**A POSIÇÃO DA CIÊNCIA DA INFORMAÇÃO NO DESENVOLVIMENTO DE MODELOS DE MACHINE LEARNING**

**Fabrício Amadeu Gualdani-** Doutorando no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho – UNESP, campus de Marília, Brasil.

E-mail: [fabricio.gualdani@unesp.br](mailto:fabricio.gualdani@unesp.br).

<https://orcid.org/0000-0001-7426-0831>.

**Allan Ferreira -** Mestrando no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho – UNESP, campus de Marília, Brasil.

E-mail: [allan.ferreira1983@unesp.br](mailto:allan.ferreira1983@unesp.br)

**Marcos Antonio Siqueira -** Doutorando no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho – UNESP, campus de Marília, Brasil.

E-mail: [marcos.a.siqeuira@unesp.br](mailto:marcos.a.siqeuira@unesp.br)

**Resumo:** Para refletir sobre a posição da ciência da informação no desenvolvimento de modelos de machine learning, A Federação Internacional de Associações e Instituições Bibliotecárias destacou a importância da ciência da informação buscar uma compreensão de como os modelos de machine learning funcionam. O objetivo deste estudo é gerar uma contribuição para a orientação do IFLA quando se propõe a refletir na posição da ciência da informação no desenvolvimento de modelos de machine learning. A metodologia consiste em uma pesquisa exploratória, na qual foi feito um levantamento da grade curricular de quatro universidades públicas brasileiras que oferecem o curso de biblioteconomia e foi feita uma associação entre o conteúdo abordado em determinadas disciplinas pertencentes a grade destas universidades para a formação do profissional da informação com os processos de construção de um modelo de machine learning. Os resultados obtidos demonstraram cinco etapas relacionadas ao processo de desenvolvimento de um modelo de machine learning: definição do problema e dimensionalidade com as disciplinas de serviços de referência e fontes de informação; preparação dos dados com classificação e organização da informação; treinamento do modelo com análise e representação temática da informação e teste, avaliação e aperfeiçoamento do modelo com gestão da informação e do conhecimento. Conclui-se que ao analisar as competências relacionadas à formação do profissional da informação é possível inserir este profissional em todas as etapas de desenvolvimento de um modelo de machine learning.

**Palavras-chave:** Ciência da Informação; Literacia Digital; Machine Learning; Biblioteconomia; Atuação.

**Abstract:** To reflect on the position of information science in the development of machine learning models, The International Federation of Library Associations and Institutions has highlighted the importance of information science to seek an understanding of how machine learning models work. The purpose of this study is to generate a contribution to IFLA's guidance when it sets out to reflect on the position of information science in the development of machine learning models. The methodology consists of an exploratory research, a survey was made of the curricula of four public Brazilian universities that offer the course of librarianship and an association was made between the content covered in certain disciplines belonging to the grid of these universities for the formation of the information professional with the processes of building a machine learning model. The results obtained showed five steps related to the process of developing a machine learning model: problem definition and dimensionality with the disciplines of reference services and information sources; data preparation with classification and information organization; model training with analysis and thematic representation of the information; and model testing, evaluation, and improvement with information and knowledge management. It is concluded that by analyzing the competencies related to the training of the information professional, it is possible to insert this professional in all the stages of developing a machine learning model.

**Key-words:** Information Science; Digital Literacy; Machine Learning; Librarianship; Performance.

**1 INTRODUÇÃO**

Para refletir sobre a posição da ciência da informação no desenvolvimento de modelos de machine learning é necessário inicialmente uma compreensão a respeito do que se trata a ciência da informação assim como uma compreensão geral de como esses modelos são desenvolvidos. Inicialmente, a ciência da informação possui como objeto de estudo a informação em si registrada em seus diferentes suportes, assim como o desenvolvimento e a aplicação de protocolos para a seleção, representação, organização, recuperação, disseminação e compreensão da informação, bem como a proposição e avaliação de serviços e produtos informacionais (GUALDANI; GALVÃO, 2020).

