core, border ou noise;
Elimine os pontos noise;
Trace uma borda entre cada ponto core a uma distância Eps;
Transforme cada grupo de pontos core conectados em um cluster separado;
Atribua cada ponto a um de seus clusters associados;
**fim**

De acordo com Silva (2018), esse algoritmo utliza o conceito de densidade baseada em centro, sendo a densidade de um ponto no conjunto de dados a quantidade de pontos dentro de um raio de vizinhança, incluindo o próprio ponto. Na Figura 4 detalhes dessa estratégia são apresentados.

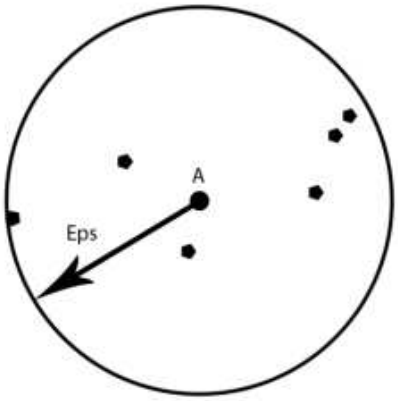
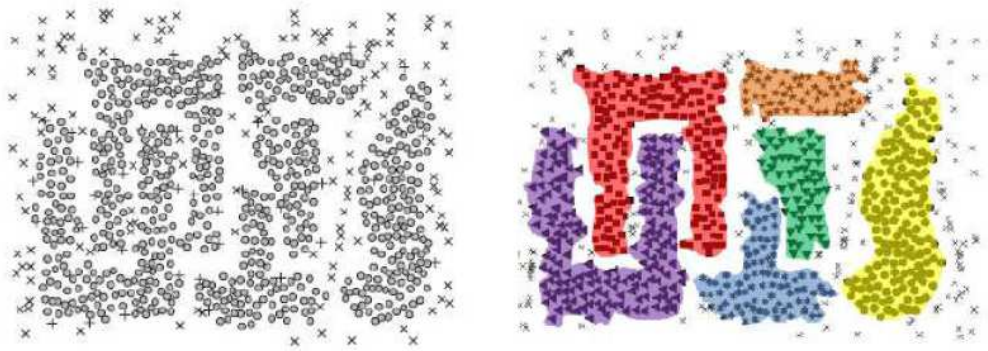



Figura 4 – Densidade baseada em centro.Fonte: Adaptado de Mendes (2017).

Semann (2013) afirma que o DBSCAN contém dois parâmetros de entrada, o raio (de vizinhança) denominado Eps e a quantidade mínima de pontos em um determinado raio descrito por MinPts. Na conceituação realizada por Mendes (2017), o 'Raio de vizinhança' determina o raio de vizinhança (Eps) para cada ponto da base de dados, sendo que após estabelecido, esse algoritmo verifica a quantidade de pontos contidos no raio (Eps) para cada ponto na base de dados, e se essa quantidade exceder certo número, um cluster é formado. O segundo parâmetro, o número mínimo de pontos (MinPts), de acordo com Mendes (2017, p. 36), "especifica o número mínimo de pontos, no dado raio (Eps),

que um ponto precisa possuir para ser considerado um ponto central e consequentemente, de acordo com as definições de cluster baseado em densidade, inicia a formação de um cluster".

Na Figura 5(a), Silva (2018) cita que "o fato de densidade baseada no centro, possibilita classificar um ponto como estando: no interior de uma região densa (ponto de centro), no limite de uma região densa (ponto de limite) ou em uma região ocupada esporadicamente (ponto de ruídos)". (SILVA, 2018, p. 32).



(a) (x) - Ponto de Ruídos (+) - Ponto de Limite (o) - Ponto de centro

(b) Grupos encontrados por DBSCAN

Figura 5 – Agrupamento DBSCAN para 3.000 pontos bidimensionais. Fonte: Adaptado de Silva (2018).

Na Figura 5(b), após a execução do DBSCAN, é apresentado o resultado, sendo possível a observação dos pontos de centro e de limite formando grupos enquanto os pontos de ruídos se apresentam afastados. Assim, de acordo com Tan, Steinbach, Kumar (2009), a principal vantagem do DBSCAN é que o algoritmo pode encontrar muitos grupos que o K-Means não poderia encontrar. No entanto, os autores afirmam que a principal desvantagem do algoritmo é trabalhar com grupos de densidades muito variadas, além de ser custoso calcular os pontos vizinhos quando requerido cálculo de proximidade entre pares.

#### 2.8.4.3 Métodos baseados em rede neurais artificiais

"Os métodos baseados em Redes Neurais Artificiais (RNAs) consistem em uma técnica que constrói um modelo matemático, de um sistema neural biológico simplificado, com capacidade de aprendizado, generalização, associação e abstração" (CASTANHEIRA, 2008, p.24). A autora afirma ainda que, como no cérebro humano, as redes neurais apresentam uma estrutura altamente paralela, composta por processadores simples (neurônios artificiais) conectados entre si.

Com sua origem na psicologia e na neurobiologia, Camilo e Silva (2009) afirmam que as RNAs consistem em simular o comportamento dos neurônios apropriados para

tarefas de percepção como o reconhecimento, classificação e autoassociação de padrões. Ainda de acordo com os autores, uma rede neural pode ser vista como um conjunto de unidades de entrada e saída conectadas por camadas intermediárias e cada ligação possui um peso associado. Os autores ainda afirmam que para conseguir uma correta classificação do registro, há o ajustamento destes pesos, no decorrer do processo de aprendizado.

De acordo com Batista (2003), quando a base de dados for constituída de dados qualitativos, o desmembramento destes dados em n atributos binários, para n valores diferentes na base qualitativa, criando um nó para cada atributo, é sugestivo, em se tratando da aplicabilidade de redes neurais. Haykin (2001) afirma que a habilidade para aprender a partir do ambiente na qual estão inseridas (ambiente de aprendizado), melhorando o desempenho da aprendizagem, é uma importante propriedade das redes neurais.

A composição da RNA é formada por várias unidades de processamento, que geralmente são conectadas por canais de comunicação (neurônios artificiais) que estão associados a determinados pesos. Os pesos são um modelo para simular os dendritos, responsáveis pelas sinapses no cérebro humano. Para um melhor entendimento, apresenta-se na Figura 6 abaixo a estrutura do neurônio biológico (sinapse) (TAFNER, 1998):

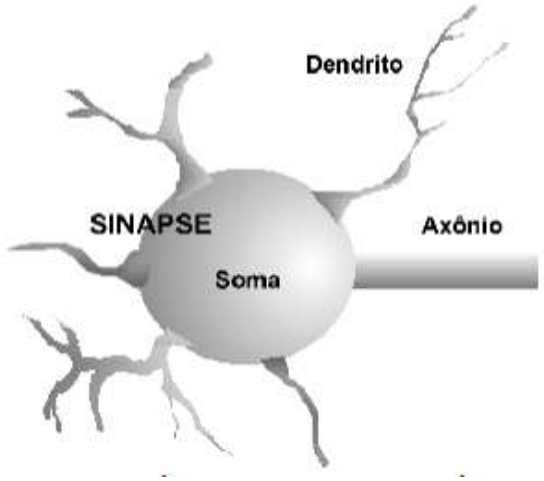



Figura 6 – Estrutura do neurônio biológico (sinapse).Fonte: Adaptado de Tafner (1998).

Na Figura 7 é apresentado a estrutura do neurônio artificial que é uma estrutura lógico-matemática que procura simular o neurônio biológico, sendo que os dendritos são substituídos pelas entradas, enquanto as ligações com o corpo celular são realizadas pelos pesos (sinapses), que captam os estímulos nas entradas para processamento através da função soma, substituindo o limiar de disparo pela função transferência.

Tafner (1998) afirma que, durante os estímulos, ao alterar os valores dos pesos representativos, isso pode influenciar o resultado do sinal de saída. As entradas, simulando uma área de captação de estímulos, podem se conectar a muitos neurônios, tendo como resultado uma série de resultados, sendo cada saída representada por um neurônio.

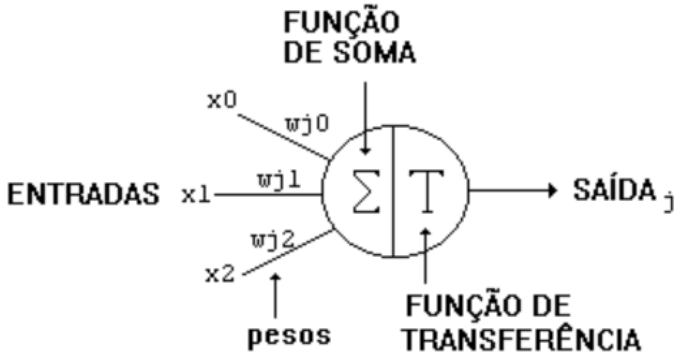



Figura 7 – Estrutura do neurônio artificial (sinapse). Fonte: Adaptado de Tafner, (1998).

Na Figura 8 é apresentado um exemplo de RNA de 2 camadas com 4 entradas e 2 saídas. Nessa Figura é possível observar que as variantes da RNA são muitas, que através de combinações, pode-se mudar a arquitetura da aplicação conforme a demanda do projetista.

Os itens de acordo com Tafner (1998), que compõem uma RNA, sujeito às modificações, basicamente são:

- Conexões entre camadas;
- Camadas intermediárias;
- Quantidade de neurónios;
- Função de transferência;
- Algoritmo de aprendizado.

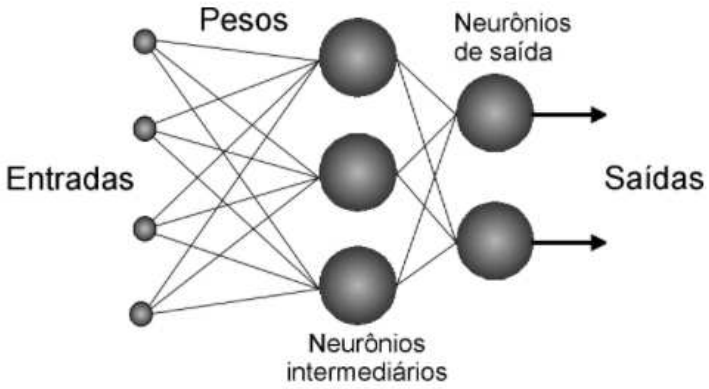



Figura 8 – Rede Neural Artificial de 2 camadas com 4 entradas e 2 saídas. Fonte: Adaptado de Tafner (1998).

"Assim as diferentes possibilidades de conexões entre as camadas de neurônios podem ter, em geral, n números de estruturas diferentes. A função da conexão em si é tornar o sinal de saída de um neurônio em um sinal de entrada de outro, orientando o sinal de saída para o mundo externo (mundo real)" (CASTANHEIRA, 2008, p.29). Com foco no estudo utilizando as RNAs, nesta pesquisa, dar-se-á destaque aos Mapas de Kohonen, que é um dos algoritmos que possibilitou o resultado comparativo no estudo.

Kohonen (2012) afirma que a metodologia de mapa de auto-organização (do inglês, *"Self-Organizing Map"* – SOM) aproxima a função de densidade de probabilidade dos dados de entrada, podendo ser utilizada em agrupamentos assim como para uma visualização eficiente de dados de alta dimensão em uma rede bidimensional. Ainda, o autor em DE PINHO (2008, p. 42) define os mapas auto-organizáveis como sendo: "uma classe especial de redes neurais baseadas na aprendizagem competitiva: neurônios na camada de saída competem entre si para ativação ou disparo, e somente um neurônio é ativado a cada apresentação de padrão".

"Em que o neurônio ativado é chamado de neurônio vencedor, e somente os pesos associados a este neurônio serão atualizados", ou seja, "o vencedor ganha tudo" (DE PINHO, 2008, p.42). É possível observar o pseudocódigo do mapa de Kohonen no algoritmo 3 adaptado de Tan et al. (2019).

**Algoritmo 3:** SOM

```
Entrada: sigma, learning_rate, n_linhas, n_colunas
início
    repita
        Selecione o próximo registro;
        Determine o neurônio vencedor como o mais próximo ao registro;
        Atualize o neurônio vencedor e os vizinhos que estejam a uma
        distância sigma do neurônio vencedor;
    até Centroides não mudarem;
fim
```

O algoritmo de mapa auto-organizável de Kohonen, é uma técnica que inicialmente foi estabelecida por Teuvo Kohonen, em 1981, e consiste em uma rede neural artificial interconectada e não supervisionada que permite um mapeamento auto-ajustável do espaço de estados multidimensionais estudado (AFFONSO, 2011, p. 2).

Desse modo, o autor Doni (2004, p. 65) afirma que a topologia da rede de Kohonen possui duas camadas, na qual todas as unidades de entrada encontram-se conectadas às unidades de saída através de conexões sinápticas. Na Figura 9 é apresentado o modelo de mapa auto-organizável proposto por Kohonen:

Para a formação do mapa de Kohonen, existem três processos envolvidos, que são

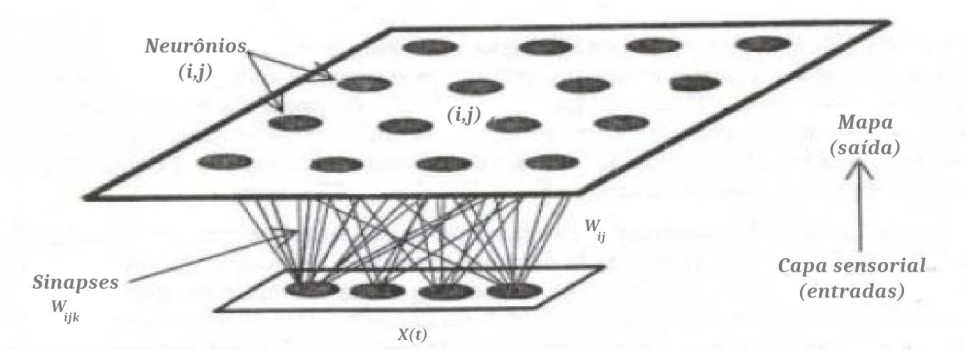



Figura 9 – O Modelo de Kohonen. Fonte: Adaptado de Haykin, (2001).

citados abaixo e detalhados a seguir (HAYKIN, 2001):

- Competição: para cada padrão de entrada, os neurônios da rede computam os seus respectivos valores de uma função discriminante. Esta função provê as bases para a competição entre os neurônios. O neurônio com maior valor da função discriminante é declarado vencedor da competição.
- Cooperação: o neurônio vencedor determina a localização espacial de uma vizinhança topológica de neurônios excitados, provendo, desta forma, as bases para a cooperação entre tais neurônios vizinhos.
- Adaptação Sináptica: este último mecanismo permite aos neurônios excitados aumentar seus valores individuais da função discriminante em relação ao padrão de entrada, através de ajustes adequados aplicados a seus pesos sinápticos. Os ajustes feitos são tais que, a resposta do neurônio vencedor à subsequente aplicação de um padrão similar de entrada é realçada.

