

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO RIO GRANDE
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

APRENDIZADO DE MÁQUINA USANDO PCA EM DADOS DE ATIVIDADE CEREBRAL

por

Mateus Schoffen

Dissertação para obtenção do Título de
Mestre em Modelagem Computacional

Rio Grande, Junho de 2022.

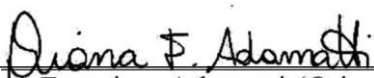
Mateus Schoffen

APRENDIZADO DE MÁQUINA USANDO PCA EM DADOS DE ATIVIDADE CEREBRAL

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal do Rio Grande - FURG, como requisito parcial para obtenção do Grau de Mestre. Área de concentração: Modelagem Computacional.

Aprovada em: 15 de Junho de 2022.

BANCA EXAMINADORA



Profa. Dra. Diana Francisca Adamatti (Orientadora – FURG)



Prof. Dr. Marilton Sanchotène de Aguiar (UFPel)



Profa. Dra. Catia Maria dos Santos Machado (FURG)



Prof. Dr. Cleo Zanella Billa (FURG)

Rio Grande - RS
2022

Ficha Catalográfica

S367a Schoffen, Mateus.

Aprendizagem de máquina usando PCA em dados de atividade cerebral / Mateus Schoffen. – 2022.

37 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-Graduação em Modelagem Computacional, Rio Grande/RS, 2022.

Orientador: Dra. Diana Francisca Adamatti.

1. Neurociências 2. EEG 3. Mineração de dados 4. Atividade cerebral 5. Aprendizado de máquina 6. PCA I. Adamatti, Diana Francisca II. Título.

CDU 004.62:612.8

Catálogo na Fonte: Bibliotecária Valéria Carlosso dos Santos Mazui CRB 10/2704

AGRADECIMENTOS

Em primeiro lugar, sou extremamente grato à minha orientadora, Profa. Dra. Diana Francisca Adamatti, por seus conselhos inestimáveis, apoio contínuo e paciência durante meus estudos no mestrado. Seu conhecimento e experiência me encorajaram todo tempo de minha pesquisa acadêmica e vida diária.

Minha gratidão se estende à Universidade Federal do Rio Grande, pela oportunidade de realizar meus estudos no Centro de Ciências Computacionais e Instituto de Matemática, Estatística e Física. Sou grato ao Programa de Pós-Graduação em Modelagem Computacional pelas experiências extraordinárias que me foram proporcionadas, além de oportunidades de crescimento profissional e pessoal.

Gostaria de agradecer aos meus amigos, colegas de laboratório, colegas de trabalho e equipe de pesquisa do LAMSA pelo tempo precioso que passamos juntos no laboratório e em ambientes sociais. Agradeço também aos meus amigos pelo incentivo e apoio durante todos os meus estudos.

Gostaria de expressar minha gratidão a minha esposa, sem sua tremenda compreensão e incentivo nos últimos anos, seria impossível para mim concluir este estudo. Agradeço aos meus pais, cujo amor e apoio constantes me mantém motivado e confiante. Minha família merece uma gratidão infinita: meu pai por me ensinar bons valores e a nunca desistir, minha mãe por me ensinar a bondade e o amor. À minha família, em especial minha Bisavó de 114 anos, Dona Margarida Blanco, que compartilha seus ensinamentos comigo, lhe dedico, inclusive isso.

Mais importante ainda, sou grato pelo apoio incondicional, inequívoco e amoroso de todos ao meu redor. Minhas realizações e sucessos são porque pessoas bondosas acreditaram em mim. A todos que me alegraram e me inspiraram, sou eternamente grato pelo amor e apoio incondicionais durante todo o processo do mestrado e todos os dias.

RESUMO

Este trabalho tem um cunho interdisciplinar, abordando as áreas da Neurociência e Aprendizado de Máquina e tem como objetivo analisar técnicas de Aprendizado de Máquina aplicadas em dados coletados através de eletroencefalograma em um estudo de caso onde o sujeito está realizando uma atividade de matemática ou está relaxando. Uma pesquisa na literatura demonstra que existem outros estudos envolvendo PCA e técnicas de mineração de dados, porém, nenhuma comparando diversas técnicas de aprendizado de máquina supervisionado e não-supervisionado a fim de analisar a acurácia resultante após o uso da técnica de PCA. Foram utilizadas técnicas de mineração de dados que são capazes de fazer o processamento de grandes quantidades de dados, facilitando a obtenção das informações desejadas. Para o trabalho foi utilizado uma base de dados existente, ao qual realizou-se o pré-processamento e aplicação da técnica de PCA nos dados brutos. Após aplicação de técnicas de aprendizado de máquina, os resultados mostraram que a acurácia se mantém a mesma, tendo pouca variação, mesmo executando as técnicas com 5 ou 30 componentes principais na técnica de PCA, resultando na diminuição da dimensionalidade da base de dados, sem perda das características dos dados.

Palavras-chaves: EEG, mineração de dados, atividade cerebral, aprendizado de máquina, PCA.

ABSTRACT

This work has an interdisciplinary nature, linking the areas of Neuroscience and Machine Learning and aims to analyze Machine Learning techniques applied to data collected through electroencephalogram in a case study where the subject is performing a math activity or is relaxing. The literature shows that there are other studies involving PCA and data mining techniques, however, none comparing different supervised and unsupervised machine learning techniques in order to analyze the resulting accuracy after using the PCA technique. Data mining techniques were used that are capable of processing large amounts of data, making it easier to obtain the desired information. For this work, an existing database was used, to which the pre-processing and application of the PCA technique in the data was executed. After applying machine learning techniques, the results showed that the accuracy remains the same, with little variation, even executing the techniques with 5 or 30 main components in the PCA technique, resulting in a decrease in the dimensionality of the database, without loss of data characteristics.

Keywords: EEG, data mining, brain activity, machine learning, PCA.

SUMÁRIO

1. INTRODUÇÃO	10
1.1 Objetivos	11
1.2 Estrutura do Texto	11
2. EMBASAMENTO TEÓRICO	12
2.1. Sistemas BCI.....	12
2.2. Neurociência	14
2.2.1 Atividade cerebral.....	14
2.3 Mineração de dados	15
2.4 Algoritmos de aprendizado de máquina	16
2.4.1 - Algoritmos de aprendizado de máquina supervisionado.....	17
2.4.2 - Algoritmos de aprendizado não-supervisionado	18
2.5 PCA.....	19
2.6 Trabalhos Relacionados	20
3. METODOLOGIA UTILIZADA	24
4. ANÁLISE E DISCUSSÃO DOS DADOS	29
5. CONCLUSÕES E TRABALHOS FUTUROS	31
REFERÊNCIAS	33

LISTA DE FIGURAS

FIGURA 1: EXEMPLO DE TOUCA UTILIZADA PARA COLETA DOS DADOS.....	13
FIGURA 2: FAIXAS DE FREQUÊNCIA.....	13
FIGURA 3: PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS (KDD)	16
FIGURA 4: FILTROS E RESULTADO DA PESQUISA	20
FIGURA 5: APARELHO DE COLETA DE DADOS EGG MINDWAVE.	24
FIGURA 6: ETAPAS DO DESENVOLVIMENTO DO TRABALHO.....	26
FIGURA 7: COLUNAS DA BASE DE DADOS APÓS APLICAÇÃO DE PCA COM 5 PRINCIPAIS COMPONENTES	27
FIGURA 8: APLICAÇÃO DA TÉCNICA DE APRENDIZADO DE MÁQUINA (KNN) E ACURÁCIA RESULTANTE.	28

LISTA DE TABELAS

TABELA 1: ACURÁCIA DAS TÉCNICAS DE APRENDIZADO DE MÁQUINA.....	29
--	----

LISTA DE ABREVIATURAS

3D - Tridimensional.

BCI - Brain Computer Interface.

CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

EEG - Eletroencefalograma.

fMRI - Functional magnetic resonance imaging - ressonância magnética.

FURG - Universidade Federal do Rio Grande.

Hz - Hertz.

ICA - Independent Component Analyses.

IEEE - Institute of Electrical and Electronics Engineers.

KDD - Knowledge Discovery in Database.

KNN - K-Nearest Neighbors.

LDA - Linear Discriminant Analysis.

MD - Mineração de Dados.

MIDS - Mestrado em Ciência da Informação e Dados.

MLP - Multi-Layer Perceptron.

PCA - Principal Component Analysis.

PPGMC - Programa de Pós-Graduação em Modelagem Computacional.

RAPM - Raven's Advance Progressive Metric

RNA - Rede Neural Artificial.

SciELO - Scientific Electronic Library Online.

SLT - Statistical Learning Theory.

SpPCA - PCA Baseado em Subpadrão.

