Research, Society and Development, v. 9, n. 11, e86191110584, 2020
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v9i11.10584

Aplicação de algoritmos de aprendizado de máquina na classificação de Conhecimentos Especializados de Professores de Física

Appication of machine learning algorithms in the Classification of Specialized Knowledge of Physics Teachers

Aplicación de algoritmos de aprendizaje automático en la Clasificación de Conocimientos Especializados de Profesores de Física

Recebido: 24/11/2020 | Revisado: 26/11/2020 | Aceito: 02/12/2020 | Publicado: 05/12/2020

Tamara Aguiar Tavares Mascarenhas
ORCID: https://orcid.org/0000-0002-1995-8169
Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso, Brasil
E-mail: tammi.aguiar@gmail.com
Jeferson Gomes Moriel Junior
ORCID: https://orcid.org/0000-0003-1526-8002
Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso, Brasil
Raphael de Souza Rosa Gomes
ORCID: https://orcid.org/0000-0003-1591-5281
Universidade Federal de Mato Grosso, Brasil
Geison Jader Mello
ORCID: https://orcid.org/0000-0002-0991-2327
Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso, Brasil

Resumo

O sucesso da Inteligência Artificial tem atraído pesquisadores de diversas áreas para o uso de técnicas computacionais em tarefas de extração de conhecimentos de dados não estruturados, como os documentos textuais, apresentando-se como uma solução possível para a classificação de Conhecimentos Especializado de Professores de Física, que consiste em uma ferramenta analítica que descreve os Conhecimentos da Física (PK) e os conhecimentos Didáticos do Conteúdo (PCK), considerada muito importante para auxiliar na identificação e na análise de conhecimentos mobilizados pelos professores em suas práticas de ensino. Porém, a tarefa de identificação e classificação de conhecimentos presentes em documentos textuais apresentam alguns desafios, como: as investigações em documentos textuais é

trabalhosa, demorada e envolve mão de obra de pessoas especializadas. Nesse sentido, o objetivo da pesquisa é analisar a eficácia dos algoritmos utilizados na classificação automática de Conhecimentos Especializados de Professores de Física (PTSK) em textos de uma base de dados previamente classificada. O encaminhamento metodológico é de natureza quantitativa, exploratória e aplicada para prever a classe Conhecimentos da Física (PK) ou Conhecimentos Didáticos do Conteúdo (PCK) para caracterizar um conhecimento. Para isso, foram usados dois algoritmos: o *doc2vec* e o J48 e os resultados foram analisados com base nos resultados alcançados nas métricas de validação. O melhor resultado foi alcançado com o doc2vec, obtendo 88% de taxa de acerto. Com base nos resultados atingidos, pode-se concluir que a estratégia de usar inteligência artificial para a classificação automática de conhecimentos de professores de Física é uma solução plausível.

Palavras-chave: Aprendizagem de máquina; Classificação de conhecimentos; PTSK; *Doc2vec*; J48.

Abstract

The success of Artificial Intelligence has attracted researchers from different areas to use computational techniques in tasks of extracting knowledge from unstructured data, such as textual documents, presenting itself as a possible solution for the classification of Specialized Knowledge of Physics Teachers , which consists of an analytical tool that describes the Knowledge of Physics (PK) and Didactic knowledge of Content (PCK), considered very important to assist in the identification and analysis of knowledge mobilized by teachers in their teaching practices. However, the task of identifying and classifying knowledge present in textual documents presents some challenges, such as: investigating textual documents is laborious, time-consuming and involves the labor of specialized people. In this sense, the objective of the research is to analyze the effectiveness of the algorithms used in the automatic classification of Expert Knowledge of Physics Teachers (PTSK) in texts from a previously classified database. The methodological approach is quantitative, exploratory and applied to predict the class Knowledge of Physics (PK) or Didactic Knowledge of Content (PCK) to characterize knowledge. For this, two algorithms were used: doc2vec and J48 and the results were analyzed based on the results achieved in the validation metrics. The best result was achieved with doc2vec, obtaining an 88% success rate. Based on the results achieved, it can be concluded that the strategy of using artificial intelligence for the automatic classification of knowledge of Physics teachers is a plausible solution.

Keywords: Machine learning; Knowledge classification; PTSK; Doc2vec; J48.

Resumen

El éxito de la Inteligencia Artificial ha atraído a investigadores de diferentes áreas a utilizar técnicas computacionales en tareas de extracción de conocimiento a partir de datos no estructurados, como documentos textuales, presentándose como una posible solución para la clasificación de Conocimientos Especializados de Profesores de Física. , que consiste en una herramienta analítica que describe el Conocimiento de la Física (PK) y el Conocimiento Didáctico de los Contenidos (PCK), considerados muy importantes para ayudar en la identificación y análisis de los conocimientos movilizados por los docentes en sus prácticas docentes. Sin embargo, la tarea de identificar y clasificar el conocimiento presente en los documentos textuales presenta algunos desafíos, tales como: investigar documentos textuales es laborioso, requiere mucho tiempo e implica el trabajo de personas especializadas. En este sentido, el objetivo de la investigación es analizar la efectividad de los algoritmos utilizados en la clasificación automática del Conocimiento Experto de Profesores de Física (PTSK) en textos de una base de datos previamente clasificada. El enfoque metodológico es cuantitativo, exploratorio y aplicado para predecir la clase Conocimiento de la Física (PK) o Conocimiento Didáctico del Contenido (PCK) para caracterizar conocimientos. Para ello se utilizaron dos algoritmos: doc2vec y J48 y los resultados se analizaron en base a los resultados obtenidos en las métricas de validación. El mejor resultado se logró con doc2vec, obteniendo una tasa de éxito del 88%. A partir de los resultados obtenidos, se puede concluir que la estrategia de utilizar la inteligencia artificial para la clasificación automática del conocimiento de los profesores de Física es una solución plausible.

Palabras clave: Aprendizaje automático; Clasificación del conocimiento; PTSK; Doc2vec; J48.

1. Introdução

O Conhecimento Especializado de Professores é uma linha de investigação que estuda e oferece relevantes contribuições para a formação docente, visando a qualidade do ensino e, consequentemente, a valorização e o reconhecimento do professor. O foco das pesquisas nessa área é oferecer subsídios teóricos, considerados úteis e essenciais às práticas de ensino e aprendizagem do professor.

Os estudos nesse área se intensificaram a partir da década de 1980 indicando um novo momento para a formação dos professores e abrindo caminho para a discussão sobre a valorização e o reconhecimento da identidade profissional do professor, tendo como um dos

seus aspectos, a compreensão da prática docente, a partir dos conhecimentos profissionais constituídos e mobilizados nas ações desempenhadas pelo professor no ambiente escolar. (Pimenta, 1997; Malanchen, 2012; Lima et. al., 2020). A amplitude dessa temática tem demandando estudos sob diferentes enfoques acerca dos conhecimentos e saberes dos professores (Shulman, 1986; 1987; Nóvoa, 1991; Gautier, 1998; Pimenta, 2012; Contreras, 2002; Tardif, 2012; Fernandez, 2015).

Um dos autores mais referenciados nessa linha de investigação é Lee Shulman (1986). A sua pesquisa evidenciou a existência de uma rede complexa de saberes e habilidades que eram únicos do ato de ensinar, constituída de uma base de conhecimentos que envolvem: o conhecimento do conteúdo, o conhecimento pedagógico do conteúdo, o conhecimento dos alunos e de suas características, o conhecimento dos contextos educacionais e o conhecimento dos fins, propósitos e valores da educação bem como sua base histórica e filosófica. A inter-relação desses conhecimentos resultou no constructo PCK - Conhecimento Pedagógico do Conteúdo (em inglês, *Pedagogical Content Knowledge* (Shulman, 1986).

A partir dos pressupostos de Shulman (1986), diversos autores (Grossman, 1990; Park Oliver, 2008) contribuíram para o avanço e o fortalecimento dessa linha de investigação. Outros, ampliaram e adequaram o modelo para áreas específicas, como Matemática e Ciências da Natureza.

Na área da Matemática um dos modelos que se destacou foi o Conhecimento Matemático para o Ensino (em inglês - *Mathematical Knowledge for Teaching* – MKT), desenvolvido do refinamento das categorias de Shulman para docentes exclusivamente da área da Matemática (Ball; Thames; Phelps, 2008). Seguindo as investigações em ensino de Matemática, José Carrillo (2014) e o grupo SIDM identificaram limitações no MKT que resultou na criação de um novo marco teórico para o ensino da Matemática: o Modelo de Conhecimento Especializado de Professores de Matemática – MTSK (em inglês, Mathematics Teacher's Specialized Knowledge – MTSK).

O reconhecimento do modelo MTSK no mundo (Kilpatrick, Spangler, 2015) culminou na sua transposição para disciplinas da área das Ciências da Natureza; a priore, para a disciplina da Biologia com o modelo teórico Conhecimento Especializado de Professores de Biologia (em inglês, *Biology Teacher's Specialized Knowledge* – BTSK) (Luís, 2015), e mais recentemente para a Física com o Conhecimento Especializado de Professores de Física (em inglês, *Physics Teacher's Specialized Knowledge* – PTSK) (Lima, 2018) e a Química com o Conhecimento Especializado de Professores de Química (em inglês, *Chemistry Teacher's Specialized Knowledge* – CTSK) (Soares, 2019). Sendo, os dois últimos, desenvolvidos por

pesquisadoras brasileiras pertencentes ao grupo de pesquisa TSK *Group* (IFMT-CBA). Esses modelos teóricos apresentam-se como uma importante ferramenta metodológica e analítica para investigar as distintas práticas do professor, a partir das dimensões do seu conhecimento da disciplina e do seu conhecimento pedagógico (Lima, 2018).