## 2.9 Pós-processamento

A finalidade dos algoritmos utilizados na mineração de dados é encontrar conhecimentos, porém os mesmos não possuem preocupação na exibição dos resultados segundo Dias et al. (2012). Silberschatz e Tuzhilin (1995) afirmam ser vital o desenvolvimento de técnicas de apoio no fornecimento, aos usuários, somente de padrões mais interessantes na mineração de dados. No pós-processamento há a realização de avaliação e interpretação dos padrões auferidos, a fim de transmitir ao usuário confiança e compreensão (TEIXEIRA; SILVA, 2020). Ainda, Ramos e Lobo (2003) afirmam ser a mineração de dados, uma busca por padrões aplicando o uso de algoritmos, enquanto o KDD engloba a interpretação e avaliação dos resultados encontrados, atribuindo pesos ao conhecimento útil existente nos mesmos.

### 2.9.1 Interpretação

Pugliesi (2004) afirma que várias pesquisas têm sido realizadas com a finalidade de auxiliar o usuário no entendimento e utilização do conhecimento adquirido, dividindo-as em medidas de desempenho e medidas de qualidade. A autora afirma ainda que medidas de desempenho tratam-se da composição dos dados (características) e de onde podem ser utilizados (campo de atuação), referindo-se à interpretação (compreensão dos dados), enquanto as medidas de qualidade permitem a avaliação na totalidade e de como podem ser utilizados, obtendo como consequência, a aprovação, aceitação ou recusa do conhecimento extraído, referindo-se ao grau de interesse dos dados pelos usuários (confiabilidade).

Silberschatz e Tuzhilin (1995) afirmam que as medidas de grau de interesse podem ser divididas em objetivas e subjetivas. As medidas objetivas, segundo Pugliesi (2004, p. 19), "são aquelas que estão relacionadas somente com a estrutura dos padrões e do conjunto de dados" sendo que "elas não consideram fatores específicos do usuário nem do conhecimento do domínio para avaliar um padrão". Porém, de acordo com Giacomelli, (2020, p. 26), todos os envolvidos no projeto devem participar da etapa de interpretação, tanto os especialistas em KDD, quanto os especialistas de domínio da aplicação, pois deverá ser realizada uma avaliação dos resultados obtidos, podendo ter diferentes graus de interesses para um determinado padrão, sendo necessárias medidas subjetivas. Assim, Pugliesi (2004, p.19) afirma que "as medidas subjetivas consideram que fatores específicos do conhecimento do domínio e de interesse do usuário devem ser tratados ao selecionar um conjunto de regras interessantes ao usuário".

### 2.9.2 Validação de Dados

O autor Eing (2019, p. 30) afirma que a validação consiste na avaliação dos clusters, atestando a qualidade dos mesmos, tendo como objetivo fornecer para o usuário um resultado de confiança. Os algoritmos de agrupamento produzem padrões, sem considerar a existência de dados no cluster, sendo que a utilização de um mesmo algoritmo pode levar a seleção de um parâmetro a um resultado diferente.

Segundo Silva (2018, p. 34), "a validação de agrupamento aborda os processos formais que avaliam, de forma objetiva e quantitativa, os resultados da análise do agrupamento". Os índices de validação comparam o resultado de diferentes agrupamentos, assim como os resultados do mesmo algoritmo com diversos valores de *cluster* (CEBECI; KAVLAK; YILDIZ, 2017).

De acordo com Zaki e Meira Júnior (2014) há várias medidas de validade que realizam a validação de dados, nas quais podem ser divididas em três tipos principais: critério interno, externo e relativo.

Os três tipos de critérios de validação são assim definidos por Gross, (2004, p.19):

> "A validação externa analisa os agrupamentos com base em uma informação de estrutura fornecida à aplicação, logo requer um conhecimento prévio das classes do conjunto de dados ou uma expectativa de resultados. A validação interna determina se os agrupamentos são intrinsecamente apropriados aos dados; não há necessidade de conhecimento prévio, a análise ocorre considerando os próprios objetos em cada grupo e suas características. E a validação relativa comparar duas estruturas de agrupamento de dados, fornecidas pelo mesmo algoritmo de agrupamento, mas obtidas em execuções com parâmetros distintos". (GROSS, 2004, p.19)

Segundo Rousseeuw (1987) o critério de validação relativo da silhueta simplificada e sua equação está apresentada nas Equações 2.1 e 2.2 (Silva e Abreu, 2011). O termo $a_{p,i}$ descreve a dissimilaridade média do i-ésimo registro ao seu *cluster* e $b_{p,i}$ é a dissimilaridade média do i-ésimo registro ao *cluster* mais próximo. Quanto menor for o valor de ap,i e maior for o valor de $b_{p,i}$, mais próximo de 1 é Sil, implicando que o agrupamento dos dados determinado pelo algoritmo é o mais apropriado.

A silhueta original e a versão simplificada são representadas pelas mesmas equações 2.1 e 2.2. No entanto, na silhueta original o cálculo de e é oriundo da média de todas as dissimilaridades do *cluster* em questão, enquanto que na adaptação simplifica esses termos são obtidos a partir da distância dos centroides dos *clusters*. O presente trabalho utiliza primariamente o critério da silhueta simplificada para validação.

$$Sx_i = \frac{b_{p,i} - a_{p,i}}{max(a_{p,i}, b_{p,i})} \tag{2.1}$$

$$Sil = \frac{1}{N} \sum_{i=1}^{N} Sx_i \tag{2.2}$$

## 2.10 Considerações finais

Neste capítulo foi apresentada uma visão geral sobre as emoções, os jogos interativos na aprendizagem, o TEA e o aplicativo MICHELZINHO. Há também uma apresentação de todo o processo de KDD, em especial as etapas de pré-processamento, mineração de dados e pós-processamento, com destaque para a teoria dos algoritmos utilizados neste trabalho. Esses tópicos foram fundamentais para entendimento do problema e assim como definição dos algoritmos a serem investigados sobre os experimentos propostos.

# 3 Desenvolvimento

## 3.1 Introdução

Neste capítulo são tratadas as etapas do desenvolvimento dos algoritmos para realização dos experimentos. Na seção 3.2 é descrito a coleta e o tratamento dos dados, a base e os 24 atributos utilizados. Na subseção 3.2.2 é apresentado todo o pré-processamento realizado a fim de refinar a base de dados. Já a subseção 3.2.3 mostra a base de dados gerada, a qual é utilizada na mineração em si. Em prosseguimento na seção 3.3 são detalhados os experimentos de mineração realizados na base de dados do aplicativo MICHELZINHO. Por fim a seção 3.4 traz as considerações finais. A figura 10 mostra de forma sucinta as etapas desenvolvidas que serão detalhadas nas seções posteriores

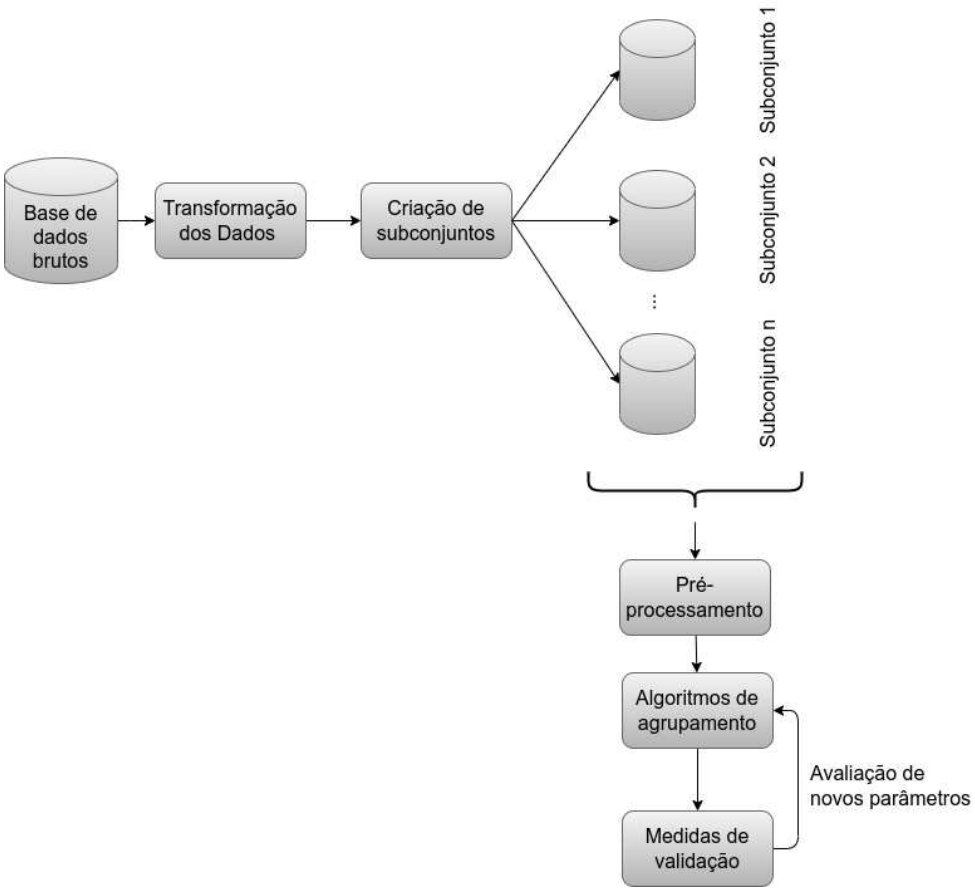



Figura 10 – Etapas do método proposto. Elaborado pelo Autor, (2023).

## 3.2 Coleta e Tratamento de Dados

### 3.2.1 Apresentação da base de dados do MICHELZINHO

A base de dados utilizada foi disponibilizada por Dantas (2022), em que é composta por dados armazenados entre o período de Abril de 2018 a Junho de 2022. Nessa base constam informações referentes aos agrupamentos dos músculos ativados pelo usuário para representar as emoções e as notas de cada emoção mapeada pelo aplicativo. Destaca-se que nessa base não há nenhuma informação armazenada das pessoas.

Assim, o desenvolvimento deste trabalho embasa-se nessa base de dados constituída de 545.044 registros, contendo individualmente 24 (vinte e quatro) atributos distintos, com variações desde códigos, diferentes expressões (alegria, raiva, medo, surpresa etc.), até os referidos músculos utilizados na representação da emoção. Dos 24 atributos mencionados, 2 (dois) são categóricos (indivíduo – ID e emoção) e os outros 22 (vinte e dois) são numéricos. Há ainda o destaque de que a base não conta com conhecimento prévio sobre a organização dos dados ou rótulos de classes (*ground truth*).

A Tabela 1 traz o detalhamento de cada atributo apresentando o nome, a descrição e o tipo, bem como a quantidade de valores ausentes e não aplicáveis (NA) existentes no conjunto de dados.

Tabela 1 – Base de dados do aplicativo MICHELZINHO®

| **ATRIBUTO** | **DESCRIÇÃO** | **TIPO** | **Nº AUSENTES** |
|---|---|---|---|
| cod | Código incremental identificador do registro | INTEGER | 0 |
| data | Data em que o registro foi gerado (Quando o usuário jogou) | DATETIME | 0 |
| alegria | Probabilidadde da emoção desenvolvida ser alegria | FLOAT | 15645 |
| tristeza | Probabilidadde da emoção desenvolvida ser tristeza | FLOAT | 15648 |
| raiva | Probabilidadde da emoção desenvolvida ser raiva | FLOAT | 15648 |
| medo | Probabilidadde da emoção desenvolvida ser medo | FLOAT | 15648 |
| desgosto | Probabilidadde da emoção desenvolvida ser desgosto | FLOAT | 15648 |
| surpresa | Probabilidadde da emoção desenvolvida ser surpresa | FLOAT | 15648 |
| emocao | Qual era a emoção objetivo daquele jogo | FLOAT | 15648 |
| paciente | Nome do usuário | STRING | 410143 |
| smile | Grau de ativação dos músculos correspondentes ao sorriso | FLOAT | 229326 |
| inerbrow | Grau de ativação dos músculos correspondentes ao movimento de cerrar a sobrancelha | FLOAT | 229326 |
| browrise | Grau de ativação dos músculos correspondentes a levantar a sobrancelha | FLOAT | 229326 |
| nose | Grau de ativação dos músculos correspondentes ao nariz | FLOAT | 229326 |
| uperlip | Grau de ativação dos músculos correspondentes ao lábio superior | FLOAT | 229326 |
| lipcorner | Grau de ativação dos músculos correspondentes aos cantos do lábio | FLOAT | 229326 |
| chinrise | Grau de ativação dos músculos correspondentes a levantar o queixo | FLOAT | 229326 |
| lippuker | Grau de ativação dos músculos correspondentes ao movimento de “biquinho” | FLOAT | 229327 |
| lippress | Grau de ativação dos músculos correspondentes a pressionar os lábios | FLOAT | 229327 |
| lipsuck | Grau de ativação dos músculos correspondentes ao movimento de sugar com a boca | FLOAT | 229327 |
| mouthopen | Grau de ativação dos músculos correspondentes a abrir a boca | FLOAT | 229327 |
| smike | Grau de ativação dos músculos correspondentes a pressionar os lábios | FLOAT | 229327 |
| eyeclose | Grau de ativação dos músculos correspondentes a fechar os olhos | FLOAT | 229327 |
| atention | Grau de atenção do usuário | FLOAT | 229327 |

Fonte: Elaborado pelo Autor (2023).