SubXPCA - PCA Baseado em Correlação de Subpadrão Cruzado.

SVM - Support vector machines.

1. INTRODUÇÃO

Os algoritmos de aprendizado de máquina são construídos de forma a aprender e fazer previsões a partir de dados. Ao contrário dos algoritmos de programação estática, que precisam de instrução humana explícita, os algoritmos de aprendizado de máquina realizam uma análise automatizada dos dados e realizam previsões a partir de amostras dos dados (CHOUDHARY, 2017).

O campo de aprendizado de máquina atingiu um desenvolvimento impressionante recentemente, com a ajuda do rápido aumento na capacidade de armazenamento de dados e o poder de processamento dos computadores (BAŞTANLAR, 2014). Junto com muitas outras disciplinas, os métodos de aprendizado de máquina têm sido amplamente empregados em neurociências.

Grandes conjuntos de dados estão cada vez mais difundidos nas mais diversas áreas. Para interpretar tais conjuntos de dados, são necessários métodos para reduzir sua dimensionalidade de forma interpretável, de modo que a maioria das informações nos dados seja preservada. A análise de componentes principais (PCA) é uma das técnicas mais antigas e mais utilizadas para reduzir a dimensionalidade de tais conjuntos de dados, aumentando a interpretabilidade e ao mesmo tempo minimizando a perda de informação característica dos dados (BRO, 2014).

A obtenção de dados de sinais cerebrais através de Eletroencefalograma (EEG) pode gerar uma quantidade enorme de dados para a realização de mineração de dados. Para melhorar a precisão e classificação da atividade de um indivíduo baseado em EEG pode-se utilizar técnicas de aprendizado de máquina (WANG, 2014).

Nos dias atuais, ainda são necessários estudos na área de mineração de dados que auxiliem na classificação dos dados de EEG para, eventualmente, realizar uma previsão satisfatória sobre o estado do indivíduo. Este trabalho une PCA e algoritmos de aprendizado de máquina para examinar a eficácia de diferentes técnicas supervisionadas e não supervisionadas a fim de contribuir para futuros trabalhos nestas áreas. O estudo de caso utilizado neste trabalho busca definir um estado em que o sujeito esteja realizando uma atividade matemática ou relaxando.

Assim, o trabalho busca entender se com a aplicação da técnica de redução da dimensionalidade da base, que no caso deste trabalho foi o PCA, traz benefícios em relação a acurácia dos algoritmos de aprendizado de máquina supervisionados e não supervisionados utilizados para análise dos dados de EEG do estudo de caso analisado neste trabalho.

1.1 Objetivos

O trabalho tem como objetivo geral analisar técnicas de Aprendizado de Máquina aplicadas em dados coletados através de eletroencefalograma, em um estudo de caso onde o sujeito está realizando uma atividade de matemática ou está relaxando. Como a base utilizada possui muitos atributos, de forma a pré-processar os dados selecionados, foi utilizada a técnica de PCA (*Principal Analysis Components*) para reduzir a dimensionalidade da mesma.

Desta maneira, tem-se como objetivos específicos:

- Análise da base utilizada como estudo de caso e tratamento dos dados para a utilização da técnica de PCA.
- Comparação entre diversas técnicas de aprendizado de máquina não-supervisionada e supervisionada, utilizando a base bruta e a base tratada com PCA, de forma a verificar seu desempenho em relação a métrica de acurácia.

1.2 Estrutura do Texto

O presente trabalho está organizado em 5 capítulos, dispostos da seguinte maneira: Capítulo 1 Introdução, onde o mesmo encontra-se dividido em duas seções, Objetivos e Estrutura do texto. No qual abordou a justificativa do projeto, o objetivo geral e objetivos específicos que foram apresentados anteriormente neste capítulo e a estrutura do texto.

O Capítulo 2, Embasamento Teórico, foi construído em cinco seções: A primeira delas sobre sistemas BCI, em que são apresentados os conceitos, tipos e técnicas utilizadas. A segunda Seção, Neurociência, encontra-se dividida em duas subseções, atividade cerebral e estado de relaxamento, nas quais contemplam definições e estudos que delimitam a presente pesquisa.

Para a terceira Seção, a Mineração de dados, no qual são evidenciados os conceitos e etapas dos processos. Na Seção 4 Algoritmos de aprendizado de máquina, descreve-se as técnicas que foram utilizadas neste estudo. Na última Seção, denominada Trabalhos Relacionados, busca-se relatar os trabalhos que vêm sendo realizados sobre a mesma temática.

No Capítulo 3, denominado Metodologia Utilizada, estão os procedimentos de pré-processamento de dados e processos de análise que fizeram parte da pesquisa, em que se apontam as etapas realizadas e os algoritmos utilizados para o desenvolvimento do trabalho.

No Capítulo 4 são apresentados os resultados encontrados na pesquisa e as discussões alicerçadas no embasamento teórico. Já o Capítulo 5 apresenta as conclusões e as reflexões sobre os resultados obtidos, bem como os trabalhos futuros propostos.

2. EMBASAMENTO TEÓRICO

Este capítulo apresenta a base teórica para o desenvolvimento do trabalho. Serão apresentadas algumas definições e conceitos que compõem a pesquisa, sendo a primeira seção sobre os sistemas BCI (*Brain Computer Interface*). Seguido da seção neurociência, que conta com duas subseções: as atividades cerebrais e o estado de relaxamento. Na terceira seção estão detalhados os conceitos e processos de mineração de dados. A seção seguinte aborda os algoritmos de aprendizado de máquina utilizados no trabalho. Na quinta seção é explicado o método de pré-processamento de dados *Principal Component Analysis* (PCA). Finalizando o capítulo, tem-se a seção sobre os trabalhos relacionados, na qual apresenta estudos que abordam uma temática parecida com este trabalho.

2.1. Sistemas BCI

Interfaces cérebro-computador (BCI - *Brain Computer Interface*) são sistemas de comunicação e controle que são usados para traduzir sinais cerebrais em comandos e mensagens, capazes de controlar aplicações como movimentar próteses, digitar letras usando um teclado virtual e ligar ou desligar luzes. “Os sistemas BCI desenvolvidos atualmente são ferramentas que podem ajudar os usuários a se comunicar e realizar atividades cotidianas, embora apresentem um sucesso limitado e ainda se encontrem principalmente em ambientes de pesquisa” (SILVEIRA, 2013, p. 15).

Como em qualquer sistema de comunicação, os sistemas BCI consistem em uma parte de entrada, que é o sinal cerebral usando eletrodos, seja invasivo ou não invasivo, e uma parte de saída, que é o comando para se comunicar com os dispositivos externos. Para WOLPAW *et al.* (2000), o BCI é um "sistema de comunicação que não depende das vias normais de saída do cérebro de nervos periféricos e músculos".

Nesse sentido, um sistema BCI fornece uma rota alternativa para o envio de comandos para computadores ou outros dispositivos. O Eletroencefalograma (EEG), por exemplo, é uma das técnicas mais utilizadas para entrada de dados nos sistemas BCI, consistindo em um sistema que permite o acesso à atividade cerebral por um "registro de atividade elétrica ao longo do couro cabeludo, produzido pelo anel de neurônios dentro do cérebro" (NIEDERMEYER, SILVA, 2005). Para que ocorra a captura desses sinais EEG, pode utilizar-se de uma touca na cabeça do indivíduo (Figura 1), onde são acoplados os eletrodos, que facilmente, fazem o contato e enviam os estímulos cerebrais para a máquina, computador ou prótese.

FIGURA 1: EXEMPLO DE TOUCA UTILIZADA PARA COLETA DOS DADOS.



FONTE: SILVA (2020).

As ondas cerebrais, captadas a partir dos sinais de EEG, podem ser classificadas por sua frequência, amplitude, forma e posição no couro cabeludo. As aplicações EEG se concentram em uma faixa estreita, a partir de 0,1 Hz a 100 Hz. Esses sinais de EEG são classificados com base em sua faixa de frequência e estão relacionados entre as ondas: Alfa, Beta, Theta, Gama e Delta (MUNIAN, 2014), conforme a Figura 2.

FIGURA 2: FAIXAS DE FREQUÊNCIA.



FONTE: O AUTOR (2020).

2.2. Neurociência

A neurociência é uma ciência interdisciplinar que surgiu na Grécia Antiga e teve um avanço significativo no século XIX devido às novas tecnologias. Desde a sua origem, ela colabora com outros campos como a educação, química, antropologia, linguística, engenharia, matemática, medicina e ciência da computação. Segundo LENT (2008), a neurociência é o estudo científico que abrange o funcionamento, desenvolvimento e estrutura do sistema nervoso. A mesma pode ser dividida em cinco grandes grupos: a neurociência molecular, celular, sistêmica, cognitiva e comportamental.