O modelo teórico de Conhecimento Especializado de Professores de Física - PTSK (Lima, 2018), é uma transposição direta do MTSK que foi adaptado da Matemática para Física. O PTSK foi desenvolvido tendo como fundamento teórico a abordagem epistemológica sócio construtivista com raízes no sócio-construtivismo pedagógico e sócio-histórico (Carrillo et al., 2014), pesquisas sobre o ensino de Física (Salem, 2012), Aprendizagem Significativa (Ausubel, 1980) e Aprendizagem Significativa Crítica (Moreira, 2017).

O modelo Conhecimento Especializado de Professores de Física (PTSK), descreve os conhecimentos que o professor de Física precisa ter para o exercício da docência. Esse modelo possui a estrutura no formato de um hexágono e é composto por dois domínios- Conhecimento da Física (PK) e o Conhecimento Didático do Conteúdo (PCK), onde cada domínio é subdividido em três subdomínios, além de apresentar as crenças dos professores caracterizadas pelas suas ações. (Lima, 2018).

No âmbito dessa linha de estudo, pesquisadores têm desenvolvido investigações desses conhecimentos especializados mobilizados pelos professores em diferentes registros que retratam práticas reais de ensino. Incluem-se nesses registros os documentos textuais, representados por: relatos de experiência, resumos, PaP-eRs - Relatório da Experiência Profissional Pedagógica (Loughran et al., 2001), artigos científicos, dissertações e teses.

No entanto, a tarefa de identificar e classificar os conhecimentos dos professores apresenta três grandes desafios: seleção manual de trechos em um grande número de documentos textuais; a identificação dos conhecimentos é desenvolvida apenas por mão de obra especializada e, a análise de conhecimentos exige um trabalho manual, intensivo e moroso.

Todavia, um novo cenário se apresenta, e a Inteligência Artificial - IA pode ser uma ferramenta importante no enfrentamento desses desafios diante do sucesso dos sistemas inteligentes, sobretudo os baseados em Aprendizagem de Máquina - AM, que envolve uma variedade de algoritmos, incluindo a árvore de decisão e as redes neurais.

A IA tem atraído pesquisadores de diversas áreas para o uso de técnicas computacionais em tarefas de extração de conhecimentos de dados não estruturados, como os documentos textuais. Apesar da AM existir há muito tempo, foi a partir da década de 80 que o

interesse por esse tema cresceu, motivada, principalmente, pelo aumento do poder computacional disponível e pela disponibilização de um grande conjunto de dados. Desde então, técnicas de AM têm sido aplicadas em diversos cenários, obtendo resultados comparáveis aos obtidos pelo ser humano (He et al., 2015).

Os estudos envolvendo documentos textuais implicam na conexão com outras áreas de conhecimento, tais como o Processamento de Linguagem Natural – NLP, um campo da ciência da computação destinado a fazer com que os computadores tenham a capacidade de entender texto e a Mineração de Textos (MT, em inglês Text Mining), que se trata de um conjunto de técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras em processos de descoberta de conhecimento, e envolve a aplicação de algoritmos computacionais para processar textos e identificar informações úteis e implícitas, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta (Morais; Ambrósio, 2007).

Na literatura, encontram-se alguns algoritmos de classificação que são usados para gerar modelos ou classificadores para descrever as classes que são utilizados com o propósito de identificar exemplos que ainda não foram classificados. O modelo é criado a partir do treinamento do classificador através de um conjunto de dados, corretamente rotulados, denominado conjunto de treinamento. A performance da classificação é obtida quando um conjunto de teste (sem rótulos), é submetido ao modelo (Santos, 2016). Alguns exemplos de algoritmos utilizados na tarefa de classificação são: Árvores de Decisão, Redes Neurais Artificiais, Aprendizado Tardio (Witten; Frank; Hall, 2005), Regras de Classificação, Aprendizado Bayesiano (Han; Kamber; Pei, 2011), Máquinas de Vetores (Maimon; Rokach, 2005; Han et. al., 2011).

Outra técnica que tem sido muito usada recentemente, é a representação de palavras em vetores a partir de modelos de redes neurais muito usada na área de PLN em tarefas de classificação. Esses vetores são gerados por meio das redes neurais recorrentes (RNN - *Recurrent Neural Network*).

Tendo em vista o potencial dessas tecnologias, a proposta desse trabalho surgiu junto ao grupo de pesquisa TSK Group - Teacher's Specialized Knowledge Research Group (IFMT-CBA), apoiados pela Rede Iberoamericana MTSK (Moriel Junior, 2014; Moriel Junior; Wielewski, 2017; Vasco; Moriel Junior; Contreras, 2017; Moriel Junior; Alencar, 2020), apresentando-se como uma pesquisa multidisciplinar na área do ensino e na ciência da computação, cujo intuito é aplicar as técnicas de árvore de decisão e redes neurais na tarefa de

classificação automática de conhecimentos especializados de professores de Física, identificados em trechos previamente classificados.

Portanto, esse trabalho se justifica porque nele é apresentado, testado e analisado um caminho mais rápido e eficiente de fazer a classificação de conhecimentos de professores de Física, apresentando-se como uma inovação científica nessa linha de pesquisa, que pode auxiliar os pesquisadores da área no desenvolvimento de estudos futuros.

Na presente pesquisa procurou-se responder a pergunta: Quais foram os resultados obtidos das técnicas de Aprendizado de Máquina aplicadas na classificação automática de conhecimentos especializados de professores de Física?

O objetivo Geral desse trabalho é analisar a acurácia das técnicas de Aprendizado de Máquina utilizadas para classificação automática de conhecimentos especializados de professores de Física. E para atender o objetivo geral, foram traçados os seguintes objetivos específicos: estudar técnicas de aprendizado de máquina empregadas em tarefas de classificação de textos; experimentar técnicas de Aprendizado de Máquina na classificação de conhecimentos; e analisar os resultados alcançados nos testes realizados.

2. Metodologia

2.1 Classificação da pesquisa

O aporte metodológico utilizado para caracterizar essa pesquisa é apresentado no Tabela 1, contendo a descrição da classificação quanto aos objetivos da pesquisa, a natureza da pesquisa, a abordagem da pesquisa, a técnica de coleta e a técnica de análise de dados.

Tabela 1. Classificação Da Metodologia Científica**.**

CLASSIFICAÇÃO QUANTO AOS OBJETIVOS DA PESQUISA	Exploratória	Os estudos exploratórios, geralmente, são úteis para diagnosticar situações, explorar alternativas ou descobrir novas ideias (ZIKMUND 2000).
CLASSIFICAÇÃO QUANTO À NATUREZA DA PESQUISA	Aplicada	É aplicada por caracterizar um trabalho prático, uma vez que os resultados são utilizados na solução de problemas que ocorrem na realidade (MARCONI & LAKATOS, 2001).
CLASSIFICAÇÃO QUANTO À ABORDAGEM DA PESQUISA	Quantitativa	É caracterizada pelo emprego da quantificação, tanto nas modalidades de coleta de informações quanto no tratamento delas por meio de técnicas estatísticas (RICHARDSON, 1999).
CLASSIFICAÇÃO QUANTO À TÉCNICA DE COLETA DE DADOS	Bibliográfica	A pesquisa bibliográfica é desenvolvida quando o pesquisador tem contato direto com material já elaborado, constituído, principalmente, de livros e artigos científicos, sendo importante para o levantamento de informações básicas sobre os aspectos ligados à temática da pesquisa. Esses materiais devem ser analisados, para obter o conhecimento necessário, para por fim elaborar um novo trabalho usando como base o material analisado anteriormente. (GIL, 1999, VERGARA, 2000; LAKATOS & MARCONI, 2001).
CLASSIFICAÇÃO QUANTO À TÉCNICA DE	Estatística descritiva	O objetivo da estatística descritiva é o de representar, de forma concisa, sintética e compreensível, a informação

ANÁLISE DE DADOS		contida num conjunto de dados. Esta tarefa, que adquire grande importância quando o volume de dados for grande, concretiza-se na elaboração de tabelas e de gráficos, e no cálculo de medidas ou indicadores que representam convenientemente a informação contida nos dados (MARCONI & LAKATOS, 2001).

Fonte: Adaptado de Oliveira (2011).

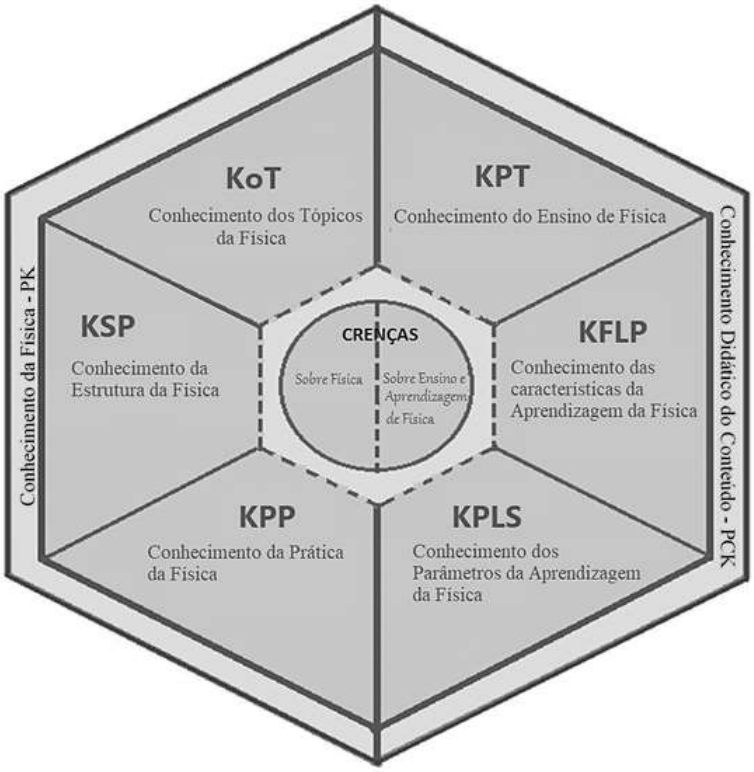
Essas características metodológicas são muito frequentes em pesquisas que envolvem a computação. Embora esse seja um estudo aplicado à educação no âmbito do ensino, há uma forte influência da objetividade e da perspectiva positivista, diante da mensuração que é realizada no processo de classificação de conhecimentos de professores de Física, o que possibilita a análise estatística dos resultados atingidos com os testes realizados.