Estudos clássicos da área pertencentes a autores como Saracevic (1995) definem a ciência da informação como um campo dedicado às questões científicas e à prática profissional voltadas para os problemas da efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, institucional ou individual do uso e das necessidades de informação. No tratamento dessas questões são consideradas de particular interesse as vantagens das modernas tecnologias informacionais. Já em Borko (1968) nos são apresentadas considerações que intrinsecamente a ciência da informação existe um diálogo, uma comunicação com outras ciências, o que sugere a autores mais modernos como Le Coadic (1996) uma natureza interdisciplinar pertencente à área. Autores ainda mais contemporâneos como Sanchez et al (2022) indicam que a ciência da informação pode ser definida como uma ciência colaborativa auxiliando na tomada de decisão e gerenciamento informacional de outras áreas, inclusive em um cenário tecnológico e digital, já que a informação se encontra presente em todos os campos do saber.

Já os modelos de machine learning consistem em uma subárea do campo da inteligência artificial que permite que as máquinas desenvolvam a capacidade de aprendizado visando tarefas específicas, como por exemplo, realizar o diagnóstico de uma doença (SANTOS, et al., 2019). Esse exemplo destaca o potencial de aprendizado de modelos de machine learning e os impactos sociais que a inteligência artificial já está causando na sociedade.

A Federação Internacional de Associações e Instituições Bibliotecárias, IFLA em inglês, em sua declaração sobre bibliotecas e inteligência artificial, destacou a importância da ciência da informação buscar uma compreensão de como os modelos de machine learning funcionam.

Nesta declaração o IFLA orienta que a chamada ‘alfabetização em inteligência artificial’ pode começar com uma busca por entender o básico sobre como a inteligência artificial e machine learning funcionam, o que eles podem e não podem fazer (IFLA, 2020). Feitas estas considerações, o objetivo deste estudo consiste em gerar uma contribuição para a orientação do IFLA quando se propõe a refletir na posição da ciência da informação no desenvolvimento de modelos de machine learning. Esta reflexão será feita por meio de um processo de literacia digital, isto é, uma espécie de alfabetização e explicação dos conceitos e fundamentos relacionados a elaboração de modelos de machine learning, fundamentos estes originalmente pertencentes à área da ciência da computação sendo transmitidos para a ciência da informação assim como será realizada uma associação sobre como cada etapa relacionada à elaboração de um modelo de machine learning comunica-se com competências pertencentes a formação do bibliotecário e do profissional da informação.

A realização deste estudo justifica-se pela relevância de apresentar aos profissionais da informação e bibliotecários a respeito de como que competências pertencentes a sua formação comunicam-se com os processos de desenvolvimento de modelos de machine learning. Desta forma, gerando maior valor para a área de dados e computação por meio do auxílio da ciência da informação que por si só possui uma natureza colaborativa.

**2 METODOLOGIA**

Os procedimentos metodológicos deste estudo, do ponto de vista de sua natureza, consistem em uma pesquisa básica pois objetiva “gerar conhecimentos novos úteis para o avanço da ciência sem aplicação prática prevista” (PRODANOV; FREITAS, 2013, p.51). Do ponto de vista dos seus objetivos trata-se de uma pesquisa exploratória pois possui como finalidade proporcionar mais informações sobre o assunto a ser investigado, possibilitando a sua definição e o seu delineamento. Quanto aos procedimentos técnicos empregados, este estudo compõem-se de uma pesquisa bibliográfica.

Para o desenvolvimento deste estudo, foram consultadas bases de dados pertencentes às áreas da ciência da informação e ciência da computação. Iniciou-se uma busca em julho de 2022 da qual a primeira base de dados acessada consistiu na BRAPCI empregando o termo \*machine learning\* com uma delimitação de tempo de 2015 até 2022 da qual foram retornados 32 artigos. Sendo que de fato, 11 se encaixaram com a proposta desta pesquisa. Em seguida, acessou-se a base de dados Scielo da qual empregaram-se os termos \* machine learning \*ciência da informação \* sendo que foram retornados 4 artigos, no entanto nenhum se encaixou com a proposta deste estudo. O objeto de análise científica considerado neste estudo, compõem-se de artigos científicos.