O presente trabalho se debruça na perspectiva da neurociência comportamental, pois esta área estuda as estruturas que estão relacionadas ao comportamento ou a fenômenos como sono, depressão, ansiedade, entre outros comportamentos. Esta linha procura estabelecer uma conexão entre o contato do organismo e seus fatores internos (pensamentos e emoções) ao comportamento visível, como a forma de falar de se postar e, até mesmo, os gestos (LENT, 2008).

2.2.1 Atividade cerebral

O estudo do cérebro pode ser realizado de diversas formas, uma das formas de mensurar a atividade cerebral é através do EEG, ao qual captura a atividade cerebral através de frequências e avalia a atividade elétrica espontânea do cérebro. É um método não invasivo. O cérebro é o principal órgão e o centro do sistema nervoso, sendo o órgão que mais demanda energia do nosso corpo. Através da atividade dos neurônios podemos ver o funcionamento de suas partes.

A comunicação entre os neurônios ocorre por meio de pulsos elétricos propagados pelas sinapses. Por conta disso, entre as primeiras tecnologias criadas para medir a atividade cerebral, estão as que se baseiam na detecção de mudanças no potencial elétrico fazendo uma leitura direta nos neurônios por meio de tecnologias invasivas. Para casos que envolvam grandes populações de neurônios, utilizam-se por técnicas não invasivas, como a eletroencefalografia (RAO, 2013).

Na eletroencefalografia, o monitoramento cerebral é amplamente utilizado em técnicas BCI, já o EEG mede a atividade elétrica neuronal em termos de potenciais do couro cabeludo. Vale a pena ressaltar, porém, que o potencial elétrico gerado por um neurônio individual é muito pequeno para ser captado por EEG; os potenciais registrados do couro cabeludo, portanto, sempre refletem o somatório da atividade síncrona de milhares ou milhões de neurônios que têm orientação espacial semelhante.

Com EEG não é possível adquirir informações em 3D da ativação neuronal, em oposição a outros métodos (por exemplo fMRI - *functional magnetic resonance imaging* - ressonância magnética). A resolução espacial do EEG também não é a melhor entre todas. No entanto, a resolução temporal é boa e a leitura de EEG dos circuitos são relativamente baratos, quando comparados a outros métodos de monitoramento de atividades cerebrais. Esse monitoramento pode nos informar algumas características do estado do indivíduo, sendo muito importante para esse trabalho classificar diversas atividades. Neste trabalho, foca-se no estado de relaxamento, para saber se o mesmo está realizando atividade de matemática ou em estado de relaxamento.

2.3 Mineração de dados

Segundo TAN *et al.* (2006), a Mineração de Dados (MD) é uma parte integral da descoberta de conhecimento em banco de dados (KDD - *knowledge Discovery in Database*). O processo de KDD, conforme apresentado em FAYYAD *et al.* (1996), é o processo de usar Métodos de Mineração de Dados para extrair conhecimento, usando um banco de dados junto com qualquer pré-processamento, sub-amostragem e transformação necessários da base de dados. Cinco etapas são consideradas:

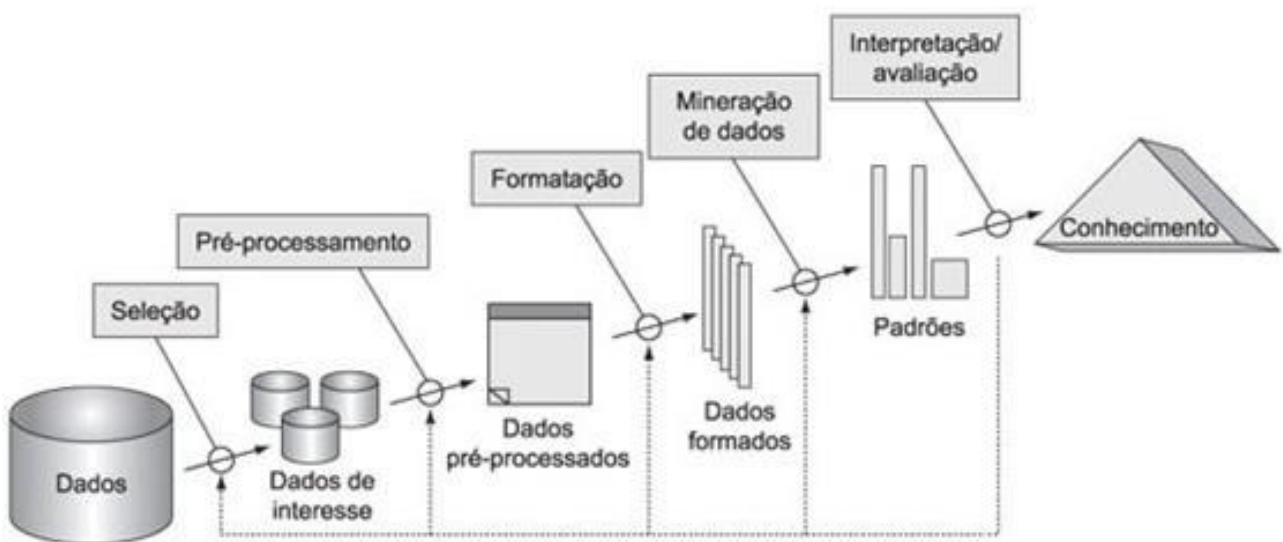
- Seleção: este estágio consiste em criar um conjunto de dados de destino ou focar em um subconjunto de variáveis ou amostras de dados;
- Pré-processamento: esta etapa consiste na limpeza e pré-processamento de dados alvo para dados consistentes;
- Transformação: esta etapa consiste em transformar os dados usando métodos de redução ou transformação de dimensionalidade;
- Mineração de Dados: esta etapa consiste na busca de padrões de interesse em uma forma de representação particular, dependendo do objetivo de mineração de dados (então geral, previsão);
- Interpretação/Avaliação: esta etapa consiste na interpretação e avaliação de padrões adquiridos.

Os modelos podem ser de dois tipos: preditivos ou descritivos. Os preditivos pretendem estimar valores futuros ou desconhecidos de variáveis de interesse, conhecidas como variáveis de destino ou dependentes, usando outros campos no banco de dados, conhecidas como variáveis independentes ou preditivas.

No entanto, os modelos descritivos identificam padrões que agrupam os dados, servindo para explorar propriedades dos dados analisados, não para prever novos dados. Dentro dos modelos preditivos, encontramos classificação e regressão, enquanto regras de associação, clustering, associações e correlações sequenciais são tarefas descritivas.

A mineração de dados gera conhecimento útil por meio da análise de dados e extração de padrões de interesse. Esse tipo de conhecimento é obtido por meio de modelos (Figura 3) que permitem encontrar relacionamentos anteriormente desconhecidos, padrões ou regras inferidas (HERNÁNDEZ *et al.*, 2004).

FIGURA 3: PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS (KDD)



FONTE: FAYYAD (1996).

2.4 Algoritmos de aprendizado de máquina

Na etapa de mineração de dados (conforme Figura 3) são utilizados algoritmos de aprendizado de máquina para obter os padrões de conhecimento. Neste trabalho são utilizados alguns algoritmos de aprendizado de máquina supervisionado e não supervisionados, sendo os supervisionados aqueles que já possuem um “rótulo” *a priori*, ou seja, um especialista da área que os dados estão relacionados definiu uma classe ou grupo para as instâncias pertencentes aos dados. E os não supervisionados, que não possuem um “rótulo”, ou seja, não há conhecimento ou definição de classe ou grupo *a priori*. A seguir são explicados os algoritmos utilizados nos testes realizados, mas a quantidade de algoritmos disponíveis na literatura é muito ampla.

2.4.1 - Algoritmos de aprendizado de máquina supervisionado

As técnicas de aprendizado de máquina supervisionado apresentadas aqui são: KNN, J48, Random Forest, SVM, Naïve Bayes e regressão logística.

Segundo ALI *et al.* (2019), o algoritmo de aprendizagem supervisionada K-Nearest Neighbors (KNN) é muito usado no campo de aprendizado de máquina para classificação de dados. Ele analisa os vizinhos mais próximos para poder classificar eles em diferentes grupos, através da distância (Euclidiana, Manhattan, Minkowski ou Ponderada).

Segundo PANDEY (2017), a técnica de KNN é amplamente utilizada em mineração de dados e aprendizado de máquina porque é simples e muito útil em diversas aplicações. A classificação é usada para prever os rótulos de pontos de dados de teste após o treinamento de dados de amostra. Nas últimas décadas, os pesquisadores propuseram muitos métodos de classificação, mas ainda assim, KNN é um dos métodos mais populares para classificação de um conjunto de dados.