### 3.2.2 Pré-processamentos Realizados

Foi realizada uma consulta ao autor do trabalho Dantas (2022), criador do aplicativo MICHELZINHO, para facilitar a tarefa de pré-processsamento indicando caminhos e complementando informações referentes à base de dados. Então, ocorreu a remoção dos conjuntos de dados dos atributos desnecessários ao processo de mineração, ratificando a base de dados em apenas 21 atributos, após o descarte das colunas relacionadas aos atributos "cod" (por ser o ID do registro da partida do jogo) e "data" e "paciente" (no caso de campo não era preenchido e foi considerado irrelevante).

Por fim o atributo emoção foi transformado em atributos binarizados, justificado na afirmativa de Tan et. al. (2019 , p. 62), de que "se os dados forem processados para fornecer recursos de nível superior" consequentemente "um conjunto muito mais amplo de técnicas de classificação pode ser aplicado para este problema", através da "criação de um novo conjunto de recursos a partir dos dados originais, conhecido como extração de dados". Desse modo foi criada uma nova coluna de acertos, pois foi considerado que essa coluna agregaria mais informação a respeito da rodada do jogo com a seguinte regra:

- Se o valor do atributo "emoção" for igual ao item de maior valor nas emoções, então define o acerto;
- Senão, configura o contrário, ou seja, o erro.

No pré-processamento dos registros foram descartados os constantes de "valores ausentes" assim como as emoções "jogo7-1" e "adivinhar" já que esses valores não representavam uma emoção-alvo. De acordo com Tan et al. (2006, p. 133) "a análise de associação requer atributos binários assimétricos onde apenas a presença do atributo é importante", portanto é "necessário introduzir um atributo binário assimétrico para cada valor categórico"

Como definido por Silva (2018, p. 40), na codificação 1-de-n é criado um atributo para cada valor categórico e o atributo que corresponde ao valor deve ser preenchido com 1 e os demais com 0. Assim, ao pré-processar o atributo "emoção", foi utilizada a técnica de "Codificação 1-de-n". Com o atributo "emoção" binarizado, visualiza-se os demais atributos em uma escala diferente, fato que poderia afetar o cálculo das medidas de distância, que no geral, são utilizadas nos algoritmos de agrupamento. Desse modo, com o intuito de evitar maiores impactos nos cálculos, todos eles foram escalados para o intervalo (0, 1).

A equação utilizada para calcular as novas escalas de distância desses atributos foi dada pela Equação 3.1:

$$d' = \frac{(d - min_d)}{(max_d - min_d)} \tag{3.1}$$

Na qual d indica o atributo a ser reescalado, $\min_d$ o menor valor da coluna do atributo e $\max_d$ o maior valor. Finalizando a etapa de pré-processamento, houve a conversão da base, substituindo o formato do arquivo tipo ".sql" pelo formato ".csv", objetivando a facilidade de leitura pelos algoritmos utilizados na pesquisa.

### 3.2.3 Base de dados gerada

Após o pré-processamento, a base de dados gerada manteve 314.636 registros, com 47 atributos, dos quais 26 foram gerados a partir da coluna "emoção" usando a binarização 1-de-n. Cita-se ainda o atributo "acertou" que foi gerado a partir do conceito de extração de dados (feature extration).

A fim de explorar as técnicas de agrupamento em diferentes níveis de visualização, efetuou-se uma subdivisão da base de dados gerada por 6 (seis) principais atributos, gerando assim as seguintes bases de dados:

- Alegria: 29 atributos dos quais 9 foram criados (acertou, is_alegria, is_alegria3, is_alegria_bolinha, is_alegria_deserto, is_alegria_pescaria, is_jogo2-alegria, is_jogo7_alegria e is_mapa-alegria), sendo 129528 registros;
- Desgosto: 25 atributos dos quais 5 foram criados (acertou, is_desgosto, is_desgosto3, is_desgosto_ratos e is_mapa-desgosto), sendo 9512 registros;
- Medo: 24 atributos dos quais 4 foram criados (acertou, is_mapa-medo, is_medo e is_medo3), sendo 22868 registros;
- Raiva: 24 atributos dos quais 4 foram criados (acertou, is_mapa-raiva, is_raiva e is_raiva3), sendo 74472 registros;
- Surpresa: 26 atributos dos quais 6 foram criados (acertou, is_mapa-surpresa, is_surpresa, is_surpresa3, is_surpresa_lake e is_surpresa_vagalumes), sendo 44828 registros; e
- Tristeza: 24 atributos dos quais 4 foram criados (acertou, is_mapa-tristeza, is_tristeza e is_tristeza3), sendo 33428 registros.

## 3.3 Métodos de Agrupamento e Medidas de Validação utilizados

Os algoritmos de agrupamento utilizados neste trabalho foram o K-Means, o DBSCAN e o Mapas de Kohonen, sendo a escolha justificada na simplicidade e eficiência além dos diferentes níveis de utilização entre eles, sendo o Mapas de Kohonen de baixo uso. Na

execução dos métodos de agrupamento, foi utilizada a biblioteca Scikit-learn, que é um módulo Python que integra uma ampla gama de algoritmos de aprendizado de máquina de última geração para problemas supervisionados e não supervisionados (PEDREGOSA et. al., 2011). De acordo com os autores, esse pacote tem como objetivo levar o aprendizado de máquina para não especialistas utilizando uma linguagem de alto nível, enfatizando que o mesmo possui uma facilidade de uso além de desempenho, documentação e consistência da API. O software possui dependências mínimas e é distribuído sob a licença simplificada, incentivando seu uso em ambientes acadêmicos e comerciais.

O método DBSCAN foi executado com alterações do raio de vizinhança (Eps) de 0,2 a 1,0, com intervalos de 0,2 de modo que todas as possibilidades de distância fossem avaliadas com um certo intevalo entre elas. O MinPts varia de 10 a 100 com grupos de 10 elementos pois empiricamente considera-se que esse é o intervalo em que há uma quantidade mínima razoável de elementos em um grupo. Finalmente, com a técnica mapas de Kohonen foi executado alterando de 2 a 36 grupos, ou seja, 6 colunas por 6 linhas de neurônios de forma a obter uma malha quadrada, ou seja, que tivesse o maior número de vizinhos possível e de modo que varresse toda a quantidade de grupos que pudessem gerar informação ao especialista do domínio. Já na investigação com o algoritmo K-Means ocorreu uma análise do número de grupos de 2 a 36. Esse valor do número de grupos ocorreu a fim de que pudesse haver uma comparação entre os grupos gerados pelos algoritmos K-Means e Kohonen.

A medida de validação utilizada neste trabalho foi a Silhueta Simplificada, que teve seu detalhamento na subseção 2.9.2. A justificativa de escolha possui seu embasamento na afirmação de Vendramin et. al. (2010), em que a silhueta se baseia, em sua versão simplificada, no cálculo das distâncias médias do registro $x_i$ ao seu grupo, $a_i$, e ao grupo vizinho mais próximo, $b_i$, podendo ser redefinidas como distâncias ao centróide do seu grupo e ao centróide do grupo vizinho mais próximo, respectivamente. Desse modo, a complexidade é linear em relação ao número de registros, sendo a eficácia da silhueta simplificada competitiva em relação à eficácia da silhueta tradicional.

## 3.4 Considerações Finais

Este capítulo descreveu em detalhes a base de dados do MICHELZINHO, que foi estudada, e todos os pré-processamentos necessários para realizar a limpeza dos dados da base. Também foram informados quais os agrupamentos utilizados e suas configurações, além do método de validação do agrupamento. No capítulo 4 estão elencados os resultados deste trabalho utilizando as propostas descritas neste capítulo, assim como a discussão dos resultados obtidos.

# 4 Resultados

## 4.1 Introdução

Neste capítulo, o objetivo é analisar os resultados obtidos na utilização dos métodos de agrupamento e suas validações conforme descrito no capítulo 3. Os resultados dos agrupamentos são avaliados utilizando o índice de Silhueta Simplificada, de modo que quanto mais próximo o valor estiver de 1 melhor é o resultado. A seção 4.2. apresenta os resultados do algoritmo K-Means, a seção 4.3. do algoritmo DBSCAN e a seção 4.4. do algoritmo Mapas de Kohonen. A seção 4.5. traz as considerações finais. Estes resultados serão apresentados em tabelas objetivando facilitar a interpretação através de gráficos associados.

## 4.2 k-means

A Tabela 2 apresenta os valores da silhueta simplificada das bases das emoções agrupadas pelo algoritmo K-Means. As linhas da Tabela indicam os diferentes valores de k utilizados e as colunas indicam os subconjuntos de dados extraídos de acordo com cada tipo de emoção. Enquanto que os valores da silhueta são o resultado da melhor avaliação dessa medida em 10 execuções.

As bases da Tabela 2 foram pré-processadas usando a codificação 1-de-n. É perceptível nas bases de cada tipo de emoção que a medida em que vai aumentando o número de grupos há pouca modificação na qualidade do agrupamento, permitindo afirmar que o número de grupos pouco influencia na qualidade do resultado do agrupamento.