Embora o estudo tenha um caráter objetivo, vale ressaltar que existe um viés na perspectiva construtivista pautada no trabalho de Piaget, pois a tecnologia entra na área da educação como um fator de colaboração, por meio de um processo automatizado, para melhorar uma atividade que tem sido desenvolvida manualmente por pesquisadores que trabalham com conhecimento especializado de professores.

2.2 Contexto e fontes

Pesquisadores que fazem parte do grupo de pesquisa TSK *Group* tem desenvolvido estudos na área de conhecimentos especializados de professores para investigar os conhecimentos que são necessários ao professor para o ensino. Desses estudos, destaca- se o Conhecimento Especializado de Professores de Física (PTSK) (Lima, 2018), que descrevesse os conhecimentos que constituem a base da docência do ensino de Física para possibilitar que o professor os aplique às variações de contexto e habilidades na sua prática de ensino. Dentre os conhecimentos presentes no modelo tem-se os domínios Conhecimento da Física (PK) e o Conhecimento Didático do Conteúdo (PCK), conforme apresentados na Figura 1.

Figura 1. Conhecimento Especializado de Professores de Física (PTSK).

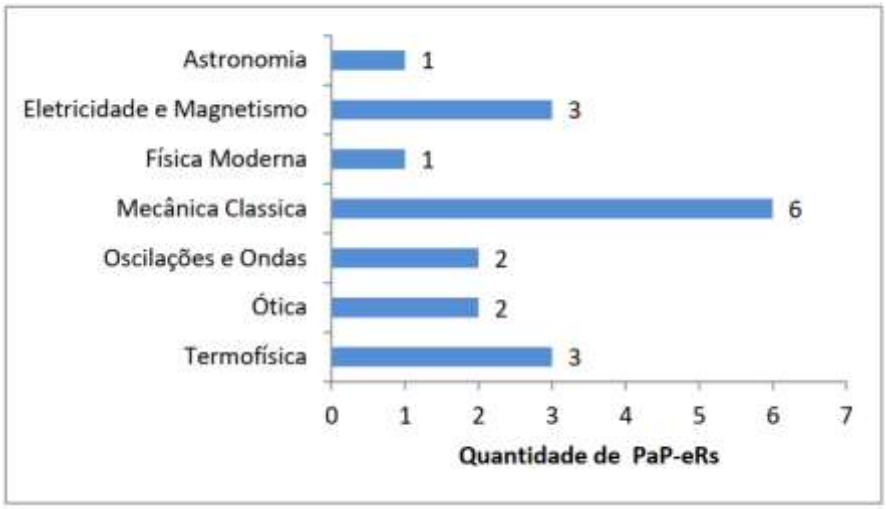


Fonte: Lima (2018, p. 108)

Para aperfeiçoar o trabalho de investigação desses conhecimentos, que tem sido realizado de forma manual pelos pesquisadores, é proposto um estudo que envolve a investigação da eficiência de algoritmos de Aprendizagem de Máquina, que seguem o paradigma de aprendizado supervisionado (Breve, 2010) na predição da classificação de conhecimentos especializados de Física.

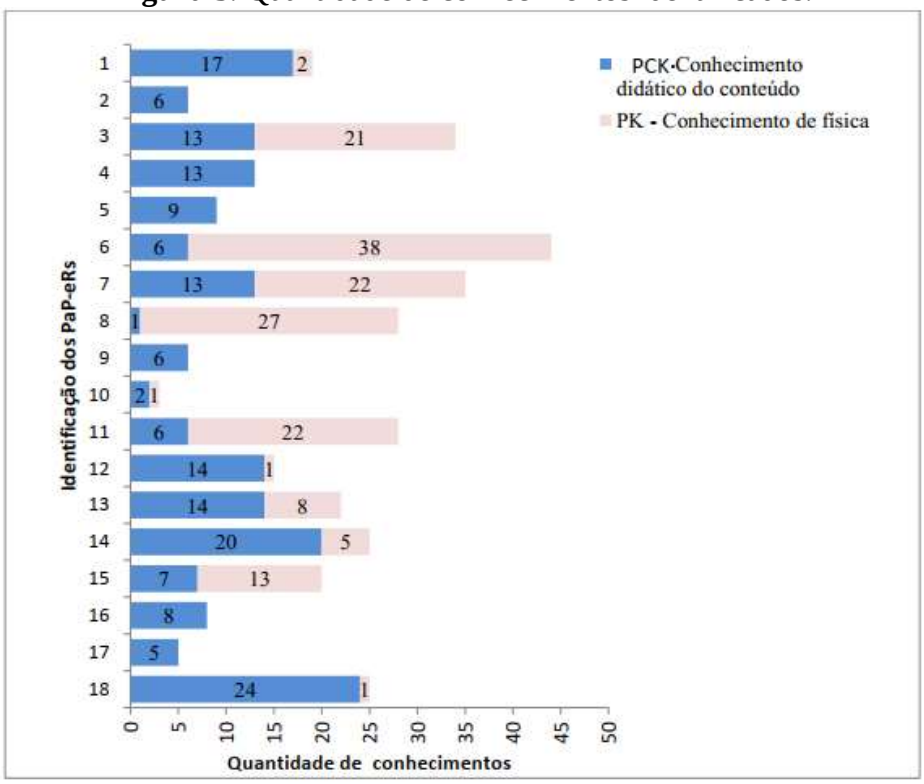
Nesse sentido, essa pesquisa se insere em trabalho prático aplicado a uma base de dados previamente classificada, construída em um trabalho anterior por Lima (2018). A composição da base de dados seguiu alguns critérios de seleção, que envolveu uma busca por publicações científicas nas diversas áreas de abrangência da Física, conforme estabelecido na Base Nacional Curricular Comum para o Ensino Médio e que apresentasse as características de um Relatório da Experiência Profissional Pedagógica - PaP-eRs (Loughran et al., 2001), definido como um documento que emerge da prática real dos professores. Nessa etapa foram selecionados 18 PaP-eRs abrangendo as unidades temáticas de Astronomia, Eletricidade e Magnetismo, Física Moderna, Mecânica Clássica, Oscilações e Ondas, ótica e Termofísica, apresentadas na Figura 2.

Figura 2. Unidades temáticas da Física.



Fonte: Lima, (2018).

Figura 3. Quantidade de conhecimentos identificados.



Fonte: Lima, (2018).

Ao todo foram identificados 346 conhecimentos, representados na Figura 3. Em azul destacam-se os conhecimentos extraídos do domínio Conhecimento Didático do Conteúdo (PCK) e em rosa destacam-se os conhecimentos extraídos do domínio Conhecimento de Física (PK).

Nessa pesquisa foram usadas duas aplicações de classificação (predição). O *doc2vec* (vetores de parágrafos) (Le & Mikolov, 2014), técnica que transforma cada documento em vetor no espaço multidimensional para realizar as representações distribuídas de frases e documentos ao aplicar rede neural para prever a probabilidade de palavras frequentes em um parágrafo e estabelecer a relação semântica entre elas (Feldman & Sanger, 2006).

Esse algoritmo foi executado no software *Intelij IDEA*, que consiste em um ambiente integrado, escrito em Java para o desenvolvimento de aplicações inteligentes, usando as bibliotecas disponíveis no *Deeplearning4j*

A outra aplicação utilizada foi o J48, uma implementação do algoritmo C4.5 (Quinlan, 1993), que constrói um modelo de árvore de decisão baseado em um conjunto de dados de treinamento, onde, a cada nó, o algoritmo escolhe o melhor atributo para subdividir o conjunto das amostras em subconjuntos homogêneos e caracterizados por sua classe. O fator principal é o ganho de informação obtida na escolha do atributo para subdivisão (Hall et al., 2009; Almeida et al., 2003; Giasson, et al. 2013).

Esse algoritmo foi executado no software Weka (*Waikato Enviroment for Knowledge Analysis*), que dispõe de uma coleção de ferramentas de visualização e algoritmos para análise de dados e modelagem preditiva.

2.3 Modelagem

2.3.1 Coleta de Dados

Os dados foram coletados de uma base de dados composta por 346 trechos, que representam os conhecimentos de professores identificados nos PaP-eRs. Esses trechos representam aproximadamente 46% no domínio Conhecimento da Física (PK) e 54% no domínio Conhecimento Didático do Conteúdo (PCK) (Lima, 2018).

Dos 346 só foram utilizados 318 porque alguns desses trechos estavam no formato de figura. Foi desenvolvido um algoritmo para ler os trecho que estavam originalmente em uma base de dados numa planilha e transformá-los no formato de texto. Esse algoritmo desprezou os trechos que estavam no formato de figura. Os dados foram divididos o treinamento e testes em um esquema de validação cruzada, que é uma técnica muito empregada em tarefas de classificação.