Decision Tree Model (J48) é um algoritmo de classificação que surgiu da necessidade de recodificar o algoritmo C4.5, que, originalmente, é escrito na linguagem C, para a linguagem Java (WITTEN *et al.*, 2005). O algoritmo tem como objetivo gerar uma árvore de decisão baseada em um conjunto de dados de treinamento, sendo este modelo usado para classificar as instâncias no conjunto de testes (LIBRELOTTO *et al.*, 2014).

O algoritmo de Decision Tree J48 é um algoritmo que usa o método de dividir e conquistar para aumentar a capacidade preditiva das árvores de decisão. Com isso, é utilizado a melhor etapa avaliada localmente, sem se preocupar se esta etapa produziu a melhor solução, cria-se um problema, e o divide em vários subproblemas, criando sub-árvores entre a raiz e as folhas (BASTOS, 2016).

Random Forest é uma evolução do algoritmo de árvore de decisão, onde uma “floresta” é criada e depois são escolhidas as melhores partes das árvores. É um algoritmo de aprendizado supervisionado que pode ser usado tanto para classificação quanto para regressão (BREIMAN, 2021). Random Forest tem uma variedade de aplicações, como mecanismos de recomendação, classificação de imagens e seleção de recursos. Ele pode ser usado para classificar candidatos a empréstimos legais, identificar atividades fraudulentas e prever doenças (ALI, 2012).

Em sua formulação clássica, o SVM (*Support Vector Machines*) é um método de aprendizado supervisionado, pois é baseado em (entrada, saída) exemplos. O SVM deriva da Teoria da Aprendizagem Estatística (SLT) e é particularmente útil quando o processo no qual as entradas são mapeadas nas saídas não é conhecido.

O SVM é um método de aprendizado amplamente utilizado em problemas de reconhecimento de padrões e regressão. As aplicações do SVM pertencem a diferentes domínios, desde a biologia computacional (BEN-HUR *et al.*, 2008) à previsão de séries financeiras (VAN GESTEL *et al.*, 2001). No contexto de confiabilidade, SVM tem sido usado, por exemplo, em CBM e detecção de falhas (WIDODO e YANG, 2007), para classificar anomalias em componentes (ZIO *et al.*, 2007), para prever a confiabilidade dos equipamentos (PAI *et al.*, 2006), entre outros.

O algoritmo Naïve Bayes é um algoritmo de aprendizado supervisionado, baseado no teorema de Bayes utilizado para resolver problemas de classificação. É usado principalmente na classificação de texto que inclui um conjunto de dados de treinamento de alta dimensão. É um dos algoritmos de classificação mais simples e eficazes que ajuda na construção de modelos de aprendizado de máquina rápidos que podem fazer previsões rápidas, sendo um classificador probabilístico, o que significa que prevê com base na probabilidade de um objeto (RISH, 2001). Alguns exemplos populares do Algoritmo Naïve Bayes são filtragem de spam, análise sentimental e classificação de artigos.

A regressão logística é uma técnica para ajustar uma superfície de regressão aos dados quando a variável dependente é dicotômica. Trata-se de fato conhecer os fatores associados a um fenômeno desenvolvendo um modelo de previsão. A popularidade deste método é bem conhecida nas ciências da saúde e na ciência humana, onde a variável a ser prevista é a presença ou a ausência de doença. Por exemplo, pode ser um estudo sobre depressão maior, onde queremos saber os fatores que melhor preveem isso, estudando variáveis como idade, sexo, autoestima ou relacionamentos interpessoais (TABACHNICK *et al.*, 2007).

2.4.2 - Algoritmos de aprendizado não-supervisionado

As técnicas de aprendizado de máquina não- supervisionado apresentadas aqui são: K-Means e MLP (*Multi-Layer Perceptron*).

Segundo GHOSH (2013), K-Means é basicamente um método de particionamento aplicado para analisar dados e tratar observações dos dados como objetos baseados em localização e distância entre vários pontos de dados de entrada. A lógica do algoritmo é dividir os objetos em clusters mutuamente exclusivos (K), de tal forma que os objetos dentro de cada cluster permaneçam tão próximos possíveis entre si, mas o mais longe possível de objetos em outros aglomerados.

O algoritmo K-Means é geralmente o método de agrupamento mais conhecido e utilizado. Existem várias extensões de K-Means na literatura, e diversas aplicações do algoritmo nos mais variados casos. Embora seja um algoritmo de aprendizado não supervisionado, ele realiza um

agrupamento para reconhecimento de padrões e aprendizado de máquina, o algoritmo K-Means e suas extensões são sempre influenciados por inicializações com um número necessário de clusters (SINAGA, 2020).

MLP (*Multi-Layer Perceptron*) é um tipo de rede neural artificial (RNA) totalmente conectado, consistindo de neurônios organizados em camadas. Pelo menos três camadas compõem o MLP: uma camada de entrada, uma camada de saída e uma ou mais camadas ocultas. O MLP é um método de fácil implementação fornecendo modelos de alta qualidade, enquanto mantém o tempo de treinamento relativamente baixo em comparação com mais métodos complexos (CAR *et al.* 2020).

O Algoritmo MLP é baseado no cálculo dos valores dos neurônios em uma camada atual como a soma ativada de saídas ponderadas de neurônios em uma camada anterior, conectada ao neurônio (BISHOP e NASRABADI, 2016). MLP é muito popular em diversos campos, como reconhecimento de fala, reconhecimento de imagem e softwares de tradução automática (SAVALIA, 2018).

2.5 PCA

A análise de componentes principais (PCA) é um algoritmo para transformar as colunas de um conjunto de dados em um novo conjunto de recursos chamados de Componentes Principais. Ao fazer isso, uma grande parte das informações em todo o conjunto de dados é efetivamente compactada em menos colunas de recursos. Isso permite redução de dimensionalidade e capacidade de visualizar a separação de classes ou clusters, se houver.

Os componentes principais apresentam propriedades importantes: cada componente principal é uma combinação linear de todas as variáveis originais, são independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados (JOHNSON, 2014; HONGYU, 2016).

A análise de componentes principais é um método multivariado de análise em estatística para descrever uma grande quantidade de dados com um número menor de quantidades relevantes, os componentes principais (JOLLIFFE *et al.*, 2016).

Fala-se em redução de dados, embora a rigor os dados não sejam reduzidos, mas apenas sua descrição. Como componentes principais, o método calcula os auto vetores da matriz de covariância dos dados e escolhe o mais importante deles. Esses auto vetores são os eixos principais do elipsoide descrito pela matriz de covariância, que representa mais ou menos a "nuvem de pontos" dos dados.

Neste estudo, os auto vetores foram utilizados nas aplicações das técnicas e o PCA aplicado foi de importação da biblioteca `sklearn.decomposition import PCA`.

2.6 Trabalhos Relacionados

Esta seção apresenta artigos que mais se aproximam da pesquisa realizada neste trabalho. A busca por trabalhos relacionados foi realizada no Portal de Periódicos da CAPES, com filtros aplicados para buscar artigos publicados nos últimos 5 anos e com as seguintes palavras-chave: `principal component analysis`, `machine learning` e `electroencephalogram`, para trabalhos escritos na língua inglesa, resultando em 904 artigos, conforme Figura 5.

FIGURA 4: FILTROS E RESULTADO DA PESQUISA

The screenshot displays the search interface of the CAPES Portal de Periódicos. It features a search filter section on the left with three criteria: 'Qualquer campo contém principal component analysis', 'E Qualquer campo contém electroencephalogram', and 'E Qualquer campo contém machine learning'. To the right, there are dropdown menus for 'Tipo de material' (set to 'Todos os itens'), 'Idioma' (set to 'Qualquer idioma'), and 'Data de publicação' (set to 'Últimos 5 anos'). Below the filters are buttons for '+ ADICIONAR OUTRO CAMPO' and 'LIMPAR'. At the bottom of the filter section is a green 'BUSCAR' button. Below the search bar, a summary bar shows '0 selecionado(s)' and 'PÁGINA 1 904 Resultados'. A link 'Personalizar meus resultados' is also visible.

FONTE: O AUTOR (2020).

Após verificar que vários trabalhos contemplavam apenas um dos temas buscados, e depois da leitura dos títulos e resumo dos artigos, os trabalhos considerados mais relevantes para essa pesquisa são apresentados a seguir.

O principal trabalho encontrado, que também aborda os temas desta pesquisa, é o artigo de GUERRERO *et al.* (2021), que apresenta um estudo onde foi aplicado PCA em dados de eletroencefalograma de indivíduos com epilepsia. O trabalho de GUERRERO *et al* (2021) propôs um

estudo, baseado na análise de Fourier, através da transformação rápida de Fourier e análise de componentes, para identificar quantitativamente padrões, de forma a diagnosticar e diferenciar entre pacientes e portadores da doença (epilepsia). Subsequentemente, a análise de componentes principais foi utilizada para classificar os pacientes, empregando as bandas de frequência como características do sinal. Ao final, foi feita uma comparação antes e depois da utilização da análise de componentes principais, onde demonstra que a precisão não aumenta após o uso de PCA nos dados de EEG.