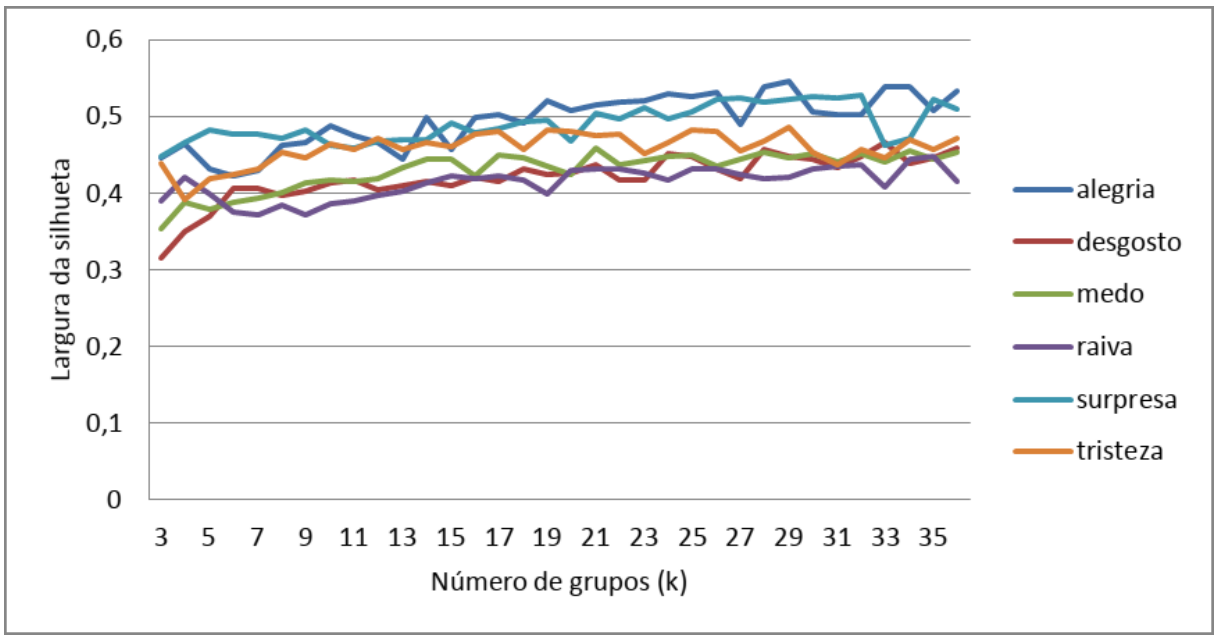
Tabela 2 – Agrupamento K-Means

| **Número de grupos (k)** | **Alegria** | **Desgosto** | **Medo** | **Raiva** | **Surpresa** | **Tristeza** |
|---|---|---|---|---|---|---|
| 3 | 0,44513271 | 0,314828465 | 0,353290605 | 0,389957622 | 0,448062771 | 0,438241053 |
| 4 | 0,464696273 | 0,349745049 | 0,387872659 | 0,421157792 | 0,465641522 | 0,391031992 |
| 5 | 0,430679051 | 0,36936603 | 0,37934963 | 0,399560712 | 0,482204768 | 0,418303352 |
| 6 | 0,422049647 | 0,406385037 | 0,388068537 | 0,374747365 | 0,477485573 | 0,423991722 |
| 7 | 0,430050885 | 0,405770483 | 0,392815995 | 0,370675937 | 0,477298308 | 0,431766545 |
| 8 | 0,461824615 | 0,396183183 | 0,400749033 | 0,384943383 | 0,47190597 | 0,453801912 |
| 9 | 0,466020037 | 0,402010055 | 0,41266592 | 0,372131293 | 0,482261434 | 0,446365482 |
| 10 | 0,487839337 | 0,413695197 | 0,416794867 | 0,386053015 | 0,463126423 | 0,463343753 |
| 11 | 0,475121565 | 0,417699461 | 0,415613154 | 0,390524266 | 0,459477852 | 0,457272961 |
| 12 | 0,466657319 | 0,403814708 | 0,419041103 | 0,397514999 | 0,46766295 | 0,47052263 |
| 13 | 0,444245461 | 0,410172747 | 0,433083535 | 0,402421615 | 0,470186004 | 0,457470751 |
| 14 | 0,499305958 | 0,414447165 | 0,44417596 | 0,413780998 | 0,469646395 | 0,466523515 |
| 15 | 0,457214214 | 0,410074051 | 0,443376498 | 0,421916798 | 0,491880578 | 0,460282446 |
| 16 | 0,499027368 | 0,420632144 | 0,421996048 | 0,41959987 | 0,477922425 | 0,477390636 |
| 17 | 0,502734451 | 0,414970203 | 0,450317583 | 0,422612258 | 0,484376934 | 0,48047015 |
| 18 | 0,491135289 | 0,430988837 | 0,446186858 | 0,416580825 | 0,492827689 | 0,456272219 |
| 19 | 0,519664827 | 0,423732642 | 0,435863691 | 0,398933915 | 0,494046013 | 0,483012201 |
| 20 | 0,507279357 | 0,425876633 | 0,423884323 | 0,430089675 | 0,468049982 | 0,480554901 |
| 21 | 0,5151622 | 0,436367205 | 0,459357878 | 0,431584688 | 0,503791007 | 0,475058895 |
| 22 | 0,519146824 | 0,416740404 | 0,436994592 | 0,431366036 | 0,495998125 | 0,476143226 |
| 23 | 0,51979279 | 0,417619614 | 0,441865026 | 0,425194426 | 0,51051873 | 0,451228305 |
| 24 | 0,530121656 | 0,450484192 | 0,448269266 | 0,416598866 | 0,496258427 | 0,465043509 |
| 25 | 0,526436181 | 0,447632755 | 0,450149076 | 0,430786383 | 0,505836027 | 0,481555994 |
| 26 | 0,532076632 | 0,431065102 | 0,435538128 | 0,430662546 | 0,523062151 | 0,479604379 |
| 27 | 0,489437767 | 0,419416239 | 0,443903894 | 0,424974627 | 0,523243009 | 0,455891508 |
| 28 | 0,538451681 | 0,456739899 | 0,452355431 | 0,41956013 | 0,518575927 | 0,467438435 |
| 29 | 0,546134201 | 0,446982501 | 0,446627142 | 0,419915714 | 0,522646511 | 0,485291608 |
| 30 | 0,506422593 | 0,444394243 | 0,451370569 | 0,432074044 | 0,526448073 | 0,4531408 |
| 31 | 0,501321198 | 0,432910965 | 0,440992719 | 0,435754267 | 0,524595083 | 0,436290024 |
| 32 | 0,502523187 | 0,447712513 | 0,45219642 | 0,437154343 | 0,526986392 | 0,456207767 |
| 33 | 0,537692745 | 0,465977765 | 0,440879995 | 0,408359072 | 0,46171041 | 0,445025801 |
| 34 | 0,539112858 | 0,438436812 | 0,455319988 | 0,444480575 | 0,471494325 | 0,469677567 |
| 35 | 0,507880712 | 0,445038122 | 0,444082951 | 0,447728206 | 0,522470853 | 0,456669209 |
| 36 | 0,532601199 | 0,459431853 | 0,453526772 | 0,41542847 | 0,509390585 | 0,470453551 |

Fonte: Elaborado pelo Autor, (2023).

A Tabela 3 apresenta a evolução da largura da silhueta simplificada usando o algoritmo K-Means correlacionando com valores de k nas bases de cada tipo de emoção. Observa-se que os maiores valores de silhueta obtidos, cerca de 0,5, ainda são baixos, o que sugere que o K-Means não conseguiu lidar com a alta dimensionalidade dos dados criando grupos com baixa coesão.

Tabela 3 – Evolução da largura da silhueta versus o número de grupos (k)




Fonte: Elaborado pelo Autor, (2023).

## 4.3 DBSCAN

Utilizando a base de dados gerada, já subdividida nas 6 (seis) bases principais dadas pelas emoções (alegria, tristeza, raiva, medo, desgosto, surpresa), efetuou-se a parametrização, por base, do DBSCAN, inserindo um raio de vizinhança (Eps) de 0,2 a 1 com intervalos de 0,2 e o MinPts variando de 10 a 10 com grupos de 10 (dez) elementos.

Após a execução do algoritmo DBSCAN, apresenta-se os resultados no software Microsoft Excel, sendo que na Tabela 4, demonstra-se, já com filtro, as linhas das bases que geraram maior e menor número de grupos (clusters).

Tabela 4 – Bases com mais e menos clusters no algoritmo DBSCAN

| **Emoção** | **eps** | **min_pts** | **clusters** | **silhueta** | **ruído** | **percentual agrupado** | **silhueta relativa** |
|---|---|---|---|---|---|---|---|
| alegria | 0.2 | 10 | 121 | 0.3519 | 28859 | 0.7772 | 0.2735 |
| medo | 0.6 | 100 | 3 | 0.5296 | 4785 | 0.7908 | 0.4188 |

Fonte: Elaborado pelo Autor, (2023).

Observa-se na Tabela 4, após a filtragem dos resultados, que a base detentora do maior número de clusters foi “alegria” com a formação de 121 grupos com 10 (dez) elementos (MinPts) em um Eps de 0,2. Já com o menor número de clusters formados (apenas 3 (três) grupos) foi a base “medo” com grupos de 100 MinPts em um Eps de 0,6. Ainda pode-se observar nos dois extremos do quantitativo de números de clusters formados que a silhueta foi maior do que 0,3 (0,03519 – alegria), porém não atingiu nem 0,6 (0,5296 – medo) de aproveitamento, inclusive apresentando altos percentuais de ruídos nos dois

casos (22,28% – alegria e 20,92% – medo). No entanto, é possível observar na Tabela 5 as 2 (duas) bases com maior e menor percentual agrupado na formação de clusters.

Tabela 5 – Bases com maior e menor percentual agrupado

| **Emoção** | **eps** | **min_pts** | **clusters** | **silhueta** | **ruído** | **percentual agrupado** | **silhueta relativa** |
|---|---|---|---|---|---|---|---|
| alegria | 1 | 10 | 17 | 0.5270 | 511 | 0.9961 | 0.5249 |
| desgosto | 0.2 | 100 | 5 | 0.8596 | 7048 | 0.2590 | 0.2227 |

Fonte: Elaborado pelo Autor, (2023).

Visualiza-se na Tabela 5 após filtragem dos resultados, que a base com maior percentual de silhueta foi "alegria" que teve 99,61% de aproveitamento ao formar 17 (dezessete) clusters de 10 MinPts em um raio de vizinhança de 1,0. Já o menor percentual de silhueta alcançado foi da base "desgosto" (25,90%) com apenas 5 (cinco) clusters formados contendo 100 elementos a um Eps de 0,2 de distância. A emoção desgosto é significativamente menos reconhecida, como relatado anteriormente em outros trabalhos na literatura, e.g. (MANCINI et al., 2018).

Nas Tabelas 4 e 5 ainda é possível observar silhueta relativa, criada com o objetivo de mensurar percentualmente a quantidade de dados que foram agrupados eficientemente pela silhueta gerada no DBSCAN.

### 4.3.1 Experimento 1

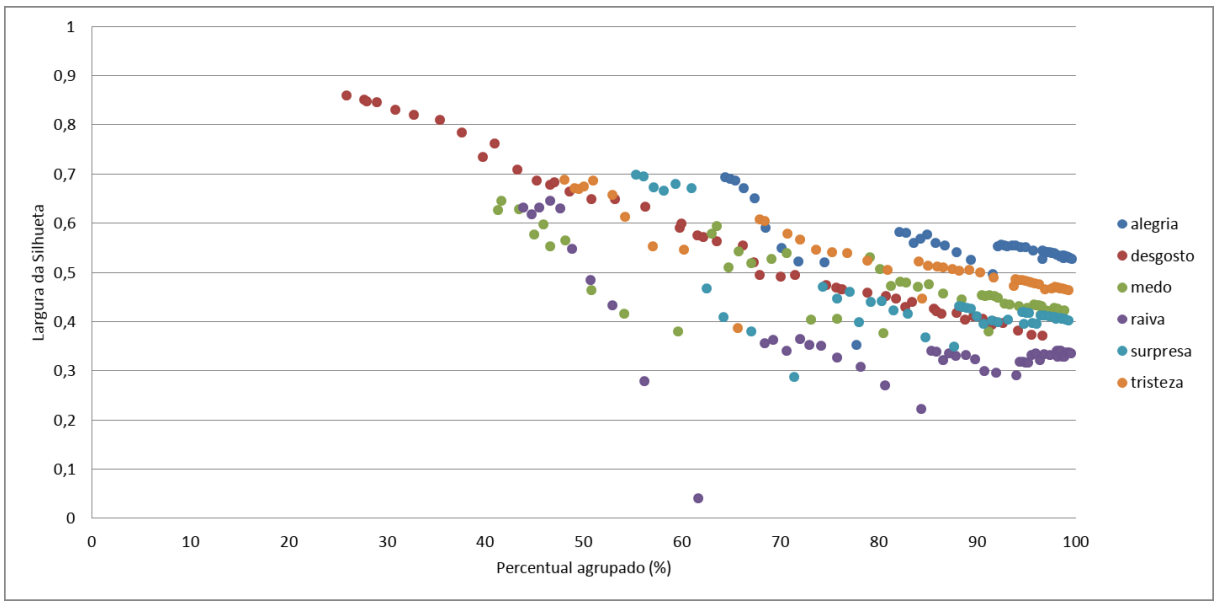
A fim de mensurar a correlação entre os dados agrupados e a largura silhueta gerada foi criado a Tabela 6.

Na Tabela 6 é possível observar que a silhueta tem o estreitamento de sua largura à medida que há um aumento no percentual dos dados agrupados, demonstrando ser variáveis de uma correlação quase que inversamente proporcional. Dada essa correlação, considerou-se que silhuetas detentoras de um percentual agrupado muito baixo deveriam ser penalizadas com base no fato de ser mais fácil identificar agrupamentos com boas larguras de silhueta utilizando apenas uma pequena parcela dos dados. Assim foi criada a largura da silhueta relativa conforme descrito na Equação 4.1:

$$\textit{Largura da silhueta relativa} = \textit{Largura da silhueta} \times \textit{percentual agrupado} \quad (4.1)$$

Considerando que a largura da silhueta relativa seja uma melhor medida de avaliação dos agrupamentos que contenham ruídos, por considerar a porcentagem dos dados que foram efetivamente agrupados, essa medida é utilizada nos experimentos sucessores.

Tabela 6 – Correlação da largura da silhueta versus do percentual agrupado (%)

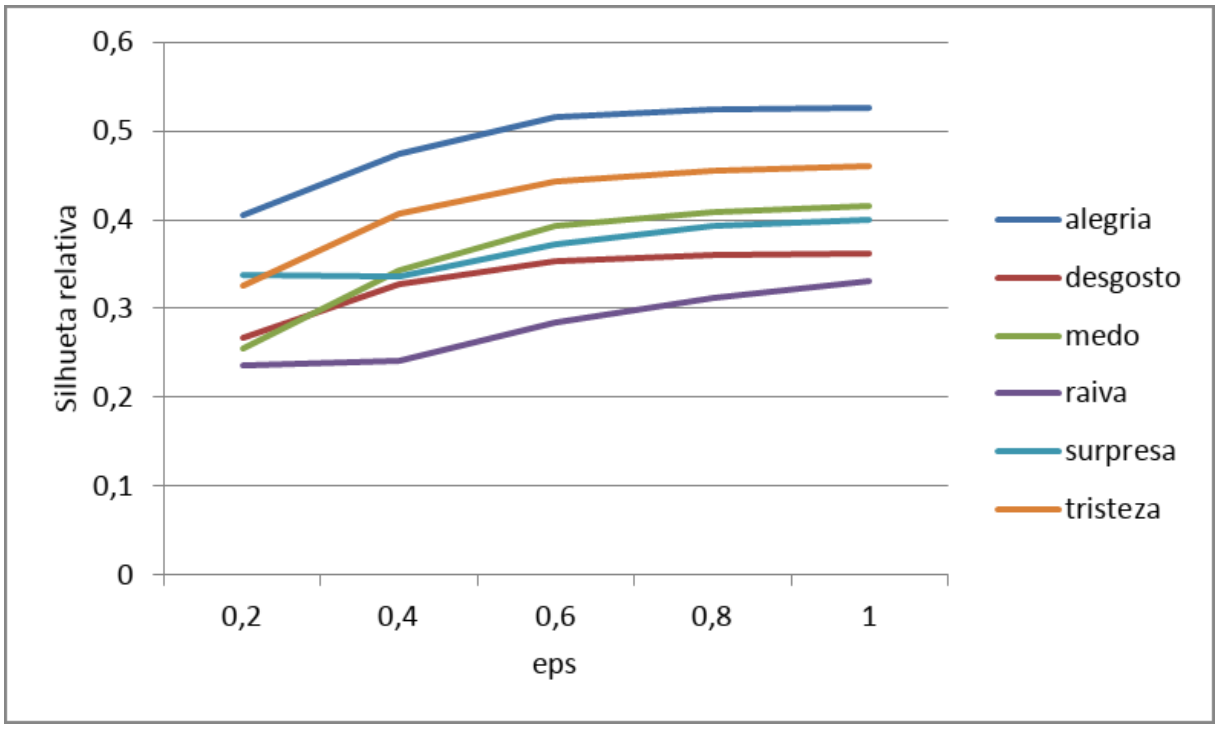



Fonte: Elaborado pelo Autor, (2023).

### 4.3.2 Experimento 2

Com objetivo de apresentar os níveis de correlação da silhueta relativa média com o Eps (raio de vizinhança), a Tabela 7 mostra a silhueta relativa média versus Eps.