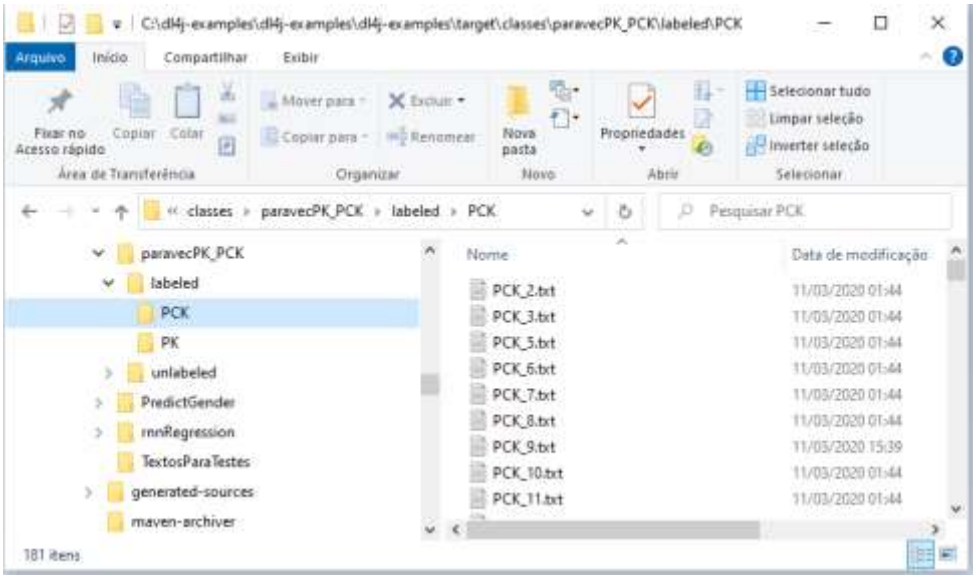
2.3.2 Pré-processamento

A finalidade dessa etapa é transformar o conjunto de documentos em uma base mais limpa, onde o trabalho de representação de documentos, o respectivo processamento dos dados e a consequente interpretação destes, possam ser feitas de maneira mais rápida e eficiente (Passini, 2012).

Porém, antes de realizar o pré-processamento dos dados, é necessário prepará-los de acordo com a particularidades específicas do ambiente de desenvolvimento.

Os dados manipulados no *InteliJ IDEA* devem possuir a extensão no formato de texto (.txt) e devem ser organizados em duas pastas: *labeled* (rotulado - dados com as categorias conhecidas pelo algoritmo) para armazenar os dados de treinamento e criar um modelo de classificação (Figura 4) e *unlebeled* (não rotulado - dados com as categorias não conhecidas pelo algoritmo) para armazenar os dados de testes e predizer a categoria a qual o trecho deve pertencer (Figura 5).

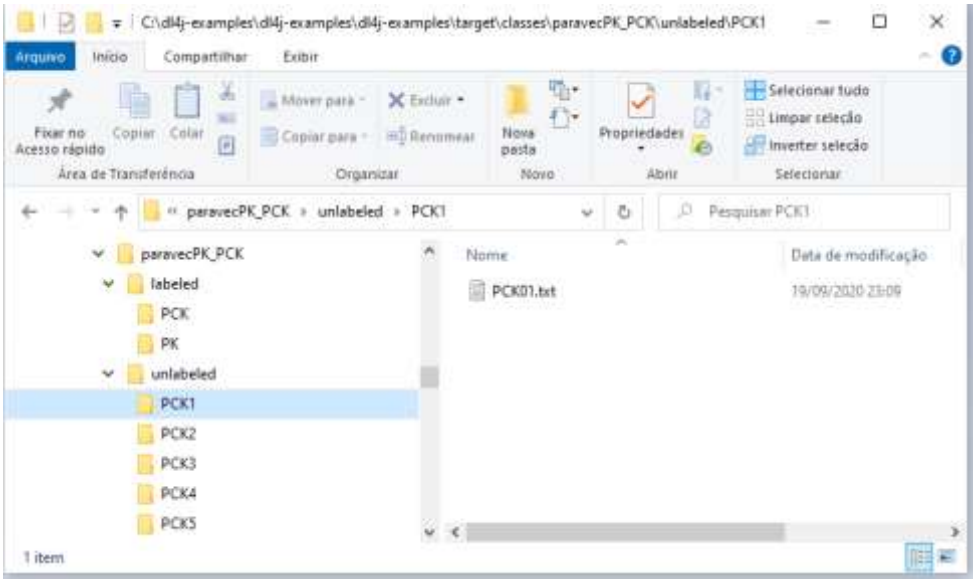
Figura 4. Organização dos arquivos usados no treinamento.



Fonte: Autores, (2020).

Na Figura 4, é possivel observar a organização dos arquivos que correspondem aos conhecimentos do domínio do Conhecimento Didático do conteúdo (PCK), organizados dentro da pasta PCK, que encontra-se dentro da pasta *labeled*. O mesmo padrão ocorre com os arquivos do Conhecimento da Física (PK). Todos os arquivo são rótulados com a sigla ao qual o conhecimento pertence e um número.

Figura 5. Organização dos arquivos usados nos testes.



Fonte: Autores, (2020).

Na Figura pode-se observar que dentro de cada pasta do *unlabeled* é armazenado apenas um arquivo que corresponde ao conhecimento que vai ser usado para o algoritmo predizer a classificação em Conhecimento da Física PK ou Conhecimento Didático da Física PCK. Todos os arquivos são rotulados conforme a classificação manual, realizada por uma especialista para facilitar a análise dos resultados.

No WEKA é utilizado como arquivo padrão para as tarefas de mineração de textos o formato Attribute Relation File Format (arff). A estrutura do arquivo é dividida em duas seções: cabeçalho e dados. O cabeçalho contém a identificação da base de dados usando o comando @relation e os atributos com o comando @attribute, definem o tipo dos dados que a base de dados contém, dando um nome a cada um deles e definindo o seu tipo e as categorias. As categorias da classificação são colocadas entre parênteses '{}'. Já os dados, representam uma lista de trechos (instâncias) organizados cada um em uma linha, que compartilham um conjunto de atributos. Os documentos para o treino tem os trechos organizados entre aspas duplas, separado por vírgula e a indicação da categoria que pertence. Caso o documento seja o de treino, utiliza-se o sinal de interrogação '?' no lugar da categoria para indicar que ela é desconhecida (Pellucci, et al., 2011; Alcântara, 2012).

Para exemplificar a estrutura do arquivo para o treinamento, segue um recorte dos arquivos que foram usados no treinamento, conforme apresentado na Figura 6.

Figura 6. Formato do arquivo para o treinamento.

```
@relation train

@attribute Documento string
@attribute classes {PK, PCK}

@data

"Voltando ao nosso procedimento, após [...] e dimensões de uma cartolina. ",PCK
"é o momento em que os estudantes [...] comparado a aulas tradicionais.",PCK
"Atividades desenvolvidas e assuntos abordados: [...] de seus elementos. ",PK
"Correlações baseadas [...] viscosidade cinemática do ar na temperatura do filme.",PK

```

Fonte: Autores, (2020).

Um exemplo do arquivo usado para teste é apresentado na Figura 7. Como pode ser observado, ao final de cada trecho, após a vírgula é inserido o sinal de interrogação, indicando que a categoria a ser classificada é desconhecida.

Figura 7. Formato do arquivo para o teste.

```
@relation teste

@attribute Documento string
@attribute classes {PK, PCK}

@data

"Na sequência da aula, fizemos uma miniexposição sobre [...] ao redor desse condutor.",?
"Nesse encontro abordamos inicialmente o conceito de [...] como mostra a Fig. 2.4.",?
"Discutimos tambem, por meio de apresentação teórica [...] a orientação desse campo.",?
"Nesse capítulo foi discutido o campo magnético ao [...] em outros dois capítulos.",?

```

Fonte: Autores, (2020).

O pré-processamento dos dados no Weka, foi definido usando o filtro *StringToWordVector* (vetor de palavras), usado para converter os atributos do tipo texto em um conjunto de atributos numéricos que representam as ocorrências das palavras contidas nos textos (Waikato, 2015, p.203). Esse filtro ainda fornece uma relação de parâmetros que podem ser configurados para fazer o tratamento dos dados na busca de melhorar o desempenho do classificador.

Dentre os parâmetros presentes no filtro foram configurados para o pré-processamento os parâmetros de tokenização, *stopwords* e as *IDFTransform* e *TFTransform*.

- TFTransform (Frequência do Termo) e *IDFTransform* (Frequência Inversa do Documento): definidos como *true* (verdade). Representam a frequência com que um termo aparece em um documento, quanto maior o TF e menor o IDF mais relevante ele será para determinado documento.
- *Stemming*: definido como *LovinsStemmer*. Reduz as palavras, retirando seu sufixo, por meio de determinadas regras que dependem do idioma, até que a mesma

fique com seu menor radical. Este processo tem como objetivo reduzir a quantidade de palavras diferentes no texto a serem tratadas (Ticom, 2007).

- *Tokenizer*: definido como *WordTokenizer*. É usado para dividir uma frase em palavras chamadas de *tokens* com base em um determinado delimitador.

O parâmetro *Stopwords* reduz a dimensão do documento, uma vez que retira do texto palavras que tem alta frequência e não acrescentam representatividade ou que sozinhas não tem significado, tais como preposições, pronomes, artigos. No entanto, esse recurso foi mantido como nulo, tendo em vista que nesse trabalho algumas dessas palavras como as preposições são importantes na identificação dos conhecimentos.