Já o trabalho de JAISWAL *et al.* (2017) direciona seus esforços em obter a localização da região do foco epiléptico durante crises epiléticas, aplicando diferentes técnicas de processamento de sinal e aprendizado de máquina em conjunto em registros intracranianos de EEG propondo duas abordagens eficazes, envolvendo PCA baseado em subpadrão (SpPCA) e PCA baseado em correlação de subpadrão cruzado (SubXPCA), utilizando SVM para detecção automática de convulsões em sinais de EEG. A extração de recursos foi realizada usando SpPCA e SubXPCA. Ambas as técnicas exploram a correlação de sub padrões dos sinais de EEG, o que auxiliou no processo de tomada de decisão. O SVM foi usado para classificação de sinais de EEG de convulsão e não convulsão e treinado com kernel de base radial. Todos os experimentos foram realizados no conjunto de dados de EEG de epilepsia com referência, o conjunto de dados consiste em 500 sinais de EEG registrados em diferentes cenários. Sete diferentes casos experimentais para classificação foram conduzidos, sendo a precisão da classificação avaliada usando validação cruzada de dez vezes. Os resultados da classificação das abordagens propostas foram comparados com os resultados de algumas das técnicas existentes propostas na literatura para estabelecer a afirmação.

O trabalho de FARSI *et al.* (2021) visou introduzir um método baseado em deep learning que pode identificar automaticamente sinais de EEG de pessoas diagnosticadas como alcoólicos. Ele também explorou se vale a pena aplicar um método de extração de recursos artesanais a técnicas de aprendizado profundo para classificação de alcoolismo. Para investigar isso, o artigo apresentou dois algoritmos baseados em aprendizado profundo para classificação de sinais de EEG de alcoólicos para comparação. No Algoritmo 1, a técnica de extração de características baseada em Análise de Componentes Principais (PCA) foi aplicada para extrair componentes representativos e, em seguida, as características extraídas são usadas como entrada para a rede neural artificial (RNA) para classificação. No Algoritmo 2, os dados brutos de EEG são usados diretamente como entradas para um método de aprendizado profundo: 'memória longa e curta ' para detecção de alcoolismo. Os resultados experimentais mostraram que o Algoritmo 2 proposto pode alcançar uma precisão de classificação média de 93%, enquanto esta precisão é de 86% para o Algoritmo 1 proposto. As avaliações comparativas com os algoritmos de última geração indicam que o Algoritmo 2 também supera outros algoritmos concorrentes na literatura.

A pesquisa de AMIN *et al.* (2017) propôs uma abordagem de reconhecimento de padrões que discrimina os sinais de EEG registrados durante diferentes condições cognitivas. Extração de características baseadas em onduleta ou wavelets, tais como, decomposição multi-resolução em coeficientes detalhados e aproximados, bem como a energia relativa de wavelets foram computadas. Um conjunto de dados de EEG de alta densidade validou o método proposto identificando duas classificações: (1) sinais de EEG registrados durante tarefas cognitivas complexas, usando o teste *Raven's Advance Progressive Metric* (RAPM); (2) Sinais de EEG registrados durante uma tarefa de linha de base (olhos abertos). Técnicas como *K-nearest neighbors* (KNN), *Support Vector Machine* (SVM), *Multi-layer Perceptron* (MLP) e Naïve Bayes foram empregadas. Os resultados obtiveram 99,11% de precisão via classificador SVM para aproximações de coeficientes (A5) de baixas frequências variando de 0 a 3,90 Hz. As taxas de precisão para coeficientes detalhados foram 98,57% e 98,39% para SVM e KNN, respectivamente; e para coeficientes detalhados (D5) derivados da faixa de sub-banda (3,90-7,81 Hz). As taxas de precisão para classificadores MLP e NB foram comparáveis em 97,11-89,63% e 91,60-81,07% para os coeficientes A5 e D5, respectivamente. Além disso, a abordagem proposta também foi aplicada no conjunto de dados públicos para classificação de duas tarefas cognitivas e obteve resultados de classificação comparáveis, ou seja, 93,33% de precisão com KNN. O esquema proposto rendeu desempenhos de classificação significativamente mais altos usando classificadores de aprendizado de máquina em comparação com a extração quantitativa de recursos existente. Esses resultados sugerem que o método de extração de características proposto classifica de forma confiável os sinais de EEG registrados durante tarefas cognitivas com um maior grau de precisão.

O trabalho de DOMA *et al.* (2020) envolveu a análise dos dados de época dos canais do sensor EEG e a realização de análises comparativas de várias técnicas de aprendizado de máquina (SVM, K-nearest neighbor, Linear Discriminant Analysis, Logistic Regression e Decision Trees) foram testados com e sem análise de componentes principais para redução de dimensionalidade. O DEAP Dataset foi usado neste estudo, que é um conjunto de dados multimodal para a análise dos estados afetivos humanos públicos. As previsões foram baseadas nos rótulos dados pelos participantes para cada um dos 40 trechos de 1 minuto de música. Os participantes classificaram cada vídeo em termos de nível de excitação, valência, gostar/não gostar, dominância e familiaridade. Os classificadores de classes binárias foram treinados no tempo segmentado, intervalos de 15s de dados de época, individualmente para cada uma das 4 classes. PCA com SVM teve o melhor desempenho e produziu um F1-score de 84,73%, com 98,01% de recall no intervalo 30 a 45 da segmentação. Os resultados comprovaram que diferentes modelos de classificação devem ser usados para identificar diferentes estados emocionais.

Esta seção apresentou os trabalhos que mais se aproximaram do tema proposto, focando no uso de aprendizado de máquina para sinais cerebrais e a utilização da técnica de PCA. Apesar de os trabalhos utilizando sinais de EEG terem explorados diversos ambientes, não foram encontrados trabalhos abordando PCA e aprendizado de máquina para classificar indivíduos quanto ao estado de relaxamento, que é o estudo de caso proposto neste trabalho.

3. METODOLOGIA UTILIZADA

Este capítulo explica a caminhada percorrida até a obtenção dos resultados do estudo, iniciando pela escolha da base de dados para o estudo de caso, o pré-processamento realizado nos dados, a aplicação do PCA, a execução dos algoritmos de aprendizado de máquina, a obtenção dos resultados e como foi realizada a análise sobre os resultados obtidos.

Devido a pandemia e algumas alterações na pesquisa, optou-se por utilizar uma base de dados existente para a execução dos algoritmos e o desenvolvimento do trabalho, esta base foi retirada do site KAGGLE.

A base de dados utilizada neste trabalho foi coletada pelo programa de Mestrado em Ciência da Informação e Dados (MIDS) da Escola de Informação da Universidade da Califórnia em Berkeley, que compartilhou publicamente o conjunto de dados coletados, usando o aparelho de BCI MindWave (Figura 6), juntamente com o código do software e o estímulo visual usado para coletar os dados.

O conjunto de dados inclui todas as leituras dos sujeitos durante a apresentação do estímulo, bem como leituras antes do início e após o término do estímulo, tudo estando disponível de forma gratuita através do link: <https://www.kaggle.com/datasets/berkeley-biosense/synchronized-brainwave-dataset>.

FIGURA 5: APARELHO DE COLETA DE DADOS EGG MINDWAVE.



FONTE: [HTTP://SUPPORT.NEUROSKY.COM/KB/MINDWAVE/MINDWAVE-DIAGRAM](http://support.neurosky.com/kb/mindwave/mindwave-diagram).

Durante a coleta foram apresentados dois estímulos ligeiramente diferentes para dois grupos diferentes. Os estímulos 1 estão disponíveis através do YouTube pelo link: <https://www.youtube.com/watch?v=zkGoPdpRvaU>; e os estímulos 2 estão disponíveis em <https://www.youtube.com/watch?v=sxqlOoBBjvc>.