Tabela 7 – Silhueta Relativa Média versus Eps



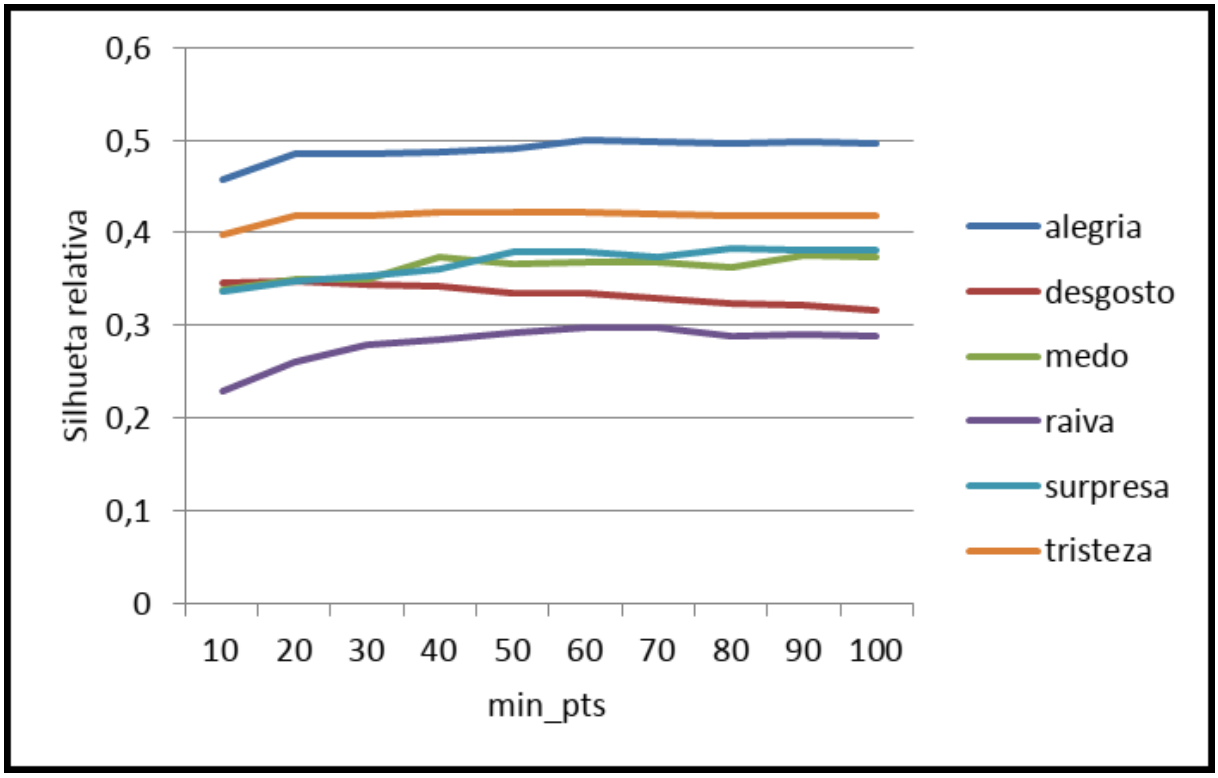

Fonte: Elaborado pelo Autor, (2023).

Na Tabela 7 nota-se que um aumento da largura da silhueta é acompanhado pelo aumento do raio (Eps). No entanto, as bases “surpresa” e “raiva” só aumentam a partir de um Eps de 0,4, sendo a média da silhueta, em um raio de 0,2 até 0,4, estável.

### 4.3.3 Experimento 3

Ainda com intuito de observar o impacto da largura da silhueta em relação aos dados inseridos, a Tabela 8 traz a visualização da correlação entre a silhueta relativa média e o MinPts (número de elementos no grupo).

Tabela 8 – Silhueta Relativa Média versus MinPts



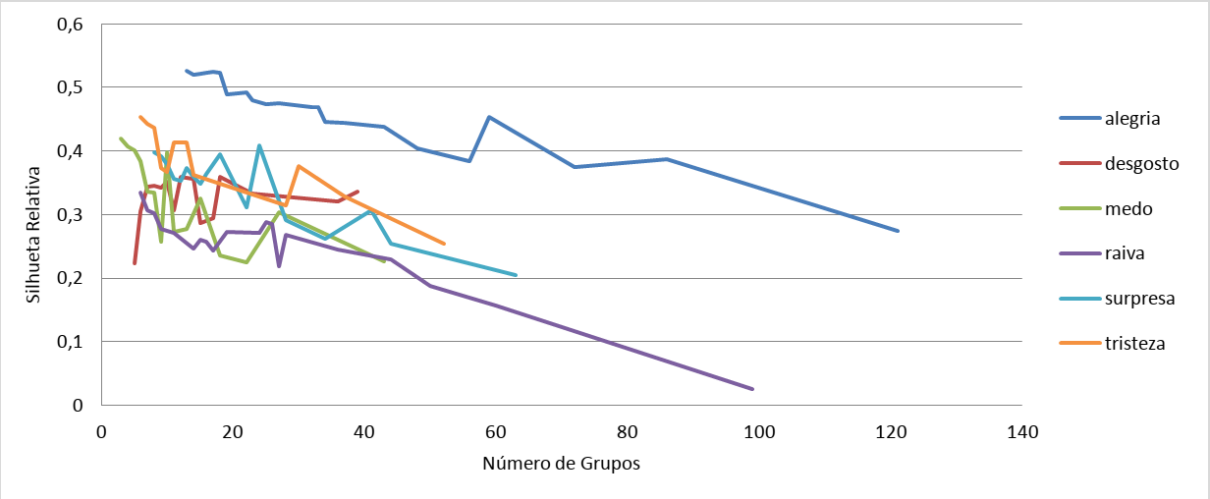

Fonte: Elaborado pelo Autor, (2023).

Na Tabela 8 é demonstrado uma tendência que todos as 6 (seis) bases agrupadas, mantém, independentemente do tamanho do grupo, uma homogeneidade na largura média da silhueta, apresentando um arco maior de crescimento nos grupos com tamanho mínimo de 10 até 20 elementos. Outro fator visual é a base “desgosto” que apresenta, contrariamente às demais, um pequeno decréscimo na largura da silhueta, estreitando à medida que os grupos formados se tornam maiores.

### 4.3.4 Experimento 4

Também foi avaliada a correlação entre a silhueta relativa média e o número de clusters (grupos) formados, conforme pode ser visualizado na Tabela 9.

Tabela 9 – Silhueta Relativa Média versus MinPts




Fonte: Elaborado pelo Autor, (2023).

Com a variação na formação de grupos, quanto às 6 (seis) bases agrupadas, o perceptível na Tabela 9 é a diferença no quantitativo formado, onde as bases "raiva" (99 grupos) e "alegria" (121 grupos) formaram muito mais grupos do que as demais bases, que figuraram em uma média de cerca de 50 grupos cada, à medida que ocorre uma diminuição da silhueta relativa média.

## 4.4 Mapas de Kohonen

Com execução de 2 a 36 grupos (6 colunas por 6 linhas de neurônios) o algoritmo Mapas de Kohonen foi executado de forma a testar todas as possibilidades de agrupamento, sendo parametrizado, com a inserção do raio de aprendizado ou largura eficiente da vizinhança (sigma $= \sigma$) de 0,3 a 1 com intervalos de 0,1, enquanto a taxa de aprendizado (learning_rate) foi inserida variando de 0,2 a 0,9 também formatada em intervalos de 0,1.

A variação de 0,1 tanto para o raio de vizinhança quanto para a taxa de aprendizado foi assi determinada com base na afirmativa de Mulier e Cherkassky (1994) de que a criação de melhores esquemas entre a taxa de aprendizado e funções de vizinhança permite contribuições mais uniformes dos dados de treinamento nas localizações das unidades, reduzindo assim as influências nos dados finais (últimos 20%) gerando grupamentos mais coesos.

Após executado o Mapas de Kohonen, os resultados são apresentados no software Microsoft Excel®, demonstrando 7296 linhas de execuções concluídas.

Na Tabela 10 é possível observar as 2 (duas) bases que apresentaram a maior e menor silhueta, após a execução do algoritmo:

Tabela 10 – Bases com maior e menor silhueta no Mapa de Kohonen

| **emoção** | **linhas** | **colunas** | **sigma** | **learning_rate** | **silhueta** | **n_clusters** |
|---|---|---|---|---|---|---|
| surpresa | 5 | 5 | 0,3 | 0,4 | 0,507497908 | 25 |
| medo | 1 | 3 | 1 | 0,6 | 0,163181734 | 3 |

Fonte: Elaborado pelo Autor, (2023).

Visualiza-se na Tabela 10 acima, após a filtragem dos resultados, que a base de maior silhueta foi a "surpresa" com um aproveitamento um pouco maior do que 50% (50,74%), sendo sua execução de 5 (cinco) linhas por 5 (cinco) colunas, com uma largura eficiente da vizinhança ($\sigma$) de 0,3 e uma taxa de aprendizado de 0,4.

Ainda é possível visualizar na mesma Tabela 10, a base que alcançou a menor silhueta (medo), com 16,31% de aproveitamento, executada com 1 (uma) linha e 3 (três) colunas, com $\sigma$ de 1 e 0,6 de learning_rate.

Enfatiza-se também que o quantitativo de grupos formados nesse algoritmo é dado em função do número de linhas multiplicado pelo número de colunas, permitindo assim, a cada base de dados trabalhada, a formação máxima de 36 grupos.

### 4.4.1 Experimento 5

Este experimento foi elaborado para identificar a influência da variação do sigma na silhueta gerada pelo Mapas de Kohonen, apresentando seus resultados na Tabela 11, abaixo:

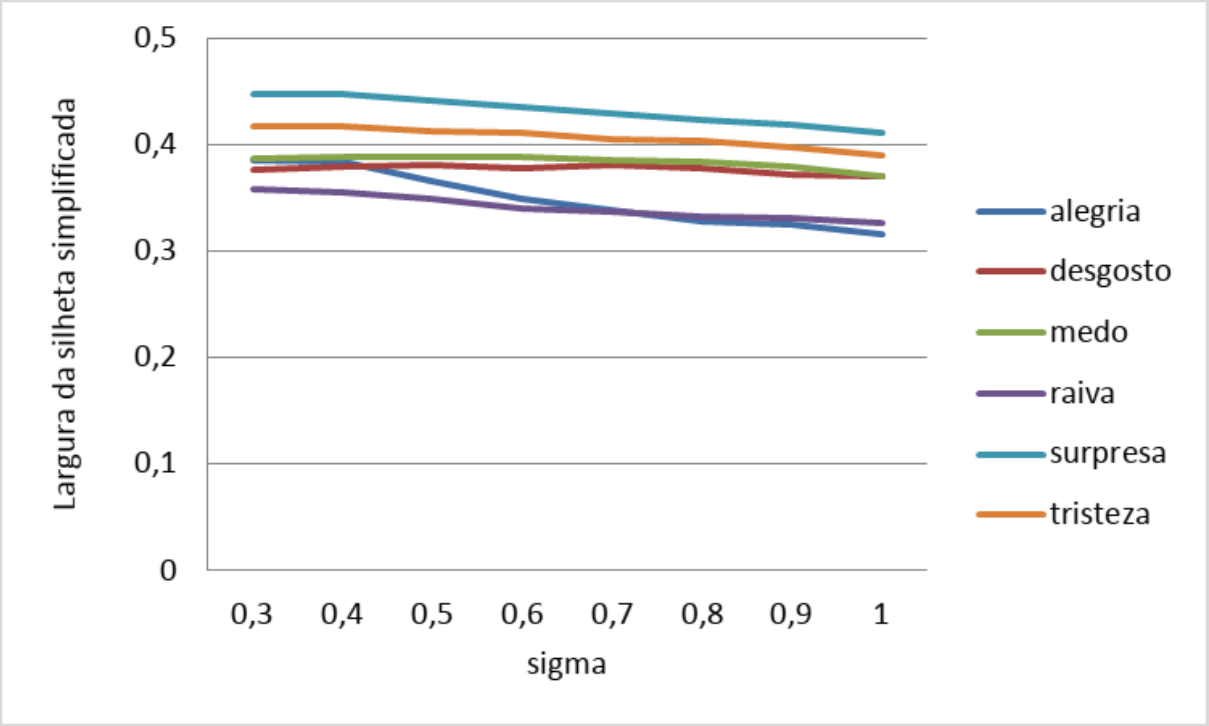
A Tabela 11 demonstra que o aumento do sigma ($\sigma$) gera a diminuição da silhueta simplificada. No entanto, deve-se enfatizar que as diminuições apresentam uniformidade em quase todas as bases, exceto na base "desgosto" onde se percebe até um leve aumento ($\sigma = 0{,}7$), mas acaba apresentando diminuição à medida que são apresentados raios de vizinhança maiores ($\sigma = 0{,}8$, 0,9 e 1). Também é perceptível a diminuição mais brusca na silhueta da base "alegria" que cai de 0,3858 com um $\sigma = 0{,}3$ para 0,3162 com $\sigma = 1$.

### 4.4.2 Experimento 6

Já a Tabela 12, abaixo, foi criada com o objetivo de apresentar a influência da variação da taxa de aprendizado na silhueta simplificada obtida:

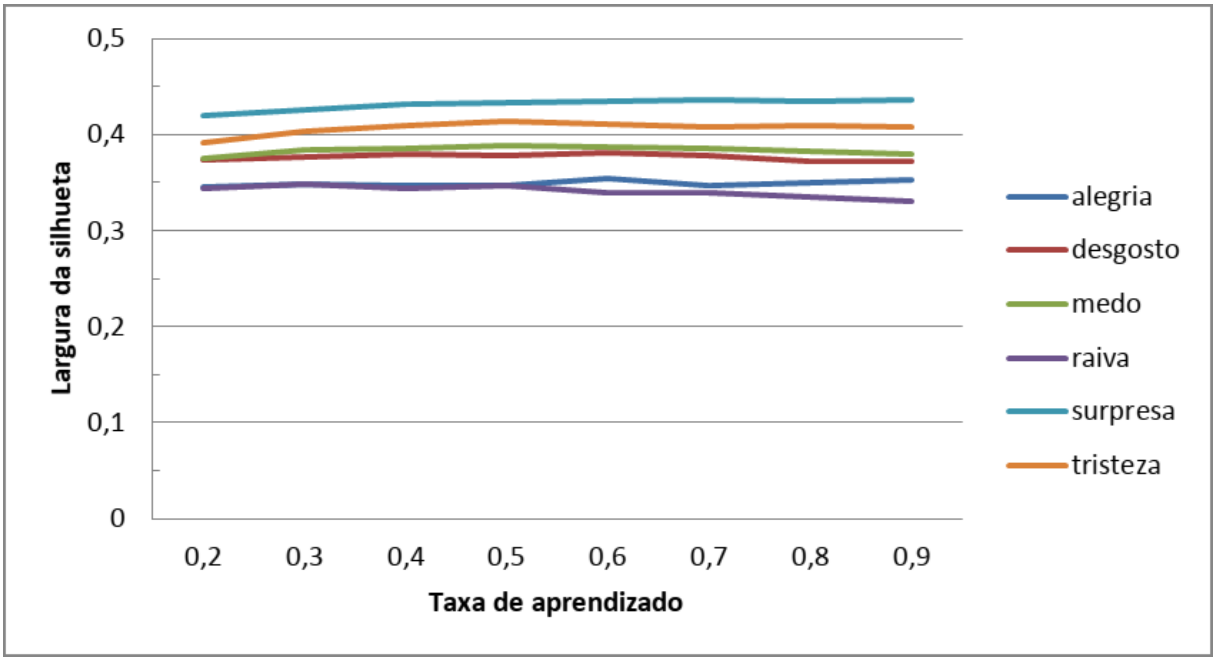
A Tabela 12 apresenta, em quase todas as bases, uniformidade no aumento da silhueta simplificada à medida que ocorre o aumento da taxa de aprendizado. As exceções ficam a cargo das bases "desgosto" e "raiva" que apresentam aumento da influência concomitantemente com o aumento da taxa de aprendizado, entre 0,5 e 0,6, mas sofrem redução da influência ao ter aumentada a learning_rate para o intervalo de 0,7 a 0,9.