Após a realização do processo de revisão de literatura, foi feito um levantamento da grade curricular de quatro universidades públicas brasileiras que oferecem o curso de biblioteconomia: a Universidade de São Paulo de Ribeirão Preto (USP); a Universidade Estadual Paulista Júlio de Mesquita Filho de Marília (UNESP); a Universidade Federal de São Carlos (UFSCAR) e a Universidade Federal do Rio Grande do Sul (UFRGS) e foi feita uma associação entre o conteúdo abordado em determinadas disciplinas pertencentes a grade destas universidades para a formação do profissional da informação, o bibliotecário com os processos de construção de um modelo de machine learning.

No que diz respeito às limitações deste estudo, esta pesquisa se concentrou apenas em estudos publicados em periódicos brasileiros e escritos em língua portuguesa, desconsiderando estudos internacionais.

**3 AS COMPETÊNCIAS DE UM PROFISSIONAL DA INFORMAÇÃO**

Para apresentar algumas das principais competências relacionadas à formação do bibliotecário e do profissional da informação, foram consultadas as grades curriculares do curso de biblioteconomia no site oficial de quatro universidades públicas brasileiras: a Universidade de São Paulo de Ribeirão Preto a (USP); a Universidade Estadual Paulista Júlio de Mesquita Filho de Marília (UNESP); a Universidade Federal de São Carlos (UFSCAR) e a Universidade Federal do Rio Grande do Sul (UFRGS) sendo que foram identificadas quatro principais competências que apesar das variações das titulações das disciplinas, referem-se à mesma fundamentação: serviços de referência e fontes de informação; classificação e organização da informação; gestão da informação e do conhecimento e análise e representação temática da informação. Na tabela abaixo é possível visualizar a síntese das disciplinas apresentadas por estas quatro universidades, assim como a descrição da competência relacionada a cada disciplina.

**Tabela 1-** Principais competências relacionadas à formação de um profissional da informação

| **Disciplina** | **Descrição** |
| --- | --- |
| Serviços de referência e fontes de informação | Identificação, análise e uso de fontes gerais de informação. |
| Classificação e organização do conhecimento | Organização e tratamento da informação enquanto aplicação de conhecimentos teóricos e práticos nos processos de análise, síntese, condensação, representação e recuperação do conteúdo informacional |
| Análise e representação temática da informação | Compreensão em profundidade de determinado aspecto ou assunto visando a sua recuperação |
| Gestão do conhecimento e da informação | Conceitos, princípios, compartilhamento do conhecimento, ferramentas e modos de conversão. |

**Fonte:** Elaborado pelos autores (2022)

Uma vez adquirida uma maior clareza a respeito da composição da área da ciência da informação e algumas das suas principais competências, torna-se possível introduzir as etapas de desenvolvimento de um modelo de machine learning.

**4 AS ETAPAS DE DESENVOLVIMENTO DE UM MODELO DE MACHINE LEARNING**

É possível resumir as etapas de desenvolvimento de um modelo de machine learning nos processos de coleta e preparação dos dados usados no treinamento, escolha do algoritmo apropriado, treinamento do modelo de machine learning e a validação e talvez o reinício do processo (SANTOS, et al., 2019). Para detalhar essas quatro etapas será usado um exemplo hipotético de desenvolvimento de um modelo de machine learning capaz de diagnosticar doenças. É necessário esclarecer que o treinamento de um modelo de machine learning é muito específico, ou seja, o modelo de machine learning treinado para diagnosticar a doença X não será capaz de diagnosticar a doença Y. Em inteligência artificial e machine learning os projetos possuem um grau elevado de especificidade.

**4.1 COLETA E PREPARAÇÃO DOS DADOS**

No exemplo de desenvolvimento de um modelo de machine learning capaz de diagnosticar uma doença específica a etapa de coleta e preparação dos dados usados no treinamento, envolveria, por exemplo, coletar os dados de exames médicos de, hipoteticamente, cinco milhões de pessoas que tiveram ou não determinada doença. Na prática, para este exemplo, uma planilha com dezenas ou até centenas de colunas com cinco milhões de linhas seria criada com casos de pessoas reais. A criação desta planilha depende de um profissional da área da ciência da computação, que terá que coletar os dados de diferentes fontes e reunir tudo em uma única planilha. Também é necessário entender cada uma dessas colunas indicando o tipo de dado e a importância deste, o que na ciência da computação, é denominado dicionário de dados.