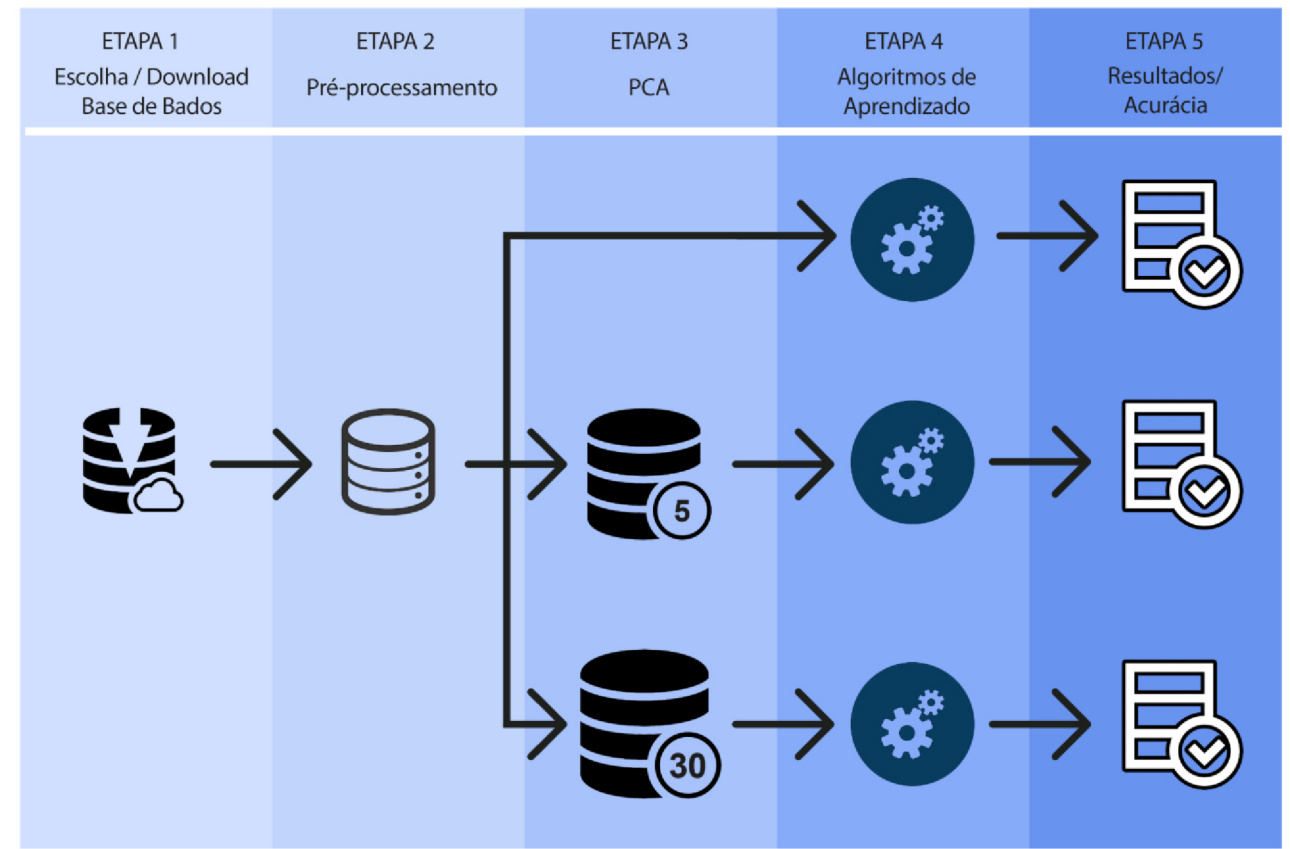
Para ambos os estímulos, um grupo de cerca de 15 pessoas visualizaram estes estímulos ao mesmo tempo, enquanto os dados de EEG estavam sendo coletados. Os estímulos que cada pessoa visualizou estão disponíveis na base de dados como `subject-metadata.csv`. Além disso, os tempos sincronizados para ambos os estímulos foram salvos no arquivo `stimulus-timing.csv`.

Para cada participante, também foi coletado anonimamente alguns outros metadados (via um questionário), como se eles já viram ou não o vídeo exibido durante o estímulo (um anúncio do superbowl), sexo, se viram ou não ícones ocultos exibidos durante o exercício de contagem de cores e a cor escolhida durante o exercício de contagem de cores. Tudo isso pode ser encontrado em `subject-metadata.csv`. Também foi coletado o tempo de visualização (em `indra_time`) de todos os eventos de estímulo, tanto para a sessão 1 quanto para a sessão 2. Esses tempos estão incluídos em `stimulus-times.csv`.

A base de dados contém 105.73 MB de dados separados em 13 colunas e possui 30.013 instâncias, para o estudo foram utilizados dos dados da coluna `raw_data`, que contém os dados dos eletrodos (um vetor de 512 valores) e da coluna que contém o rótulo (label) da atividade, com valores `relax` ou `math`, que seriam o estado de relaxamento e realizando atividade matemática, respectivamente. Para este trabalho, a base original é tratada com o nome de base sem tratamento.

Para facilitar o entendimento da metodologia utilizada neste trabalho, as atividades foram divididas em 5 etapas como demonstra a Figura 7.

FIGURA 6: ETAPAS DO DESENVOLVIMENTO DO TRABALHO.



FONTE: O AUTOR (2022).

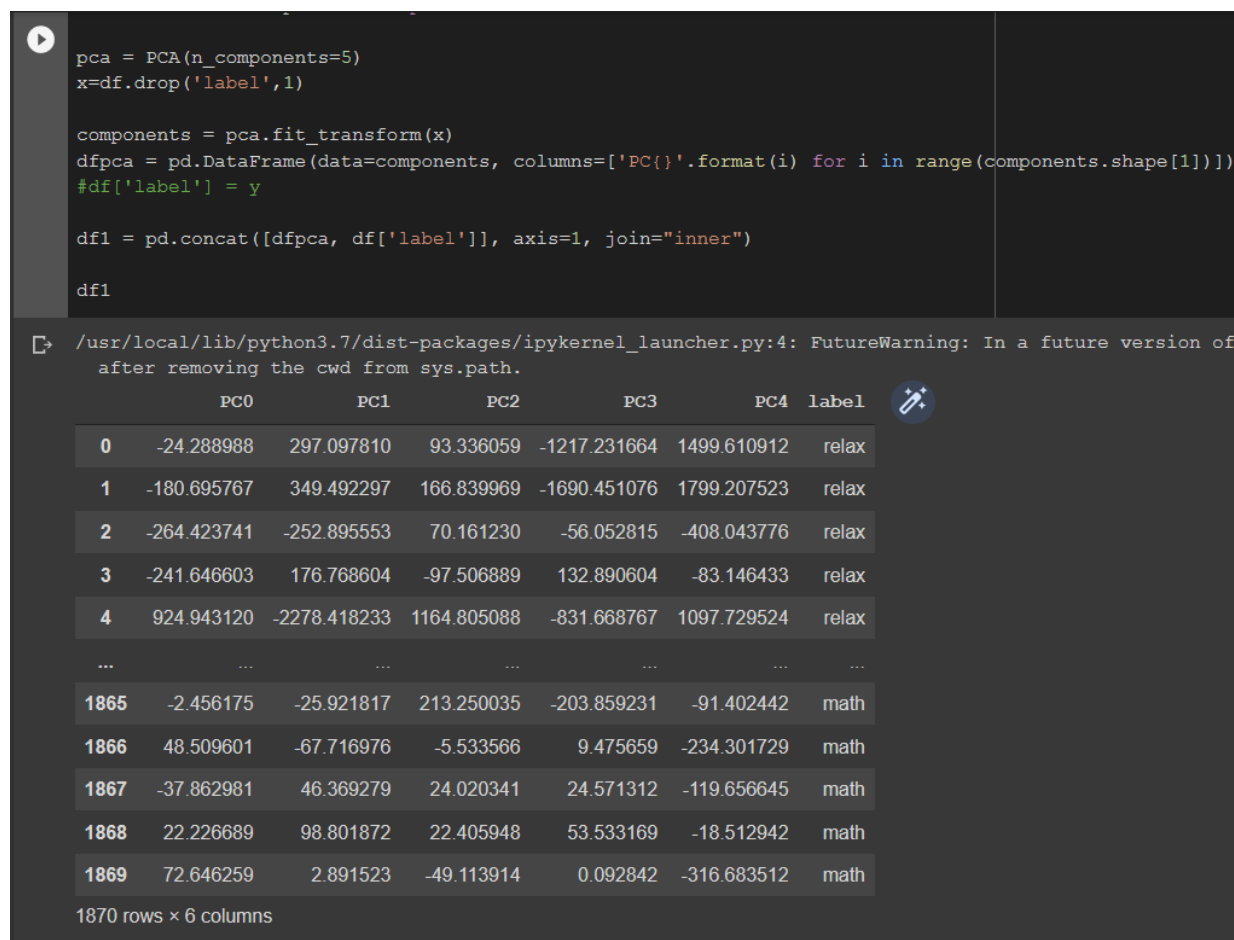
Na etapa 1 busca-se a base de dados para realizar os estudos. Após pesquisa em sites de base de dados de livre acesso, foi escolhida a base de dados disponível publicamente no site *Kaggle*, pois possui informações estruturadas de como foram realizadas as coletas e exemplos já aplicados à base.