Tabela 11 – Influência da variação do Sigma na silhueta




Fonte: Elaborado pelo Autor, (2023).

Tabela 12 – Influência da variação da taxa de aprendizado na silhueta

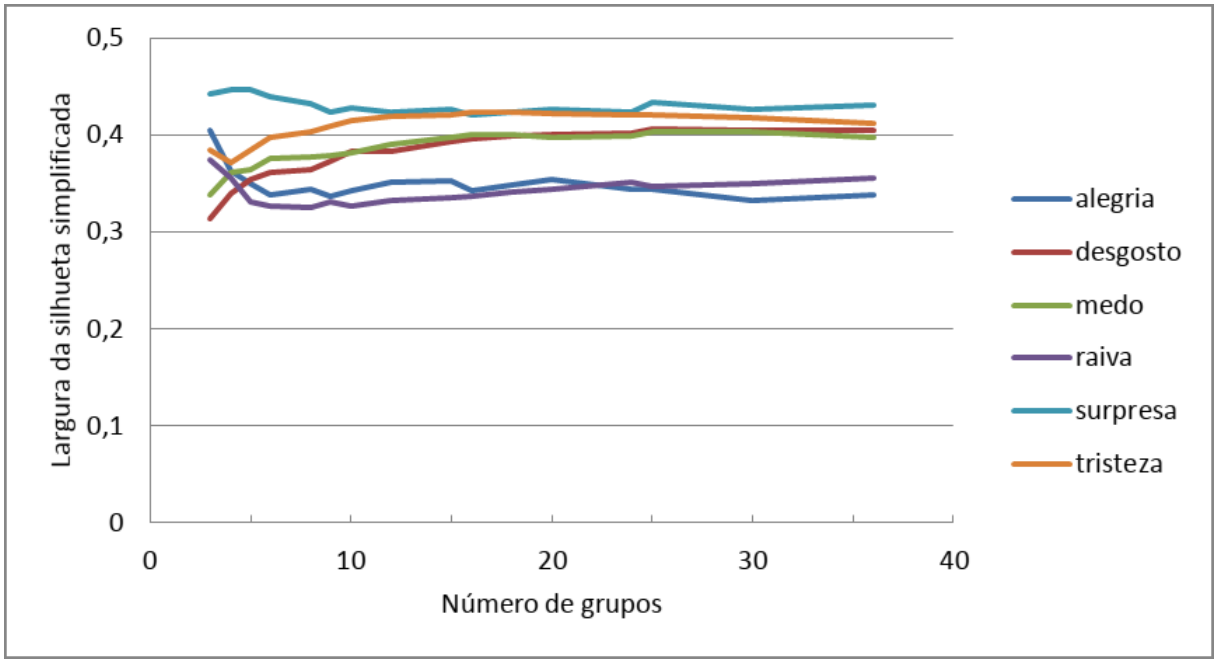



Fonte: Elaborado pelo Autor, (2023).

### 4.4.3 Experimento 7

A Tabela 13, abaixo, foi elaborada a fim de demonstrar a evolução da silhueta simplificada:

Tabela 13 – Evolução da silhueta em relação ao número de grupos




Fonte: Elaborado pelo Autor (2023).

Observa-se na Tabela 13 que na formação de 3 até 12 grupos, as bases “desgosto”, “medo” e “tristeza” tiveram um aumento da silhueta simplificada, enquanto, contrariamente, as bases “alegria”, “raiva” e “surpresa” apresentam uma redução da silhueta. No entanto, na formação de 15 a 36 grupos, visualiza-se uma manutenção uniforme da silhueta simplificada em todas as bases geradas.

## 4.5 Considerações Finais

Este capítulo apresentou os resultados obtidos pelos algoritmos K-Means, DBSCAN e Mapas de Kohonen no agrupamento da base de dados do MICHELZINHO. Considerando a variação na largura da silhueta entre 0,4 a 0,6, de forma mediana, em todos os experimentos, demonstra que os resultados dos agrupamentos, utilizando os 3 (três) algoritmos, foram relativamente relevantes, mesmo tendo sido detectada uma alta dimensionalidade das bases geradas. Os experimentos mostraram que a variação dos parâmetros de entrada não causou grande impacto na variação dos valores da silhueta.

No entanto, ainda há de se considerar que como uma grande parte da base de dados é proveniente de usuários com o espectro autista, é preciso levar em conta a grande variabilidade comportamental dessa característica, pois segundo Denney e Neuringer (1998) o estudo da variabilidade como dimensão do comportamento operante é fundamental para o desenvolvimento de estratégias de ensino para populações com dificuldades de aprendizagem, sendo que Fialho (2017) diz que reforçar respostas diferentes ou pouco frequentes

aumenta os índices de variabilidade.

Esse foi um trabalho inicial e seus resultados podem ser melhorados, mas o principal objetivo foi analisar os resultados de diferentes tipos de agrupamentos usando algoritmos clássicos sobre dados obtidos de um jogo que explora a representação das emoções. Enfatiza-se, portanto, que algoritmos mais robustos e modernos podem ser usados para melhores avaliações da silhueta.

# 5 Conclusão

Nesta pesquisa foram apresentadas técnicas de mineração de dados aplicadas à base de dados do jogo sério MICHELZINHO desenvolvido pelo autor Dantas (2022). Esse jogo permite obter informações dos músculos ativados da face durante a representação das emoções.

No processo de avaliação das técnicas de mineração de dados, a etapa de pré-processamento (transformação e validação dos dados) foi desenvolvida em linguagem Python a fim de codificar os dados utilizando a binarização 1-de-n, gerando diferentes bases de dados, o que possibilitou a exploração individualizada de cada uma das emoções elencadas pelo aplicativo. Neste trabalho foram investigados três algoritmos na mineração de dados, o K-Means, do tipo particional, o DBSCAN baseado em densidade e o Mapas de Kohonen, que por sua vez, é baseado em mapas auto-organizáveis. Assim, houve a variação em cada uma das execuções, dos parâmetros de entrada, a fim de identificar um melhor ajuste à base escolhida, corroborando em uma validação mais eficiente.

Nota-se que na validação dos agrupamentos realizados pelo K-Means e Mapas de Kohonen foi utilizada a silhueta simplificada. Já na validação dos dados agrupados pelo DBSCAN realizou-se uma adaptação da silhueta simplificada considerando a natureza do algoritmo em encontrar ruídos, chamada de silhueta relativa.

Por fim é possível analisar o comportamento desses dados sobre esses algoritmos fornecendo dados relevantes em relação ao agrupamento, visto que na maioria dos experimentos o valor da largura da silhueta ficou entre 0,4 e 0,6. Desse modo é possível encontrar alguns indicativos que podem direcionar psicólogos, tutores e pesquisadores da área. Pressupõe-se ainda que também é possível levantar quais hipóteses podem ser avaliadas em trabalhos futuros.

## 5.1 Dificuldades Encontradas

Considera-se relevante ressaltar as principais dificuldades encontradas neste trabalho:

- Implementação do algoritmo de agrupamento baseado em grafos, CAST, pois foi necessária a leitura do mesmo depois de implementado, foi necessário descartar o algoritmo CAST já que esse tem complexidade quadrática;
- Foram usadas várias máquinas com diferentes tamanhos e foram feitas inclusive sucessivas solicitações a AWS de forma que liberasse mais capacidade com instâncias

maiores, por fim executando com uma c6g.metal que tem 128GB de memória RAM e 64vCPU de processamento.

Por fim vale ressaltar que apesar do curso de Ciência da computação como um todo ter contribuído para a superação dessas dificuldades, a disciplina de Análise de Algoritmos ajudou especialmente na avaliação do desempenho da largura da silhueta e algoritmo CAST,bem como a disciplina de Tópicos Especiais de Inteligência Artificial, na qual foram apresentados os conceitos de mineração de dados que foram utilizados ao longo de todo esse trabalho.

## 5.2 Principais Contribuições

Este trabalho apresenta as seguintes contribuições:

- Desenvolvimento de uma ferramenta, em python, capaz de realizar o pré-processamento das bases de dados do MICHELZINHO, podendo inclusive ser adaptada para utilização em outras bases de dados;
- Estudo de agrupamentos usando diferentes algoritmos e variando os parâmetros de entrada;
- Adaptação do método da largura da silhueta, através da mensuração da quantidade de ruídos presentes nos dados agrupados, aumentando a eficiência do processo de agrupamento, utilizando a silhueta relativa.

## 5.3 Trabalhos Futuros

Como trabalhos futuros, recomenda-se:

- Utilização de diferentes técnicas de agrupamento, como os baseados em hierarquia e baseados em grafos;
- Identificação do usuário na base do MICHELZINHO, a fim de encontrar diferentes comportamentos do mesmo ao longo da utilização do aplicativo;
- Comparar os resultados de agrupamentos, utilizando algoritmos variados, a fim de possibilitar a identificação de semelhanças que possam indicar nichos de usuários, que possam ser atendidos com um mesmo tipo ferramental.

# Referências

A. Ben-Dor, R. Shamir and Z. Yakhini. Clustering gene expression patterns. 1999 J. Comput. Biol., vol. 6. Disponível em <https://dl.acm.org/doi/pdf/10.1145/299432.299448>. Acesso em: 26 jan 2023

AFFONSO, Gustavo Souza.Mapas auto - organizáveis de kohonen (SOM) aplicados na avaliação dos parâmetros da qualidade da água. 2011. Dissertação (Mestrado em Tecnologia Nuclear - Reatores) - Instituto de Pesquisas Energéticas e Nucleares, Universidade de São Paulo, São Paulo, 2011. doi:10.11606/D.85.2011.tde-31102011-101817. Disponível em: <https://teses.usp.br/teses/disponiveis/85/85133/tde-31102011-101817/publico/2011AffonsoMapas.pdf>. Acesso em: 2021-12-27.

ALMEIDA, Cleibson Aparecido de. et al. Melhoria na qualidade de dados com a aplicação de "data cleaning" na base de dados de acidentes aeronáuticos da aviação civil brasileira. AtoZ: novas práticas em informação e conhecimento, [S.l.], v. 5, n. 2, p. 72-79, dez. 2016. ISSN 2237-826X. Disponível em: <https://revistas.ufpr.br/atoz/article/view/47303>. Acesso em: 06 nov. 2021. doi:http://dx.doi.org/10.5380/atoz.v5i2.47303.

ALMEIDA, S. S. A.; MAZETE, B. P. G. S.; BRITO, A.R.; VASCONCELOS, M. M.. Transtorno do espectro autista. Residência Pediátrica. 2018;8 (0 Supl.1):72-78 DOI: 10.25060/residpediatr-2018.v8s1-12. Disponível em: <https://residenciapediatrica.com.br/detalhes/345/transtorno%20do%20espectro%20autista>. Acesso em: 08 dez. 2022.

AMERICAN PSYCHIATRIC ASSOCIATION (2014). DSM-5: Manual Diagnóstico e Estatístico de Transtornos Mentais (5ª Ed). Porto Alegre: Artmed. Disponível em: <http://www.niip.com.br/wp-content/uploads/2018/06/Manual-Diagnosico-e-Estatistico-de-Transtornos-Mentais-DSM-5-1-pdf>. Acesso em: 09 jan. 2022.

AMO, S. de; ROC, C. Curso de Data Mining - Universidade Federal de Uberlândia, 2003. Disponível em: <http://www.fatecead.com.br/tei/semana08-1_livro_mineracaodados.pdf>. Acesso em: 06 nov. 2021.

BAPTISTA, C. R.; BOSA, C. Autismo e educação: Reflexões e propostas de intervenção. [S.l.]: Artmed Editora, 2002.

BARON-COHEN, S. et al. Empathizing and systemizing in autism spectrum conditions. Handbook of autism and pervasive developmental disorders, JohnWiley  Sons Inc. New Jersey, v. 1, p. 628–639, 2005.

BARROS, Everton ROMÃO, Wesley CONSTANTINO, Ademir SOUZA, Celso. (2011). Pré-processamento para mineração de dados sobre beneficiários de planos de saúde suplementar. Journal of Health Informatics. Disponível em: <https://www.researchgate.net/publication/277733249_Pre-processamento_para_mineracao_de_dados_sobre_beneficiarios_de_planos_de_saude_suplementar>. Acesso em: 07 nov. 2021.

BATISTA, Gustavo Enrique de Almeida Prado Alves. Pré-processamento de dados em aprendizado de máquina supervisionado. 2003. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2003. doi:10.11606/T.55.2003.tde-06102003-160219. Acesso em: 11 dez. 2021.