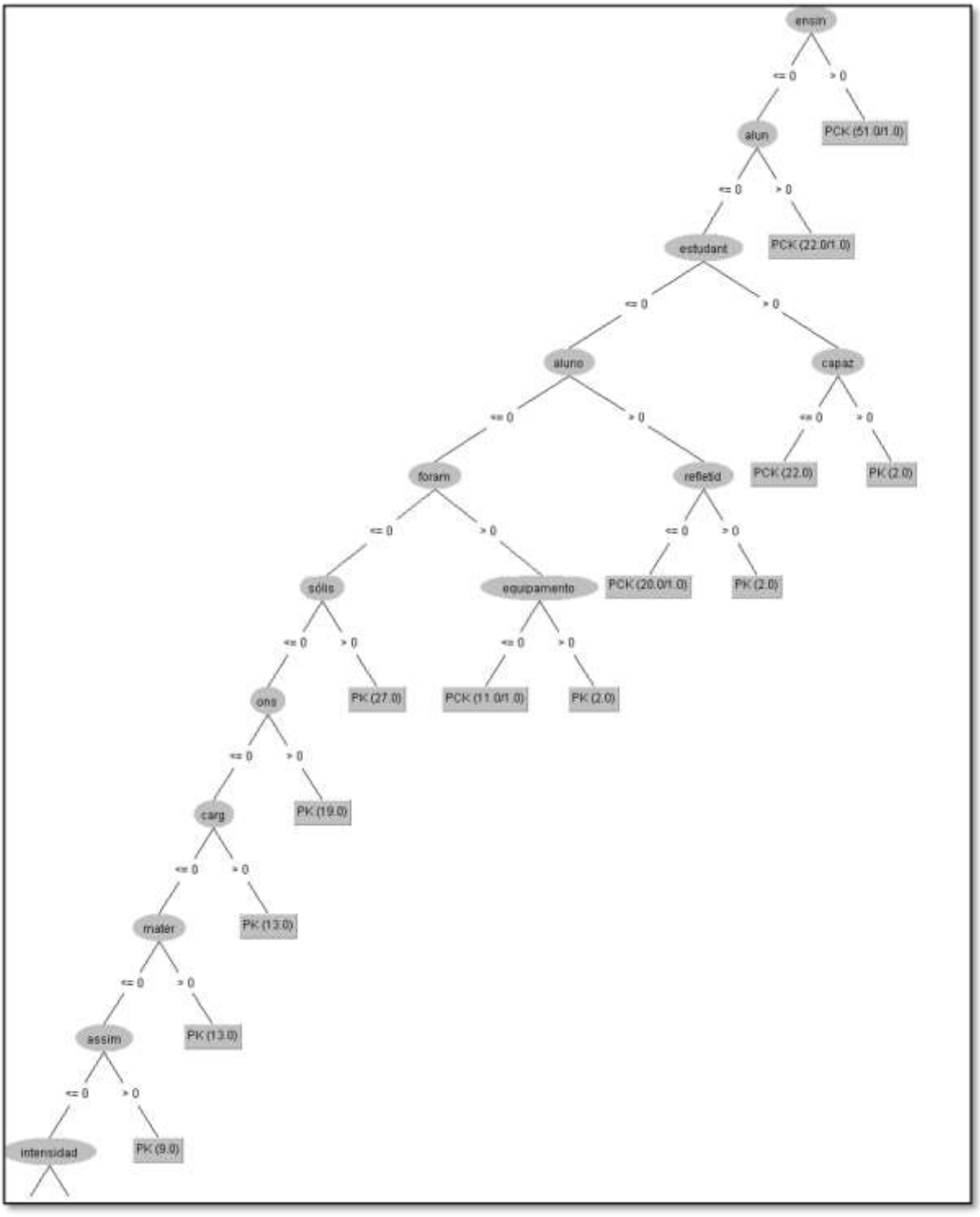
2.3.3 Extração de padrões

O modelo *doc2vec* faz uso de uma rede neural do tipo RNN e foi gerado a partir da arquitetura memória distribuída e saco de palavras distribuído com base na biblioteca disponível no *deeplearning4j*. O modelo apresenta tamanhos do vetores que variam conforme a quantidade de épocas definidas.

O modelo gerado pelo J48 constrói um modelo de árvore de decisão baseado no conjunto de dados de treinamento, que é utilizado para classificar as instâncias (trechos) do conjunto de teste. O algoritmo seleciona o atributo que mais eficientemente subdivide o conjunto das amostras em subconjuntos homogêneos, caracterizados por sua classe. O critério utilizado para a seleção consiste no ganho de informação obtida na escolha do atributo para subdivisão (Quinlan, 1993; Hall et al., 2009).

O tamanho da árvore foi 51 e o número de folhas geradas foram 26. Na Figura 8 é apresentada a árvore gerada pelo algoritmo J48. A grande quantidade de níveis e nós-folha se deve ao fato de não ter sido aplicado o filtro s*topwords* para reduz a dimensão do documento. Nesse trabalho algumas s*topwords* são relevantes para a análise textual, portanto a remoção desses termos pode retirar o sentido semântico de expressões.

Figura 8. Árvore de decisão – J48.



Fonte: Weka Classifier Tre Visualizer.

2.4 Validação e Testes

2.4.1 Validação cruzada

A validação cruzada é uma técnica muito empregada em tarefas de predição para avaliar a capacidade de generalização de um modelo gerado a partir de um conjunto de dados. Essa técnica busca estimar o quão preciso é o desempenho do modelo para um novo conjunto de dados (Santana, 2020).

Uma das maneiras de fazer a divisão desses dados é usando o método *cross validation,* que consiste em dividir os dados em 70-30 aleatoriamente e tem como principal objetivo evitar problemas de aleatoriedade, permitindo treinar e testar o modelo com todos os dados disponíveis para evitar a variância, com isso temos um resultado mais robusto (Santana, 2020).

Na validação cruzada *K-Fold*, o conjunto de dados inicial é dividido aleatoriamente em k subconjuntos ou *folds* ($D1$, $D2$, ..., Dk), de tamanho aproximadamente igual. Treinamento e teste são realizados k vezes, e para cada iteração i, o subconjunto Di é utilizado como teste, e os demais subconjuntos são utilizados para o treinamento do modelo (Han, et. al., 2011).

Algumas pesquisas tem apresentado as melhores estimativas da taxa de erro com a utilização de $k = 10$ (Witten & Frank; 2005).

2.4.2 Métricas de avaliação

A avaliação dos modelos de classificação pode ser realizada por meio de diferentes métricas que devem ser consideradas de acordo com as características dos dados. Um erro de classificação ocorre quando o valor predito pelo classificador é diferente da Classe real.

Para avaliar o desempenho do modelo do algoritmo J48 foi usado como referência a matriz de confusão, criada com base nas classes estabelecidas para a predição. Em tarefas de classificação, uma matriz de confusão é uma tabela que permite a visualização do desempenho do algoritmo. As linhas são representadas pelas classes verdadeiras, enquanto as colunas pela predição do algoritmo.

A matriz de confusão gerada pelo modelo foi adaptada, contemplando as classes do domínio do Conhecimento da Física (PK) e do domínio do Conhecimento Didático do Conteúdo (PCK), como apresentados na Figura 9.

Figura 9. Matriz de confusão do contexto da pesquisa.

PREDIÇÃO
O que o modelo prediz

REAL Se é PK ou PCK		PK	PCK
	PK	**V_{PK}**	**F_{PCK}**
	PCK	**F_{PK}**	**V_{PCK}**

Fonte: Autores, (2020).

Onde:

V_{PK}: Verdadeiro PK;

F_{PK}: Falso PK;

V_{PCK}: Verdadeiro PCK;

F_{PCK}: Falso PCK.

O desempenho geral de um modelo de predição pode ser calculado através da sua exatidão (também conhecida como acurácia geral) que é medida pela quantidade de acertos de classificação cometidos dividido pelo número total de casos na amostra utilizada para o teste (Ye, 2003).

- **Acurácia (Taxa de acerto)**: Mede o desempenho médio do classificador (Equação 1). Onde, N é a quantidade total de amostras ($V_{PK} + V_{PCK} + F_{PK} + F_{PCK}$).

$$Acurácia = \frac{V_{PK} + V_{PCK}}{N} \quad (1)$$

A Taxa de Verdadeiros V_{PK} representa a quantidade de conhecimentos classificados corretamente como PK (V_{PK}) dividido pela quantidade de conhecimentos classificados corretamente como PK (V_{PK}) mais a quantidade de conhecimentos classificados incorretamente como PCK (F_{PCK}), como apresentado na equação 2.

- **Taxa de Verdadeiros PK:**

$$Taxa\ V_{PK} = \frac{V_{PK}}{V_{PK} + F_{PCK}} \quad (2)$$

De forma similar, pode-se calcular a Taxa de Verdadeiros V_{PCK} que é apresentada na equação 3.

- **Taxa de Verdadeiros PCK:**

$$Taxa\ V_{PCK} = \frac{V_{PCK}}{V_{PCK} + F_{PK}} \quad (3)$$

A Taxa de Falsos F_{PK} representa a quantidade de conhecimentos classificados incorretamente como PK (F_{PK}) dividido pela quantidade de conhecimentos classificados incorretamente como PK (F_{PK}) mais a quantidade de conhecimentos classificados corretamente como PCK (V_{PCK}), como apresentado na equação 4.

- **Taxa de Falsos PK:**

$$Taxa\ F_{PK} = \frac{F_{PK}}{F_{PK} + V_{PCK}} \quad (4)$$

De forma similar, é calculada a Taxa de Falsos (F_{PCK}), como apresentada na equação 5.