Na etapa 2 foi realizado o pré-processamento dos dados, onde o primeiro passo foi a organização da base de dados removendo as colunas `id`, `indra_time`, `browser_latency`, `reading_time`, `attention_esence`, `meditation_esence`, `eeg_power`, `signal_quality`. Deixando apenas os dados brutos da atividade cerebral e a coluna onde indica qual estado o indivíduo se encontra, podendo este ser um estado de relaxamento (934 instâncias) ou realizando uma atividade matemática (936 instâncias). Após isso, a base de dados ficou com 513 colunas, sendo elas 512 colunas com dados numéricos da atividade cerebral e a última coluna a atividade que o indivíduo estava realizando no momento da coleta (para técnicas supervisionadas).

Na terceira etapa foi aplicado a técnica de PCA na base, utilizando os auto vetores, resultando primeiramente em 5 principais componentes, e na segunda vez que foi aplicado, configura-se para

resultar uma base com 30 principais componentes, com isso diminuindo a dimensionalidade da base de dados de 513 colunas para 6 e para 31, respectivamente, pois a última coluna sempre contém os dados do estado do indivíduo no momento da coleta (*relax*, *math*), como demonstra a Figura 8. Contudo, para os algoritmos de aprendizado de máquina não supervisionado, a coluna *label* não é utilizada.

FIGURA 7: COLUNAS DA BASE DE DADOS APÓS APLICAÇÃO DE PCA COM 5 PRINCIPAIS COMPONENTES



FONTE: O AUTOR (2022).

Na etapa seguinte, foram aplicadas técnicas de aprendizado de máquina supervisionado e não-supervisionado, tanto na base de dados original quanto na base que resultou 5 principais componentes e 30 principais componentes. Utilizando sempre 70% da base como treino e 30% da base como teste, como demonstra a Figura 9.

FIGURA 8: APLICAÇÃO DA TÉCNICA DE APRENDIZADO DE MÁQUINA (KNN) E ACURÁCIA RESULTANTE.

```
[4] X = df1.drop('label',1)
    y = df1.label

    # Split dataset into training set and test set
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1) # 70% training and 30% test

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: In a future version of pandas all arguments
    """Entry point for launching an IPython kernel.
```

```
[5] #Import knearest neighbors Classifier model
    from sklearn.neighbors import KNeighborsClassifier

    #Create KNN Classifier
    knn = KNeighborsClassifier(n_neighbors=5)

    #Train the model using the training sets
    knn.fit(X_train, y_train)

    #Predict the response for test dataset
    y_pred = knn.predict(X_test)
```

```
[6] # Model Accuracy, how often is the classifier correct?
    print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

Accuracy: 0.5383244206773619
```

FONTE O AUTOR (2022).

Todos algoritmos foram executados em Python através do Google Colab e da utilização de bibliotecas como scikit-learn e pandas. Na última etapa, após a execução de todas as técnicas de aprendizado de máquina, foram coletadas as acurácias dos algoritmos aplicados em uma tabela para a comparação dos resultados.

Os algoritmos selecionados para a pesquisa foram: KNN (K - Nearest Neighbors), SVM (Support Vector Machine), Random Forest, Naive Bayes, Decision Tree Model (J48), Logistic Regression, MLP (Multi layer perceptron) e K-means, por serem os mais populares e devido ao amplo material disponível. Como para este estudo de caso, especificamente, não foram encontrados trabalhos similares, optou-se por utilizar algoritmos base de aprendizado de máquina para entendimento dos dados e de como a técnica de PCA pode ser utilizada nesta base. Os testes realizados utilizaram a base não balanceada.

A última etapa, de análise de resultados obtidos, foi realizada a partir da análise de acurácia entre as três bases de dados testadas: bruta, 5 componentes e 30 componentes, e a verificação do uso de PCA junto a esse estudo de caso.

4. ANÁLISE E DISCUSSÃO DOS DADOS

Após a programação e execução dos algoritmos para cada uma das três bases de dados, foram agrupadas as acurácias das técnicas de aprendizado de máquina para a comparação, resultando na Tabela 1.

O resultado obtido demonstra que a técnica de Naive Bayes foi a que mais apresentou melhora na acurácia após a aplicação do PCA. Em contrapartida, o MLP teve a acurácia reduzida em ambas aplicações do PCA. Há casos em que a aplicação de PCA resultando 30 componentes principais aumentou a acurácia da técnica de aprendizado de máquina, como podemos ver em SVM e Random Forest.

TABELA 1: ACURÁCIA DAS TÉCNICAS DE APRENDIZADO DE MÁQUINA

TÉCNICA SUPERVISIONADA	BASE	PCA 5 PC	PCA 30 PC
KNN	59.18	53.83	57.04
Random Forest	55.79	59.71	61.85
SVM	49.19	49.01	53.29
Naive Bayes	47.23	54.18	50.98
Decision Tree Model (J48)	54.36	52.40	56.50
Logistic Regression	53.29	55.25	52.94
TÉCNICA NÃO SUPERVISIONADA	BASE	PCA 5 PC	PCA 30 PC
MLP	55.97	51.69	53.29
k-means	50.10	50.05	50.58

FONTE: O AUTOR (2022).

Comparando as técnicas de aprendizado supervisionadas com as não supervisionadas, a acurácia diminui nos experimentos utilizando 5 principais componentes em todas as técnicas não supervisionadas, quanto as técnicas supervisionadas têm-se o Random Forest, Naive Bayes e Logistic Regression, que possuem um aumento de, no mínimo, 2%.

Entre as técnicas de aprendizado de máquina não supervisionadas, percebe-se que a acurácia não atinge nem 56% em nenhuma das técnicas utilizadas, sendo a mais alta 55.97% para a técnica MLP, executada na base bruta. Quanto às técnicas supervisionadas, tem-se o Random Forest que

atinge quase 62% de acurácia, quando executado com 30 principais componentes e o KNN, atingindo 59.19% na base bruta.

As técnicas que tiveram menos oscilação entre a acurácia nos três experimentos foram o k-mean e o SVM. Levando em consideração o experimento aplicando o PCA com 5 principais componentes, o Random Forest foi a técnica com maior acurácia seguido por Logistic Regression, ambas técnicas de aprendizado de máquina supervisionada.

Contudo, para a base de dados de EGG utilizada neste estudo, e com os algoritmos testados, não se verifica uma melhora significativa da aplicação da técnica de PCA. Mesmo o trabalho mais similar realizado a este, de Guerrero *et al.* (2021), não obteve resultados tão promissores quanto o desejado. E o trabalho de Guerrero *et al.* (2021) utiliza uma base com menor quantidade de atributos (apenas 19 eletrodos). Desta forma, ainda são necessários estudos para verificar se a técnica de PCA pode auxiliar na descoberta de conhecimento em sinais do tipo EEG.

Mesmo contendo uma diferença entre a quantidade de eletrodos utilizada no trabalho de Guerrero *et al.* (2021) e este trabalho (19 e 512 eletrodos), os resultados indicam que a técnica de PCA pode realizar a diminuição da dimensionalidade dos dados, sem a perda das principais características dos dados de EEG. Contudo, não foi possível, a partir das técnicas de aprendizado de máquina utilizadas, obter-se uma boa acurácia para os dados.

Desta forma, sabendo que o uso de PCA em dados de EEG não aumenta e nem diminui a acurácia de forma significativa, a técnica pode ser utilizada para problemas em que é preciso uma redução da dimensionalidade da base de dados, de maneira a poupar recursos computacionais ou diminuir tempo de processamento e custos.

5. CONCLUSÕES E TRABALHOS FUTUROS

A correta compreensão e interpretação de dados de EEG é de suma importância para uso em conjunto com técnicas de aprendizado de máquina para futuros trabalhos que possam impactar em tratamentos de doenças que atingem o cérebro. O EEG, sendo a técnica mais econômica e simples para se obter dados sobre a atividade cerebral, continua tendo amplo uso em estudos sobre aprendizado de máquina, devido ao fato de serem dados com uma grande quantidade de informações que podem ser extraídas. Desta maneira, a importância de continuar os estudos procurando melhorar as tecnologias já existentes, a fim de encontrar novas aplicações para este tipo de dado.

Este trabalho ajudou a demonstrar que para dados de EEG são necessárias mais pesquisas para compreender a sua complexidade e descobrir a melhor técnica para previsão do estado de um indivíduo, levando em consideração apenas os dados obtidos da atividade cerebral. Sendo assim, o trabalho alcançou o objetivo proposto, de analisar técnicas de aprendizado de máquina aplicadas em dados coletados através EEG, em um estudo de caso onde o sujeito está realizando uma atividade de matemática ou está relaxando.