BENUTE, Gláucia R. G. (Org). Transtorno do espectro autista (TEA): desafios da inclusão. Volume 2 – São Paulo: Setor de Publicações - Centro Universitário São Camilo, 2020. – (Coleção Ensaios sobre Acessibilidade). 50 p. Disponível em: <https://saocamilo-sp.br/_app/views/publicacoes/outraspublicacoes/nape_volume_02_13abr_FINAL.pdf>. Acesso em: 09 jan. 2022.

BORGES, V. R. P. Comparação entre as técnicas de agrupamento k-means e fuzzy c-means para segmentação de imagens coloridas. VII Encontro Nacional de Computação – Catalão, Goiás. 2010. Disponível em: <https://www.researchgate.net/publication/341593009_Comparacao_entre_as_Tecnicas_de_Agrupamento_K-Means_e_Fuzzy_C-Means_para_Segmentacao_de_Imagens_Coloridas>. Acesso em: 24 dez. 2021.

BRASIL. Ministério da Educação. Dispõe sobre a educação especial, o atendimento educacional especializado e dá outras providências. Decreto 7611 de 17 de novembro de 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Decreto/D7611.htmart11>. Acesso em: 08 jan. 2022.

BRASIL. Presidência da República. Lei Geral de Proteção de Dados Pessoais – Lei nº 13.853, de 8 de julho de 2029. Altera a Lei nº 13.709, de 14 de agosto de 2028, para dispor sobre a proteção de dados pessoais e para criar a Autoridade Nacional de Proteção de Dados, e dá outras providências. Disponível em: <http://www.planalto.gov.br/ccivil_03/$_{A}to2019-2022/2019/Lei/L13853.htmart1 > .Acessoem : 15ago.2022$.

CALIL, Leonardo CARVALHO, Deborah SANTOS, Celso GOMES VAZ, Maria Salete. (2008). Mineração de dados e pós-processamento em padrões descobertos. Publication UEPG - Ciências Exatas e da Terra Agrarias e Engenharias. 14. 207-215. 10.5212/Publi. Exatas.v.3i2.207215. Acesso em: 06 nov. 2021.

CAMILO, Cássio Oliveira SILVA, João Carlos Da. Mineração de dados: Conceitos, Tarefas, Métodos e Ferramentas. Instituto de Informática – Universidade Federal de

Goiás – Technical Report – RT-INF_001-09 – Relatório Técnico – 2009. Disponível em: <https://rozero.webcindario.com/disciplinas/fbmg/dm/RT-INF_001-09.pdf>. Acesso em: 05 dez. 2021.

CARVALHO, Luís A. V. De. Data mining: a mineração de dados no marketing, medicina, economia, engenharia e administração. Rio de Janeiro: Ciência Moderna, 2005.

CASSIANO, Keila Mara. Análise de séries temporais usando Análise Espectral Singular (SSA) e clusterização de suas componentes Baseada em Densidade. Tese (Doutorado em Engenharia Elétrica) – Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2014. Disponível em: <https://www.maxwell.vrac.puc-rio.br/colecao.php?strSecao=resultadonrSeq=24787@1>. Acesso em: 27 nov. 2021.

CASSIANO, Keila SOUZA, Reinaldo PESSANHA, José. Combinação de clusterização baseada em densidade com análise espectral singular de séries temporais - uma aplicação à velocidade do vento. XVII Simpósio de Pesquisa Operacional e Logística da Marinha – SPOLM, Vol 1, n. 1 – Agosto 2014. Disponível em: <http://pdf.blucher.com.br.s3-sa-east-1.amazonaws.com/marineengineeringproceedings/spolm2014/127220.pdf>. Acesso em: 27 nov. 2021.

CASTANHEIRA, L. G. Aplicação de técnicas de mineração de dados em problemas de classificação de padrões. Dissertação (Mestrado) Universidade Federal de Minas Gerais, 2008. Disponível em: <https://www.ppgee.ufmg.br/documentos/Defesas/777/Dissertacao_LucianaCastanheira.pdf>. Acesso em: 07 nov. 2021.

CEBECI, Zeynel; KAVLAK, Alper Tuna; YILDIZ, Figen. Validation of fuzzy and possibilistic agrupamento results. In: Artificial Intelligence and Data Processing Symposium (IDAP), 2017 International. IEEE, 2017. p. 1-7.

CELINSKI, T. M. Métodos de agrupamento: uma abordagem comparativa com aplicação em segmentação de imagens de profundidade. Dissertação. 142p. (Mestrado em Informática), Universidade Federal do Paraná – UFPR, Curitiba, 1998. Disponível em: <http://ri.uepg.br/riuepg/bitstream/handle/123456789/912/DISSERTA%C3%87%C3%83O_TatianaMontesCelinski.pdf?sequence=1>. Acesso em: 24 dez. 2021.

CHIU, S.; TAVELLA, D. Introduction to Data Mining. 2nd editio. ed. [S.l.]: Pearson, 2008. 137–192 p. ISBN 9780133128901.

CHRISTENSEN D, BAIO J, BRAUN K, BILDER D, CHARLES J, CONSTANTINO J, et al. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged

8 Years – Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012; 2018

DANTAS, Adilmar C.; MELO, Sara de; NEVES, Leandro; MILESSI, Taynara; NASCIMENTO, Marcelo Z. do . Michelzinho: Jogo sério para o ensino de habilidades emocionais em pessoas com autismo ou deficiência intelectual. In: XXX Simpósio Brasileiro de Informática na Educação (Brazilian Symposium on Computers in Education), 2019, Brasília. Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE 2019), 2019. p. 644.

DANTAS, Aldimar Coelho. Métodos Computacionais para Aprendizagem das Emoções em Indivíduos com Autismo. Qualificação (Doutor em Ciências de Computação) - Faculdade de Computação, Universidade Federal de Uberlândia, Uberlândia, 2020.

DANTAS, Adilmar Coelho. Abordagem computacional para aprimoramento das habilidades com as emoções em indivíduos com autismo. 2022. 115 f. Tese (Doutorado em Ciência da Computação) - Universidade Federal de Uberlândia, Uberlândia, 2022. DOI http://doi.org/10.14393/ufu.te.2022.211.

DANTAS, Adilmar C, NASCIMENTO, Marcelo Z. Recognition of Emotions for People with Autism: An Approach to Improve Skills. International Journal of Computer Games Technology, v. 2022, p. 1-21, 2022.

DAPOGNY, A. et al. Jemime: a serious game to teach children with asd how to adequately produce facial expressions. In: IEEE. 2018 13th IEEE International Conference on Automatic Face  Gesture Recognition (FG 2018). [S.l.], 2018. p. 723–730.

Denney, J.,  Neuringer, A. (1998). Behavioral variability is controlled by discriminative stimuli. Animal Learning  Behavior, 26 (2), 154-162.

DE PINHO, Anderson Guimarães. Mineração de dados com mapas de Kohonen: Uma abordagem no setor financeiro. Revista Pensamento Contemporâneo em Administração. 2008, 2(1). p. 39-49. Disponível em:
<https://www.redalyc.org/articulo.oa?id=441742832004>. Acesso em: 27 dez. 2021.

DIAS, M. M. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. Acta Scientiarum. Technology, v. 24, p. 1715-1725, 2002.

DIAS M.;YAMAGUCHI K.; RABELO E. FRANCO C. Visualization Techniques: Which is the Most Appropriate in the Process of Knowledge Discovery in Data Base?. Book edited by Adem Karahoca, ISBN 978-953-51-0748-4. under CC BY 3.0 license, Sept. 2012. Disponível em: <https://www.intechopen.com/chapters/39033>. Acesso em: 01 jan. 2022.

DONG, G. LI J. Interestingness of discovered association rules in terms of neighborhood-based unexpectedness. In: Proceedings of the 2nd Pacific-Asia Conf. Knowledge Discovery and Data Mining, PAKDD, Lecture Notes in Artificial Intelligence, LNAI, Volume 1394, Melbourne, Australia, p. 72-86. Disponível em: <https://corescholar.libraries.wright.edu/cgi/viewcontent.cgi?article=1356context=knoesis>. Acesso em: 01 jan. 2022.

DONI, Marcelo Viana. Análise de Cluster: Métodos Hierárquico e de Particionamento. Trabalho de Graduação Interdisciplinar II (Bacharelado em Sistemas de Informação) - Universidade Presbiteriana Mackenzie, São Paulo, 2004. Disponível em: <http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF>. Acesso em: 27 nov. 2021.

EKMAN, Paul – Trad. SZLAK, Carlos. A linguagem das emoções: Revolucione sua comunicação e seus relacionamentos reconhecendo todas as expressões das pessoas ao redor. São Paulo: Lua de Papel, 2011. Disponível em: <https://moodle.ufsc.br/pluginfile.php/3731743/mod_resource/content/3/Paul%20Ekman%20-%20A%20Linguagem%20das%20Emoc%CC%A7o%CC%83es%20-%20Capi%CC%81tulos%201%20a%204.pdf>. Acesso em: 18 ago. 2021.

EKMAN, P.; FRIESEN,W. V.; HAGER, J. C. The facial action coding system. Research Nexus eBook, Salt Lake City, UT, 2002.

EYNG, Alini Marangoni. Análise de agrupamento pelos métodos hierárquico aglomerativo e particional Fuzzi utilizados para Educational Data Mining em dados de educação à distância. Trabalho de Conclusão de Curso (Bacharel em Ciência da Computação), UNESC - Universidade do Extremo Sul Catarinense, Criciúma. 2019. Disponível em: <http://repositorio.unesc.net/bitstream/1/8174/1/ALINI%20MARANGONI%20EYNG.pdf>. Acesso em: 02 jan. 2022.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: The Knowledge Discovery and Data Mining Conferences. [s.n.], 1996. v. 96, n. 34, p. 226/231. Disponível em: <http://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf>. Acesso em: 24 dez. 2021.

FAYYAD, U.; PIATESKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge Discovery in databases. Artificial Intellingence Magazine, v. 17, n.3, p. 37, 1996.

FAYYAD, U. et al. Advances in knowledge discovery and data mining. American Association for Artificial Intelligence. Menlo Park, CA: MIT Press. 1996.

FERNANDES, Naraline Alvarenga. Uso de jogos educacionais no processo de ensino e de aprendizagem. Trabalho de Conclusão de Curso (Especialista em Mídias da Educação) – CINTED/UFRGS – Centro Interdisciplinar de Novas Tecnologias na Educa-

ção da Universidade Federal do Rio Grande do Sul, Alegrete/RS, 2010. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/141470/000990988.pdf>. Acesso em: 16 jan. 2022.

FIALHO, Juliana. A variabilidade comportamental no autismo. Portal Comporte-se – Psicologia Ac. 2017. Disponível em: <https://comportese.com/2017/09/18/variabilidade-comportamental-no-autismo/>. Acesso em: 31 dez. 2022.

FUNDAÇÃO ABRINQ. População estimada pelo IBGE segundo faixas etárias. 2021. Disponível em: <https://observatoriocrianca.org.br/cenario-infancia/temas/populacao/1048-populacao-estimada-pelo-ibge-segundo-faixas-etarias?filters=1,1620>. Acesso em: 08 jan. 2022.

FUNDAÇÃO ABRINQ. População estimada pelo IBGE segundo faixas etárias. 2021. Disponível em: <https://observatoriocrianca.org.br/cenario-infancia/temas/populacao/1048-populacao-estimada-pelo-ibge-segundo-faixas-etarias?filters=1,1621>. Acesso em: 08 jan. 2022.

GIACOMELLI, Daniela Freitas. Técnicas de agrupamento aplicadas aos dados de acidentes de trabalho [recurso eletrônico]. Dissertação (Mestrado em Ciências de Computação) - Faculdade de Computação, Universidade Federal de Uberlândia, Uberlândia, 2020.

GOLDSCHMIDT, R.; PASSOS, E. Data Mining. [S.l.]: Elsevier Brasil, 2015.

GOLDSCHMIDT, R.; PASSOS, E. Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações. [S.l.]: Elsevier Brasil, 2017.

GOMES P, LIMA L, BUENO M, ARAÚJO L, SOUZA N. Autism in Brazil: a systematic review of family challenges and coping strategies. Rio de Janeiro: J Pediatria. 2015; 91:111-21. Disponível em: <https://www.scielo.br/j/jped/a/wKsNY3ngvLDcRZ5bxWCn47v/?format=pdflang=pt>. Acesso em: 08 jan. 2022.

GROSS, J. L. G. URSA: um framework para agrupamento de dados e validação de resultados. Monografia (Bacharel em Ciência da Computação), Universidade Federal do Rio Grande do Sul, Porto Alegre. 2014. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/110328/000952575.pdf?sequence=1>. Acesso em: 24 dez. 2021.

GROSSARD, C. et al. Serious games to teach social interactions and emotions to individuals with autism spectrum disorders (asd). Computers Education, Elsevier, v. 113, p. 195–211, 2017.

HAIR, J. F. et al. Análíse multivariada de dados. 6 ed. Porto Alegre: Bookman, 2010. 688p.

HAN, J. KAMBER, M.. Data Mining: Concepts and Techniques - Second Edition. Elsevier. 2006.

HAN, J.; PEI, J.; KAMBER, M. Data Mining: Concepts and techniques. [S.l.]: Elsevier, 2011. 83-445 p.

HAND, D. J. Principles of Data Mining. Drug safety, Springer, v. 30, n. 7, p. 621-622, 2007. Https://doi.org/10.2165/00002018-200730070-00010.

HAYKIN, S.. Redes Neurais - Princípios e prática. Paulo Martins Engel. 2ª ed. Porto Alegre: Editora Bookman, 2001.

HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. Understanding robust and exploratory data analysis. New York: John Wiley Sons, 1983. 447p.

HOWES O, ROGDAKI M, FINDON J, WICHERS R, CHARMAN T, KING B, et al. Autism spectrum disorder: Consensus guidelines on assessment, treatment and research from the British Association for Psychopharmacology. J Psychopharmacol. 2017; 1-27. Disponível em: <https://www.bap.org.uk/pdfs/BAP_Guidelines-ASD.pdf>. Acesso em: 09 jan. 2022.