- **Taxa de Falsos PCK:**

$$Taxa\ F_{PCK} = \frac{F_{PVK}}{F_{PVK} + V_{PK}} \quad (5)$$

- **Precisão**: mede a porcentagem de acertos entre as observações classificadas como positivas (Equação 6).

$$Precisão = \frac{V_{PK}}{V_{PK} + F_{PK}}$$

(6)

- **Sensibilidade (*Recall*):** também chamado de taxa de verdadeiro positivo, mede a porcentagem das observações positivas que foram corretamente classificadas (Equação 7).

$$Sensibilidade = \frac{V_{PK}}{V_{PK} + F_{PCK}}$$

(7)

F1 *score*: essa medida busca um equilíbrio entre a sensibilidade e a precisão (Equação 8).

$$F1 - score = \frac{2 * Precisão * Sensibilidade}{Precisão * Sensibilidade}$$

(8)

2.4.3 Descrição dos testes de validação

Os testes de validação foram realizados com os algoritmos de classificação *doc2vec* e J48. A realização dos testes seguiu as seguintes etapas:

1. Divisão do conjunto de treinamento baseado no método *cross validation K-Fold.*
2. Apresentação dos resultados alcançados por meio das métricas apresentadas na seção anterior.

Os resultados obtidos com o doc2vec se darão pela medida da acurácia e os resultados do J48 serão apresentados conforme as métricas de desempenho apresentadas na seção anterior.

Para atingir os objetivos do estudo de forma eficaz, nesse capítulo foram descritas todas as etapas do percurso metodológico para extrair o conhecimento no estudo desenvolvido, essencial para garantir o alcance do que foi delineado no trabalho.

3. Resultados e Discussão

Neste capítulo são apresentados os resultados dos testes de validação realizados com os algoritmos *doc2vec* e o J48. O conjunto de dados utilizado para o treinamento e testes foram apresentados na Seção 3.3.1, contendo 318 conhecimentos de professores de Física, identificados em trechos de PaP-eRs. A avaliação dos resultados, consistiu da análise estatística do desempenho dos classificadores através dos índices obtidos nas métricas apresentadas na Seção 3.4.2.

3.1 Doc2Vec

Os testes com *doc2vec* foram executados no *InteliJ IDEA*. Os parâmetros da rede neural configurados para os testes foram a taxa de aprendizado, a taxa mínima de aprendizado e as épocas. O conjunto total de dados foi dividido aleatoriamente em 3 subconjuntos denominados de subconjunto 1, subconjunto 2 e subconjunto 3, seguindo o critério de validação cruzada (70 – 30), sendo 70% para treino e 30% para testes, conforme apresentado na Seção 3.4.1. Assim, a cada interação, um novo conjunto de dados é formado para treino e testes. Em cada subconjunto foram realizados 5 testes, com valores diferentes atribuídos aos parâmetros da rede neural para observar o desempenho do classificador em cada situação. Os valores dos parâmetros definidos para cada teste são apresentados na Tabela 2.

Tabela 1. Definição dos parâmetros no doc2vec.

Parâmetros	Teste 1	Teste 2	Teste 3	Teste 4	Teste 5
Taxa de aprendizado	0.025	0.025	0.025	0.05	0.025
Taxa mínima de aprendizado	0.001	0.001	0.01	0.001	0.001
Épocas	5.000	20.000	5.000	20.000	40.000

Fonte: Autores, (2020).

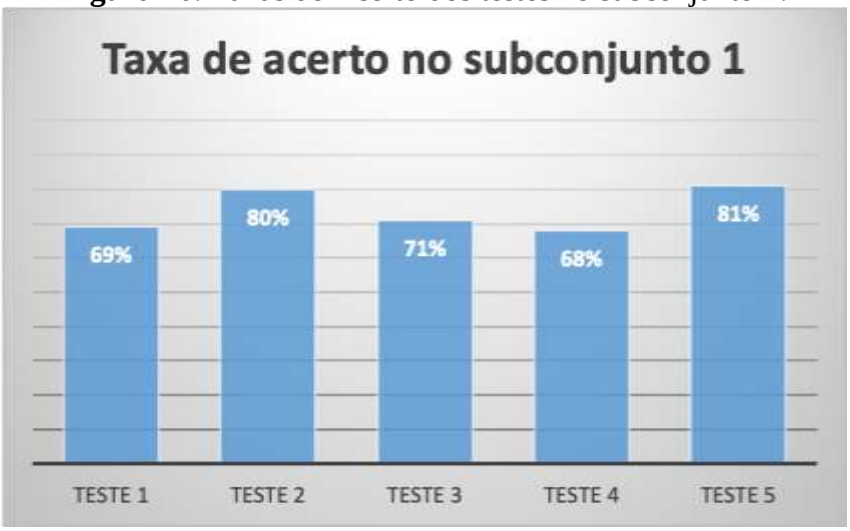
3.1.1 Validação com o subconjunto 1

No teste 1, as configurações definidas nos parâmetros: taxa de aprendizagem = 0.025, taxa mínima de aprendizagem = 0.001 e épocas = 5, alcançaram um resultado de 69% de taxa de acerto. Aumentando o número de épocas para 20 e mantendo os outros parâmetros, a taxa de acerto aumentou para 80%, teste 2. No teste 3, a taxa mínima de aprendizagem foi alterada

para 10 vezes o valor utilizado no teste 1 e a taxa de acerto não melhorou muito, atingindo 71%.

Observou-se após os 5 testes que quanto maior o número de épocas, melhor é a taxa de acerto, como pode ser visto no teste 5 (Figura 10). Entretanto, no teste 4, mesmo quadruplicando o número de épocas a taxa de acerto atingiu o pior resultado dentre os testes do subconjunto 1. Nesse teste a taxa de aprendizagem foi duplicada demonstrado que o aumento nessa taxa pode levar a um conjunto sub-ótimo, como foi comentado na seção 3.3.2. Todas as taxas de acerto dos testes realizados são apresentados no gráfico da Figura 10.

Figura 10. Taxas de Acerto dos testes no subconjunto 1.

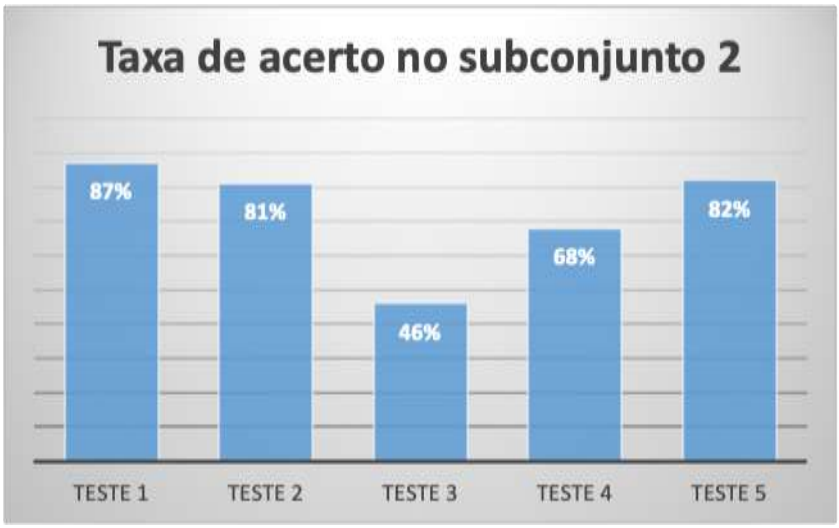


Fonte: Autores, (2020).

3.1.2 Validação com o subconjunto 2

Os resultados da validação com o subconjunto 2 seguiram a mesma tendência do subconjunto 1, com exceção o teste 3.

Figura 11. Taxas de Acerto dos testes no subconjunto 2.



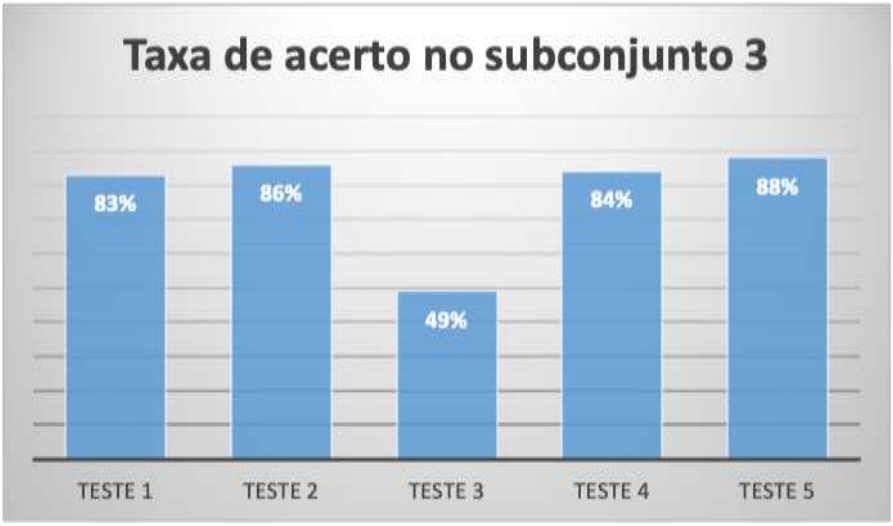
Fonte: Autores, (2020).

A conclusão desses testes é que a taxa mínima de aprendizagem não pode ser muito alta. Nesse teste, a taxa mínima de aprendizado foi aumentada 10 vezes e o resultado da taxa de acerto caiu para 46%. No gráfico da Figura 11 são apresentados os resultados alcançados em cada teste.

3.1.3 Validação com o subconjunto 3

Nesses testes foi observado a mesma tendência dos resultados alcançados nos testes anteriores. Os resultados dos testes realizados no subconjunto 3 são apresentados na forma de gráfico na Figura 12.