Os resultados obtidos a partir deste trabalho poderão ser utilizados para estudos e comparação de novas técnicas que possam surgir a fim de realizar mineração de dados de EEG, possibilitando assim o desenvolvimento de melhores sistemas ou dispositivos, assim auxiliando na compreensão das atividades cerebrais ou para realização de classificações (definições de padrões).

Embora o PCA, em sua forma padrão, seja uma ferramenta de análise descritiva de dados amplamente utilizada e adaptável, ele também possui muitas adaptações próprias, que o tornam útil para uma ampla variedade de situações e tipos de dados em vários contextos. Neste trabalho, mesmo não aumentando significativamente a acurácia das técnicas de aprendizado de máquina, o PCA contribuiu para a redução da base original sem a perda de características dos dados e melhorando a velocidade de processamento devido a diminuição da quantidade de dados processados. Relacionado ao desempenho de processamento, não foi realizada uma análise minuciosa, pois a base utilizada não foi muito grande, mas percebeu-se que o processamento das técnicas de aprendizado de máquina entre os dados da base bruta e das bases com 5 ou 30 componentes principais foi realizada em tempos diferentes.

Para trabalhos futuros pretende-se utilizar outras bases de dados de EEG para futuras comparações e também aplicação de técnicas de balanceamento de dados e *deep learning* na base de dados atual. Trabalhos futuros podem explorar outras técnicas de aprendizado de máquina (além de *deep learning*) para obtenção de informações sobre a atividade cerebral, bem como outras técnicas de redução de dimensionalidade, como ICA (*Independent Component Analysis*) ou LDA (*Linear Discriminant Analysis*) podem ser utilizadas para testes e comparações. Além disso, pode-se testar

diferentes quantidades de exemplos de referências para testar a sensibilidade dos resultados da aplicação.

REFERÊNCIAS

- ALI, Jehad et al. Random forests and decision trees. **International Journal of Computer Science Issues (IJCSI)**, v. 9, n. 5, p. 272, 2012.
- ALI, Najat; NEAGU, Daniel; TRUNDLE, Paul. Evaluation of k-nearest neighbor classifier performance for heterogeneous data sets. **SN Applied Sciences**, v. 1, n. 12, p. 1-15, 2019.
- AMIN, Hafeez Ullah et al. Classification of EEG signals based on pattern recognition approach. **Frontiers in computational neuroscience**, v. 11, p. 103, 2017.
- BAŞTANLAR, Yalin; ÖZUYSAL, Mustafa. Introduction to machine learning. **miRNomics: MicroRNA biology and computational analysis**, p. 105-128, 2014.
- BASTOS, Narusci S.; ADAMATTI, Diana F.; BILLA, Cleo Z. Discovering patterns in brain signals using decision trees. **Computational Intelligence and Neuroscience**, v. 2016, 2016.
- BEN-HUR, Asa et al. Support vector machines and kernels for computational biology. **PLoS computational biology**, v. 4, n. 10, p. e1000173, 2008.
- BISHOP, Christopher M.; NASRABADI, Nasser M. **Pattern recognition and machine learning**. New York: springer, 2006.
- BREIMAN, Leo. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.
- BRO, Rasmus; SMILDE, Age K. Principal component analysis. **Analytical methods**, v. 6, n. 9, p. 2812-2831, 2014.
- CAR, Zlatan et al. Modeling the spread of COVID-19 infection using a multilayer perceptron. **Computational and mathematical methods in medicine**, v. 2020, 2020.
- CHOUDHARY, Rishabh; GIANEY, Hemant Kumar. Comprehensive review on supervised machine learning algorithms. In: **2017 International Conference on Machine Learning and Data Science (MLDS)**. IEEE, 2017. p. 37-43.
- DOMA, Vikrant; PIROUZ, Matin. A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals. **Journal of Big Data**, v. 7, n. 1, p. 1-21, 2020.

FARSI, Leila et al. Classification of alcoholic EEG signals using a deep learning method. **IEEE Sensors Journal**, v. 21, n. 3, p. 3552-3560, 2020.

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

GUERRERO, Maria Camila; PARADA, Juan Sebastián; ESPITIA, Helbert Eduardo. Principal Components Analysis of EEG Signals for Epileptic Patient Identification. **Computation**, v. 9, n. 12, p. 133, 2021.

GHOSH, Soumi; DUBEY, Sanjay Kumar. Comparative analysis of k-means and fuzzy c-means algorithms. **International Journal of Advanced Computer Science and Applications**, v. 4, n. 4, 2013.

HERNÁNDEZ, José; FERRI, C.; RAMÍREZ, M. J. Introducción a la Minería de Datos. 2004.

HONGYU, Kuang; SANDANIELO, Vera Lúcia Martins; DE OLIVEIRA JUNIOR, Gilmar Jorge. Análise de componentes principais: resumo teórico, aplicação e interpretação. **E&S Engineering and science**, v. 5, n. 1, p. 83-90, 2016.

JAISWAL, Abeg Kumar; BANKA, Haider. Epileptic seizure detection in EEG signal using machine learning techniques. **Australasian physical & engineering sciences in medicine**, v. 41, n. 1, p. 81-94, 2018.

JOHNSON, Richard Arnold et al. **Applied multivariate statistical analysis**. London, UK:: Pearson, 2014.

JOLLIFFE, Ian T.; CADIMA, Jorge. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 374, n. 2065, p. 20150202, 2016.

LENT, R. **Neurociência da mente e do comportamento**. Rio de Janeiro, Guanabara Koogan, 2008, 372 p.

LIBRELOTTO, Solange Rubert; MOZZAQUATRO, Patricia Mariotto. Análise dos algoritmos de mineração J48 e Apriori aplicados na detecção de indicadores da qualidade de vida e saúde. **Revista Interdisciplinar de Ensino, Pesquisa e Extensão-RevInt**, v. 1, n. 1, 2014.

MUNIAN, S; SIVALINGAM, S; JAYARAMAN, Vinoth. **Analysis of Real Time EEG Signals**. 2014. 44 f. Dissertação (Mestrado) - Curso de Electrical Engineering, School Of Computer Sciences, Physics And Mathematics, Linnaeus University, Växjö, Sweden, 2014.

NIEDERMEYER, Ernst; DA SILVA, FH Lopes (Ed.). **Electroencephalography: basic principles, clinical applications, and related fields**. Lippincott Williams & Wilkins, 2005.

PANDEY, Amit; JAIN, Achin. Comparative analysis of KNN algorithm using various normalization techniques. **International Journal of Computer Network and Information Security**, v. 9, n. 11, p. 36, 2017.

PAI, Ping-Feng; HONG, Wei-Chiang. Software reliability forecasting by support vector machines with simulated annealing algorithms. **Journal of Systems and Software**, v. 79, n. 6, p. 747-755, 2006.

RAO, Rajesh PN. **Brain-computer interfacing: an introduction**. Cambridge University Press, 2013.

RISH, Irina et al. An empirical study of the naive Bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**. 2001. p. 41-46.

SAVALIA, Shalin; EMAMIAN, Vahid. Cardiac arrhythmia classification by multi-layer perceptron and convolution neural networks. **Bioengineering**, v. 5, n. 2, p. 35, 2018.

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

SILVEIRA, Josimara de Ávila. **Análise de sinais cerebrais utilizando árvores de decisão**. 2013. Dissertação de Mestrado (Programa de Pós-Graduação em Modelagem Computacional). Universidade Federal do Rio Grande.

TABACHNICK, Barbara G.; FIDELL, Linda S.; ULLMAN, Jodie B. **Using multivariate statistics**. Boston, MA: pearson, 2007.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**, Boston: Person Education. 2006.

VAN GESTEL, Tony et al. Financial time series prediction using least squares support vector machines within the evidence framework. **IEEE Transactions on neural networks**, v. 12, n. 4, p. 809-821, 2001.

WANG, Xiao-Wei; NIE, Dan; LU, Bao-Liang. Emotional state classification from EEG data using machine learning approach. **Neurocomputing**, v. 129, p. 94-106, 2014.

WIDODO, Achmad; YANG, Bo-Suk. Support vector machine in machine condition monitoring and fault diagnosis. **Mechanical systems and signal processing**, v. 21, n. 6, p. 2560-2574, 2007.

WITTEN, Ian H. et al. Practical machine learning tools and techniques. In: **DATA MINING**. 2005. p. 4.

WOLPAW, Jonathan R. et al. Brain-computer interface technology: a review of the first international meeting. **IEEE transactions on rehabilitation engineering**, v. 8, n. 2, p. 164-173, 2000.

ZIO, Enrico et al. A support vector machine integrated system for the classification of operation anomalies in nuclear components and systems. **Reliability engineering & system safety**, v. 92, n. 5, p. 593-600, 2007.