JAIN, A.K. (2009) Data clustering: 50 years beyond K-means. Pattern Recognition Letters. (2009), doi:10.1016/j.patrec. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0167865509002323>. Acesso em: 12 dez. 2021.

KOHONEN, Teuvo. Essentials of the self-organizing map. Neural networks : the official journal of the International Neural Network Society. (2012). 37. 10.1016/j.neunet. 2012.09.018. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0893608012002596>. Acesso em: 26 dez. 2021.

LACAVA, P. G. et al. Using assistive technology to teach emotion recognition to students with asperger syndrome: A pilot study. Remedial and Special Education, Sage Publications Sage CA: Los Angeles, CA, v. 28, n. 3, p. 174–181, 2007.

LAROSE, D. T. Discovering Knowledge in Data: an Introduction to Data Mining. John Wiley and Sons, Inc, 2005.

LAROSE, D. T. Discovering knowledge in data: an introduction to Data Mining. [S.l.]: John Wiley Sons, 2014.

LEE, Huei Diana. Seleção de atributos importantes para a extração de conhecimento de bases de dados. Tese (Doutorado em Ciências de Computação e Matemática Computaci-

onal) - Instituto de Ciências Matemáticas e de Computação, University of São Paulo, São Carlos, 2005. doi:10.11606/T.55.2005.tde-22022006-172219. Acesso em: 01 nov. 2021.

LENT, R. Neurociência da mente e do comportamento. Rio de Janeiro: Guanabara Koogan, 2013.

LI, K.-H. et al. The effects of applying game-based learning to webcam motion sensor games for autistic students' sensory integration training. Turkish Online Journal of Educational Technology-TOJET, ERIC, v. 11, n. 4, p. 451–459, 2012. Disponível em:<https://files.eric.ed.gov/fulltext/EJ989340.pdf>. Acesso em: 16 jan. 2021.

MACÁRIO FILHO, Valmir. Algoritmos particionais semissupervisionados com ponderação automática de variáveis. Tese (Doutorado em Ciência da Computação) - Universidade Federal de Pernambuco, Recife, 2015. 165 f. Disponível em: <https://repositorio.ufpe.br/bitstream/123456789/15260/1/TESE%20Valmir%20Macario%20Filho.pdf>. Acesso em: 24 dez. 2021.

MACHADO, Felipe Nery Rodrigues. Tecnologia e projeto de Data Warehouse. São Paulo: Editora Érica, 2004.

MANCINI, G., BIOLCATI, R., AGNOLI, S., Andrei, F., Trombini, E. (2018). Recognition of facial emotional expressions among Italian pre-adolescents, and their affectivereactions. Frontiers in Psychology, 9, 1303. Disponível em: <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01303/full>. Acsso em 26 jan. 2023.

MENDES, Jakelson Carreiro. Agrupamento de Dados e suas Aplicações. Monografia (Bacharelado em Ciência da Computação) - Universidade Federal de Maranhão, São Luís, 2017. Disponível em: <https://monografias.ufma.br/jspui/bitstream/123456789/3570/1/JAKELSON-MENDES.pdf>. Acesso em: 24 dez. 2021.

MELO, A.. Emoções no período escolar: estratégias parentais face à expressão emocional e sintomas de internalização e externalização da criança. Dissertação de mestrado, Universidade do Minho, Porto. 2005. Disponível em: <http://repositorium.sdum.uminho.pt/bitstream/1822/4926/1/TESE%20MESTRADO%20ANA%20MELO.pdf>. Acesso em: 14 mai. 2022.

METRI, P.; GHORPADE, J.; BUTALIA, A. Facial emotion recognition using context based multimodal approach. International Journal of Emerging Sciences, Springfield Publishing Corporation, v. 2, n. 1, p. 171–183, 2012.

F. Mulier and V. Cherkassky, Learning rate schedules for self-organizing maps, Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 -

Conference C: Signal Processing (Cat. No.94CH3440-5), 1994, pp. 224-228 vol.2, doi: 10.1109/ICPR.1994.576908.

NASS, C.; BRAVE, S. Emotion in human-computer interaction. In: The human-computer interaction handbook. [S.l.]: CRC Press, 2007. p. 94–109.

OLIVEIRA, K. G.; SERTIÉ, A. L. Transtornos do espectro autista: um guia atualizado para aconselhamento genético. Einstein, 15(2): 233-8, 2017. Disponível em: <https://www.scielo.br/j/eins/a/YMg4cNph3j7wfttqmKzYsst/?lang=pt>. Acesso em: 08 jan. 2022.

OMS. Autism Spectrum Disorders, Key Facts@ONLINE. 2021. Disponível em: <http://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders>. Acesso em: 08 jan. 2022.

PEDREGOSA et al.. Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, 2011. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.htmlexamples-using-sklearn-metrics-silhouette-score>. Acesso em: 17 ago. 2022.

PESSINI, Adriano et al. O uso de Jogos Sérios na Educação em Informática: Um Mapeamento Sistemático. Nuevas Ideas en Informática Educativa TISE. p.537-541. 2014. Disponível em: <http://www.tise.cl/volumen10/TISE2014/tise2014_submission_105.pdf>. Acesso em: 23 jan. 2022.

PINTO, Amâncio da Costa. Psicologia Geral. Lisboa: Universidade Aberta, 2001. 340 pg.

PRENSKY, M. Digital game-based learning. Computers in Entertainment (CIE), ACM, v. 1, n. 1, p. 21–21, 2003.

PUGLIESI, J. B. Pós-processamento de Regras de Regressão. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. 2004. Disponível em: <https://teses.usp.br/teses/disponiveis/55/55134/tde-20082015-111001/publico/JaquelineBrigladoriPugliesi.pdf>. Acesso em: 01 jan. 2022.

RAHM, E., DO, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3– 13. Recuperado de http://sites.computer.org/debull/A00dec/ issue1.htm

RAMOS, C. M.; LOBO, F. Descoberta de conhecimentos em base de dados. Dos Algarves, n. 12, p. 53Ű59, 2003.

RAMOS, D; CRUZ, D. M. (org). Jogos digitais em contextos educacionais. São Paulo: CRV, 2018.

RAMOS, D. K.; SILVA, G. A.; MACEDO, C. C. Jogos digitais e emoções: um estudo exploratório com crianças. Revista Pedagógica, Chapecó, v. 22, p. 1-21, 2020. DOI: https://doi.org/10.22196/rp.v22i0.4314

RIVA, Carmen. Novos tempos, novas crianças. 2009. Disponível em: <http://www.dihoje.com.br/dihoje2009/?pg=noticiaid=1360>. Acesso em: 16 jan. 2022.

ROUSSEEUW, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computacional and Applied Mathematics, v. 20: pp. 53-65. Disponível em: <https://www.sciencedirect.com/science/article/pii/0377042787901257>. Acesso em: 15 nov. 2022.

SAMPAIO, S.; FREITAS, I. B de (Orgs.). Transtornos e dificuldades de aprendizagem: entendendo melhor os alunos com necessidades educativas especiais. Rio de Janeiro: Wak editora, 2011.

SAVI, Rafael; ULBRICHT, Vania R. Jogos digitais educacionais: benefícios e desafios. UFRGS. Porto Alegre. 2008. Disponível em: <https://seer.ufrgs.br/index.php/renote/article/view/14405/8310>. Acesso em: 16 jan. 2022.

SEMAAN, G. S. Algoritmos para o Problema de Agrupamento Automático. Tese (Doutorado) Tese de Doutorado, Instituto de Computação, Universidade Federal Fluminense, 2013. Disponível em: <http://www.ic.uff.br/index.php/pt/pos-graduacao/teses-e-dissertacoes>. Acesso em: 24 dez. 2021.

SILBERSCHATZ, A.; TUZHILIN, A. On subjective measures of interestingness in knowledge Discovery. Proceedings of the First International Conference on Knowledge Discovery and Data Mining – KDD'95 1, 275-281. Disponível em: <https://www.aaai.org/Papers/KDD/1995/KDD95-032.pdf>. Acesso em: 01 jan. 2022.

SILVA, Danilo Arantes da. Aplicações de técnicas de pré-processamento e agrupamento na base de dados de benefícios previdenciários do Ministério Público do Trabalho. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) - Universidade Federal de Uberlândia, Uberlândia, 2018.

SILVA, G. A. da; RAMOS, D. K.; RIBEIRO, F. L. Formação Inicial de Professores à Distância para o Uso das Tecnologias Digitais: Um Estudo dos Projetos Pedagógicos dos Cursos de Licenciatura da Universidade Aberta do Brasil/UFSC. Anais do Simpósio Ibero- Americano de Tecnologias Educacionais, 2019.

SILVA, Michel de Almeida. O pré-processamento em mineração de dados como método de suporte à modelagem algorítmica. Dissertação (Mestrado em Modelagem Computacional de Sistemas) - Universidade Federal do Tocantins, Brasil, 2014.

SILVA, Tassyo Tchesco ABREU, Andrêssa Finzi de. Avaliação de algoritmos para estimação do número de grupos em problemas de mineração de dados. In: VII ENTEC - Encontro de Tecnologia da UNIUBE, Universidade de Uberaba, Campus Aeroporto, Uberaba, MG. 2018. Disponível em: <https://www.uniube.br/eventos/entec/2011/arquivos/sistemas5.pdf>. Acesso em: 15 nov. 2022.

SOUZA, J. C., FRAGA, L. L., OLIVEIRA, M. R., BUCHARA, M. S., STRALIOTTO, N. C., ROSÁRIO, S. P., REZENDE, T. M. (2004). Atuação do psicólogo frente aos transtornos globais do desenvolvimento infantil. Psicologia: Ciência e Profissão, 24(2), 24-31. doi: 10.1590/S1414-98932004000200004. Disponível em: <http://pepsic.bvsalud.org/pdf/pcp/v24n2/v24n2a04.pdf>. Acesso em: 09 jan. 2022.

SOORYA L, CARPENTER L, EL-GHOROURY N. Diagnosing and managing autism. In: American Psichological Association. 2018. Disponível em: <https://www.apa.org/topics/diagnosing-managing-autism.pdf>. Acesso em: 15 jan. 2022.

STRODL, Stephan RAUBER, Andreas RAUCH, Carl HOFMAN, Hans DEBOLE, Franca AMATO, Giuseppe. (2006). The DELOS Tested for Choosing a Digital Preservation Strategy. 4312. 323-332. 10.1007/11931584_35. Disponível em: <https://www.researchgate.net/publication/220705685_The_DELOS_Testbed_for_Choosing_a_Digital_Preservation_Strategy/link/0fcfd50c5afb116b10000000/download>. Acesso em: 01 nov. 2021.

TAFNER, Malcon A. Redes Neurais Artificiais: Aprendizado e Plasticidade. Revista Cérebro e Mente, Universidade Estadual de Campinas, mar/mai 1998. Disponível em: <https://cerebromente.org.br/n05/tecnologia/plasticidade2.html>. Acesso em: 19 dez. 2021.

TAN, P.-N. et al. Introduction to data mining. [S.l.]: Pearson Education India, 2006.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. Introdução ao datamining: mineração de

dados. [S.l.]: Ciência Moderna, 2009.

TAN, Pang-Ning; STEINBACH, Michael; KARPATNE, Anuj; KUMAR, Vipin. Introduction to Data Mining. 2 ed. New York: Pearson, 2019.

TANG, S.; HANNEGHAN, M.; RHALIBI, A. E. Introduction to games-based learning. Games Based Learning Advancements for Multi-Sensory Human Computer Interfaces. New York: IGI Global, 2009.

TEIXEIRA, Carlos A. A. SILVA, Simone V.. A Mineração de Dados como ferramenta de apoio à tomada de decisão aplicada à área de gerenciamento de projetos de uma indústria. In: XVIII SEGET – Simpósio de Excelência em Gestão e Tecnologia, Faculdades Dom

Bosco. Rezende, RJ. 2020. Disponível em: <https://www.aedb.br/seget/arquivos/artigos20/28830381.pdf>. Acesso em: 01 jan. 2022.

THIBES-UEM, P. A.; ALENCAR-UEM, G. A. R. D.; AZEVEDO-UEM, F. C. D. Estratégia de intervenção: Ensinando emoções. In: V Congresso Brasileiro Multidisciplinar de Educação Especial. Novembro de 2009. Londrina: Paraná. Disponível em: <http://www.uel.br/eventos/congressomultidisciplinar/pages/arquivos/anais/2009/169.pdf>. Acesso em: 15 jan. 2022.

TSAI, F. S. A visualization metric for dimensionality reduction. Expert Systems with Applications. Elsevier Ltd, v. 39, n. 2, p. 1747–1752, fev. 2012. ISSN 09574174. Disponível em: <https://reader.elsevier.com/reader/sd/pii/S0957417411011924?token=A51C4D87F4962F7D586D2D20A434F60A5AE019B66FD503C48725D1660AD2EB27060F3533784283289CD7963A9E96378BoriginRegion=us-east-1originCreation=20211107152252>. Acesso em: 07 nov. 2021.

VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. (2010). Relative clustering validity criteria: A comparative overview. Department of Computer Sciences of the University of São Paulo at São Carlos. May 2010. Disponível em: <https://www.cs.nmsu.edu/ hcao/paper/icredits/2010_RelativeClusteringValidityCriteriaAComparativeOverview.pdf>. Acesso em: 15 nov. 2022.

XU, Rui; WUNSCH, Don. Clustering. John Wiley  Sons, 2008.

ZAKI, Mohammed J.; MEIRA JR, Wagner; MEIRA, Wagner. Data Mining and analysis: fundamental concepts and algorithms. Cambridge University Press, 2014.ZAR, J. H. Biostatistical analysis. 4th ed. New Jersey, Prentice-Hall, Inc., 1999. 663p +212

ZUEGE, Tiago Jasper. Aplicação de técnicas de mineração de dados para detecção de perdas comerciais na distribuição de energia elétrica. Monografia (Bacharelado em Engenharia da Computação) - Faculdade UNIVATES, Universidade do Vale do Taquari, Lajeado, 2018.