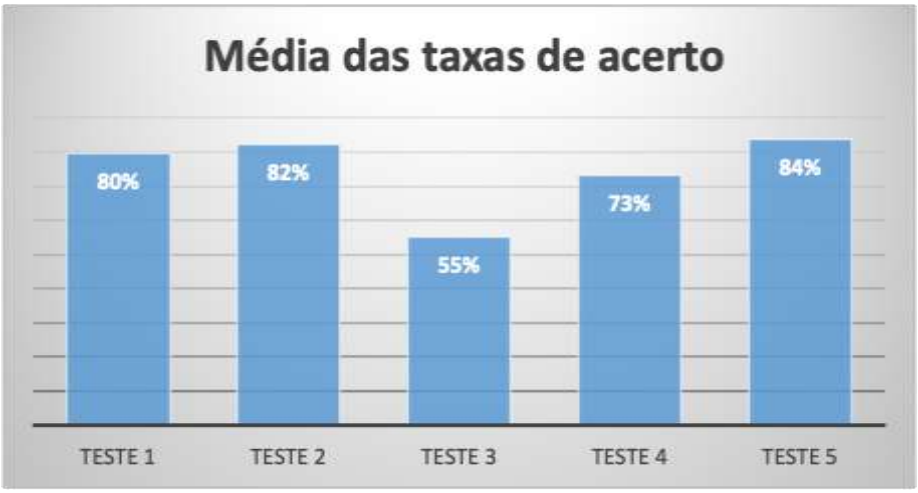
Figura 12. Taxas de Acerto dos testes no subconjunto 3.



Fonte: Autores, (2020).

Dentre os testes realizados nos três subconjuntos o quinto teste foi o que apresentou a melhor taxa de acerto, como apresentado no gráfico da Figura 13. Para esse conjunto de dados, manteve-se as taxas a taxa de aprendizado e taxa mínima de aprendizado iguais ao ajustado para o teste 1 e o número de épocas foi aumentado a oito vezes. É claro que se for realizado muitos outros teste e analisando a correlação entre a variação da taxa de acerto em relação às variações dos parâmetros de entrada (taxas e número de épocas) no teste pode-se chegar aos valores ótimos mas isso não é o objetivo desse trabalho.

Figura 13. Média das taxas de acerto do três subconjuntos.



Fonte: Autores, (2020).

3.2 J48

Os testes com o algoritmo J48 foram realizados no Weka usando a técnica *Cross Validation* com *K-Fold = 3 e K-Fold =10*. Os parâmetros utilizados para a extração dos conhecimentos foram os mesmos apresentados na Seção 3.3.2.

3.2.1 Validação com K-Fold = 3

Na Figura 14 é apresentada a matriz de confusão gerada pelo modelo contendo os valores reais e preditos.

Figura 14. Matriz de confusão *K*-Fold = 3.

PREDIÇÃO
O que o modelo prediz

REAL (Se é PK ou PCK)		PK	PCK
	PK	**V_{PK} = 101**	**F_{PCK}= 46**
	PCK	**F_{PK} = 31**	**V_{PCK} = 140**

Fonte: Autores, (2020).

Os resultados da matriz de confusão são utilizados na avaliação do desempenho do classificador. Nesse teste a acurácia foi de 75%, conforme a equação (1).

$$\boldsymbol{Acurácia = \frac{101 + 140}{318} = 0{,}757}$$

A taxa de verdadeiro V_{PK} foi de 68%, conforme a equação (2).

$$Taxa\ V_{PK} = \frac{101}{101 + 46} = 0{,}687$$

A taxa de falso F_{PK} foi de 18%, conforme a equação (4).

$$Taxa\ F_{PK} = \frac{31}{31 + 140} = \ 0{,}181$$

O modelo alcançou uma precisão de 76%, calculado conforme a equação (6).

$$Precisão = \frac{101}{101 + 31} = 0{,}765$$

A sensibilidade alcançada foi de 68%, calculado conforme a equação (7).

$$Sensibilidade = \frac{101}{101 + 46} = 0{,}687$$

O F1-score (equação 8) que atribui o mesmo grau de importância para as métricas de Precisão e Sensibilidade foi de 72%.

$$F1 - score = \frac{2 * (0{,}765 * 0{,}687)}{0{,}765 + 0{,}687} = 0{,}724$$

De forma similar, foram realizados os cálculos com os valores da classe do domínio do Conhecimento Didático do Conteúdo PCK. Os resultados obtidos com esse modelo para as classes do domínio do Conhecimento da Física PK e do domínio do Conhecimento Didático do Conteúdo PCK são apresentados na Tabela 3.

Tabela 3. Resultados da validação com *K*-Fold = 3.

Taxa V	Taxa F	Precisão	Sensibilidade	F1- score	Classe
0,687	0,181	0,765	0,687	0,724	**PK**
0,819	0,313	0,753	0,819	0,784	**PCK**
0,758	0,252	0,758	0,758	0,756	**Média**

Fonte: Autores, (2020).

3.2.2 Validação com K-Fold = 10

O segundo teste foi realizado com o *k-Fold* = 10, nesse caso a acurácia foi de 76%. Na Figura 15 é apresentada a matriz de confusão gerada com esse modelo e na Tabela 4 são apresentados os resultados obtidos.

Figura 15. Matriz de confusão *K-Fold* = 3.

PREDIÇÃO
O que o modelo prediz

REAL: Se é PK ou PCK	PK	PCK
PK	**V_{PK} = 108**	**F_{PCK} = 39**
PCK	**F_{PK} = 36**	**V_{PCK} = 135**

Fonte: Autores, (2020).

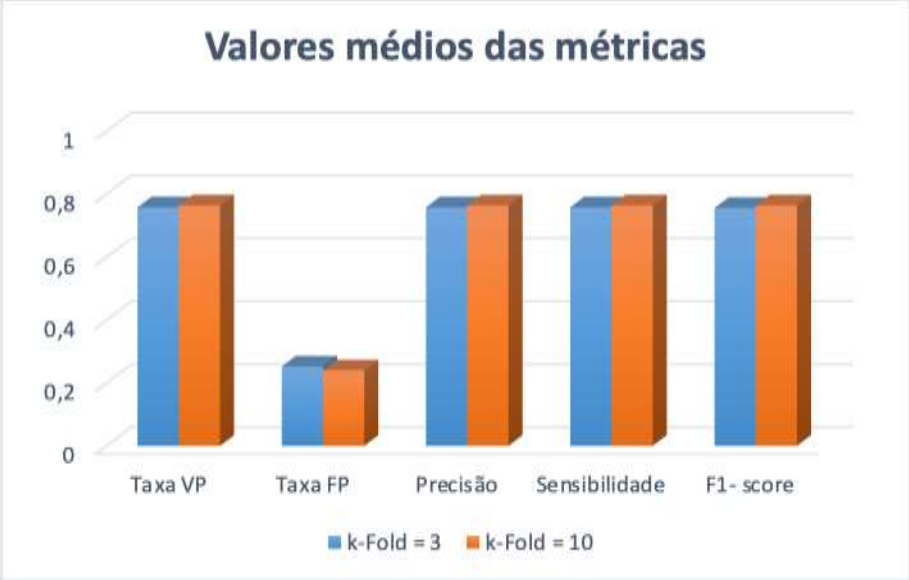
Tabela 4. Resultados da validação com *K-Fold* = 10.

Taxa VP	Taxa FP	Precisão	Sensibilidade	F1- score	Classe
0,735	0,211	0,750	0,735	0,742	**PK**
0,789	0,265	0,776	0,789	0,783	**PCK**
0,764	0,240	0,764	0,764	0,764	**Média**

Fonte: Autores, (2020).

A partir dos resultados apresentados nos dois testes, observa-se que os modelos atingiram uma acurácia com índices muito próximos, o primeiro atingiu 75% e o segundo 76%. Nesse sentido, embora os autores Witten e Frank (2005) indiquem que a utilização de *K-Fold* =10 apresente melhores resultados, nesse trabalho o *K-Fold* = 3 e o *K-Fold* = 10 apresentaram resultados muito próximos, como pode ser observado nas Tabelas 3 e 4 e no gráfico da Figura 16.

Figura 16. Valores médios das métricas.



Fonte: Autores, (2020).

Fazendo um comparativo entre os testes realizados com o *doc2vec* e o J48, constata-se que os melhores resultados foram alcançados com o *doc2vec*, com uma taxa média de acerto de 84% enquanto que o J48 apresentou 75%. Entretanto, a usabilidade do J48 com o Weka é muito melhor do que o doc2vec no ambiente do *InteliJ IDEA*.

4. Considerações Finais

Nesse trabalho foi realizado a análise da eficácia de dois algoritmos de aprendizado de máquina, *doc2vec* e J48 na classificação automática de conhecimentos especializados de professores de Física. Para isso foi realizado um estudo de caso em uma perspectiva exploratória e aplicada, que contou com a utilização de técnicas de mineração de textos e de processamento de linguagem natural para prever a classe que caracterize um Conhecimento da Física (PK) ou Conhecimento Didático do Conteúdo (PCK).

Os testes realizados com o algoritmo doc2vec foram desenvolvidos no InteliJ IDEA e os teste com o J48 foram desenvolvidos no Weka. No primeiro caso o ambiente permite uma maior flexibilidade para modificar os parâmetros, entretanto, exige mais codificação. Os testes realizados com o algoritmo J48 foram desenvolvidos no Weka. Esse ambiente simplificou o trabalho e permitiu a utilização de filtros de pré-processamento. Ambos os testes apresentaram bons resultados. No *doc2vec* a taxa média de acerto foi de 84% enquanto que no J48 foi de 75%.

Além da taxa de acerto, também foram calculados a precisão e a sensibilidade de cada modelo que tiveram, respectivamente, os resultados médios de 75% e 76%.

Com base nos resultados atingidos, pode-se concluir que a estratégia de usar inteligência artificial para a classificação automática de conhecimentos de professores de Física é uma solução plausível. É importante ressaltar que a base de dados utilizada no treinamento continha apenas 318 trechos, o que é considerado pequena para treinamento em aprendizado de máquina.

Esse trabalho é válido e traz benefícios para os pesquisadores da linha de investigação em Conhecimento Especializado de Professores de Física na medida em que a utilização dos métodos aqui apresentados possibilitam a redução do tempo gasto de classificação de conhecimentos.

Ainda que as técnicas tenham sido aplicadas aos conhecimentos de professores de Física, com os resultados obtidos, pode-se concluir que a mesma abordagem pode ser aplicada aos outros modelos teóricos de conhecimento especializado de professores.

Por fim, verificou-se que foi possível atingir o objetivo proposto, utilizando-se aprendizado de máquina para a classificação de documentos em língua portuguesa em classes pré-definidas, mesmo em condições onde as classes apresentam semelhanças e a quantidade de documentos não seja muito grande.

Sugere-se que em trabalhos futuros sejam realizados novos testes com outros classificadores, verificando-se a matriz de confusão resultante e eventuais melhorias nos resultados pela combinação de resultados de cada classificador. O uso de técnicas de extração de informações e a classificação com redes neurais pode apresentar resultados promissores para a área.

Referências

Almeida, L. M., Padilha, T. P. P., Oliveira, F. L.de & Previero, C. A. (2003). *Uma Ferramenta para Extração de Padrões*. REIC. Revista Eletrônica de Iniciação Científica.

Ausubel, D. P., Novak, J. D. & Hanesian, H. (1980). *Psicologia educacional*. Tradução Eva Nick. Rio de Janeiro: Interamericana.

Ball, D. L., Thames, M. H. & Phelps, G. (2008). *Content Knowledge for Teaching: What Makes It Special?* Journal of teacher education.

Breve, F. A. (2010). *Aprendizado de máquina em redes complexas*. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

Carrillo, J., Avila, D. I. E., Mora, D. V. & Medrano, E. F. (2014). *Un marco teórico para el conocimiento especializado del profesor de matemáticas*. Huelva, Espanha: Universidad de Huelva Publicaciones.

Carrillo, J., Climent, N., Contreras, L. C. & Muñoz-Catalán, M. C. (2013). *Determining Specialised Knowledge For Mathematics Teaching*. In: Ubuz, B.; Haser, C., et al. (Ed.). VIII Congress of the European Society for Research in Mathematics Education (CERME 8). 8. Antalya, Turkey: Middle East Technical University, Ankara.

Contreras, J. (2002). *A autonomia de professores*. São Paulo: Cortez.

Feldman, R., & Sanger, J. (2006). *Text Mining Handbook*. Cambridge (MA): Cambridge University Press.

Fernandez, C. (2015). *Revisitando a Base de Conhecimentos e o Conhecimento Pedagógico do Conteúdo (PCK) de Professores de Ciências*. Ensaio: Pesquisa em Educação em Ciências.
Gauthier, C. (1998). *Por uma teoria da Pedagogia: pesquisas contemporâneas sobre o saber docente*. Ijuí: Unijuí.

Giasson, E., Hartemink, A. E., Tornquist, C. G., Teske, R., & Bagatini, T. (2013) *Avaliação de cinco algoritmos de árvores de decisão e três tipos de modelos digitais de elevação para mapeamento digital de solos a nível semidetalhado na Bacia do Lageado Grande, RS, Brasil*. Ciência Rural (UFSM).

Gil, A. C. (1999). *Métodos e técnicas de pesquisa social*. (5a ed.), São Paulo.

Grossman, P. L. (1990). *The making of a teacher: teacher knowledge and teacher education.* New York: Teachers College Press.

Hall, M. (2009). *The WEKA Data Mining Software: an update*. SIGKDD Explorations Newsletter.

Han, J., Kamber, M. & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3a ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

He, K., Zhang, X., Ren, S. & Sun, J. (2015). *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*. In: Proceeding ICCV '15 Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV).

Kilpatrick, J., & Spangler, D. A. (2015) *Educating Future Mathematics Education Professors*. Handbook of International Research in Mathematics Education.

Lima, S. S. (2018). *Conhecimento especializado de professores de física: uma proposta de modelo teórico.* Dissertação (Mestrado em Ensino) - Programa de Pós-Graduação Stricto Sensu em Ensino, Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso – IFMT, Cuiabá.

Lima, S. S., Darsie, M. M. P., & Mello, G. J. (2020). *Análise comparativa dos modelos usados como ferramenta metodológica nas pesquisas sobre o Conhecimento Pedagógico de Conteúdo (PCK) de professores de Física no Brasil*. Caderno Brasileiro de Ensino de Física.

Loughran, J., Milroy, P., Berry, A., Gunstone, R. & Mulhall, P. (2001) *Documenting science teachers' pedagogical content knowledge through PaP-eRs*. Research in Science Education.

Luís, M., Monteiro, R. & Carrillo, J. (2015). *Conhecimento Especializado do Professor para Ensinar Ciências*. In: Encontro Nacional De Educação Em Ciências, XVI. Lisboa, Portugal. Anais. Lisboa: APEduC.

Maimon, O., & Rokach, L. (2005). *Data mining and knowledge discovery handbook.* Springer.

Malanchen, J., Negrão, R. & Santos, S. A. (2012). *Formação de professores: Diferentes enfoques e algumas contradições.* In: IX ANped Sul, 2012, Caxias do Sul. A pós-graduação e suas interlocuções com a educação básica.

Marconi, M. De A. & Lakatos, E. M. (2001). *Metodologia científica: ciência e conhecimento científico, métodos científicos, teoria, hipóteses e variáveis, metodologia jurídica*. 3. ed. rev. e ampl. São Paulo.

Morais, E. A. M. & Ambrósio, A. P. L. (2007) *Mineração de textos*. Relatório Técnico–Instituto de Informática (UFG).

Moreira, M. A. (2017) *Grandes desafios para o ensino da física na educação contemporânea*. Revista do Professor de Física, Brasília.

Moriel Junior, J. G. (2014). *Conhecimento especializado para ensinar divisão de frações.* 2014. 162 p. Tese de doutorado (Pós-Graduação em Educação em Ciências e Matemática – PPGECEM/REAMEC) – Universidade Federal de Mato Grosso, Cuiabá.

Moriel Junior, J. G., & Alencar, E. S. (2020). *Pesquisa e formação docente com MTSK em Mato Grosso e Mato Grosso do Sul*. Research, Society and Development.

Moriel Junior, J. G., & Wielewski, G. D. (2017). *Base de Conhecimento de Professores de Matemática: do Genérico ao Especializado*. Revista de Ensino, Educação e Ciências Humanas.

Nóvoa, A. (1991). *Para o estudo sócio-histórico da gênese e desenvolvimento da profissão docente*. Teoria & Educação.

Park, S., & Oliver, S. (2008). *Revisiting the conceptualization of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals*. Research in Science Education. NewYork.

Pellucci, P. R. S., Ribeiro, R. P., Oliveira, W. B. & Ladeira, A. P. (2011) *Utilização de Técnicas de Aprendizado de Máquina no reconhecimento de entidades nomeadas no Português*. Exacta, Belo Horizonte.

Pimenta, S. G. (2012). *Saberes pedagógicos e atividade docente*. São Paulo: Cortez.

Pimenta, S. G. (1997). *Formação de professores - saberes da docência e identidade do professor*. Revista da Educação da Aec do Brasil, São Paulo.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Sydney, Austrália: Morgan Kaufmann Publishers.

Richardson, R. J. (1999) *Pesquisa social: métodos e técnicas*. (3a ed.), São Paulo.

Salem, S. (2012). *Perfil, evolução e perspectivas da Pesquisa em Ensino de Física no Brasil*. 2012. 385f. Tese (Doutorado) – Universidade de São Paulo, São Paulo.

Santos, R. M. M. dos. (2016). *Técnicas de aprendizagem de máquina utilizadas na previsão de desempenho acadêmico*. 2016. 90 p. Dissertação (Mestrado Profissional) – Programa de Pós-Graduação em Educação, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina.

Shulman, L. (1987). *Knowledge and teaching: Foundations of the new reform*. Harvard Educational Review. Feb.

Shulman, L. (1986). *Those who understand: knowledge growth in teaching*. Educational Researcher. Washington.

Soares, S. T. C. (2019). *Conhecimento Especializado de Professores de Química: Proposta de Modelo com detalhamento do Conhecimento dos Tópicos*. Dissertação (Mestrado em Ensino) Programa de Pós-Graduação Stricto Sensu em Ensino, Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso – IFMT, em associação com Universidade de Cuiabá.

Tardif, M. (2010). *Saberes docentes e formação profissional*. (12a ed.), Petrópolis (RJ): Vozes.

Vasco, D., Moriel Junior, J. G. & Contreras, L. C. (2017). *Subdominios KoT y KSM del Mathematics Teacher's Specialised Knowledge (MTSK): definición, categorías y ejemplos*. In: III Jornadas de Investigación en Didáctica de las Matemáticas, Huelva.

Vergara, S. C. (2000). *Projetos e relatórios de pesquisa em administração*. (3a ed.), Rio de Janeiro: Atlas.

Waikato. (2020, junho 16). *Weka 3: Data Mining Software in Java*. 2015. Recuperado de: http://www.cs.waikato.ac.nz/ml/weka/.

Witten, I. H., & Frank. E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.

Ye, N. (2003). *The Handbook of Data Mining*. London: Taylor & Francis.

Zikmund, W. G. (2000). *Business research methods*. (5a ed.) Fort Worth, TX: Dryden.

Porcentagem de contribuição de cada autor no manuscrito

Tamara Aguiar Tavares Mascarenhas – 40%

Jeferson Gomes Moriel Junior – 20%

Raphael de Souza Rosa Gomes – 20%

Geison Jader Mello – 20%