**4.2 ESCOLHA DO ALGORITMO**

Na escolha do algoritmo de machine learning apropriado, diversos algoritmos prontos estão disponíveis para os profissionais da área da ciência da computação. Na maioria dos casos, os programadores de computadores trabalham com uma linguagem de programação, por exemplo, Java ou Python, que já oferecem bibliotecas de algoritmos para machine learning. Entre os profissionais de programação, essas bibliotecas são denominadas pacotes. Por exemplo, os programadores que usam a linguagem Python têm o pacote Scikit-Learn à disposição. Esse pacote pode conter diversos algoritmos já prontos. Portanto, nesta etapa, o desafio é selecionar o algoritmo mais adequado para o treinamento que se pretende realizar com o modelo de machine learning em desenvolvimento. É válido ressaltar que o algoritmo e o modelo de machine learning consistem em entidades distintas, ou seja, um algoritmo de machine learning é selecionado para contribuir no desenvolvimento de um modelo de machine learning. Se o pacote Scikit-Learn, por exemplo, não existisse na linguagem de programação Python, isso significaria que os programadores teriam que desenvolver desde o início todas as complexas fórmulas matemáticas e estatísticas necessárias para o projeto tornando o trabalho muito mais dificultoso. Ainda nesta etapa, após a escolha do algoritmo, será necessário configurar diversos parâmetros do algoritmo selecionado.

**4.3 TREINAMENTO DO MODELO**

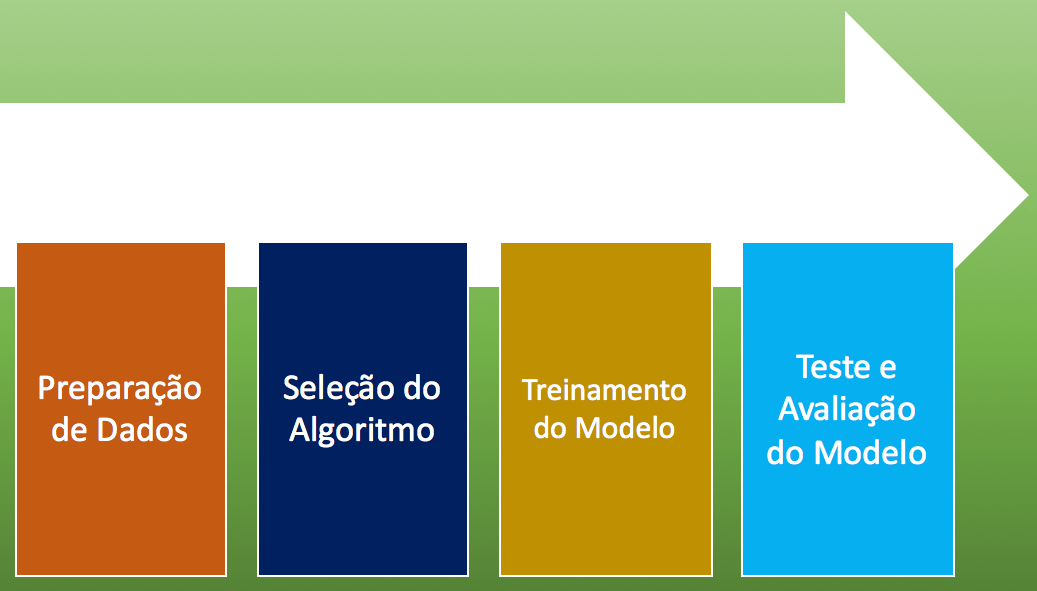
Antes de detalhar esta etapa é necessário esclarecer que o processo de treinamento de um modelo de machine learning pode levar horas ou até mesmo dias. No exemplo sobre diagnóstico de uma doença específica com uma planilha de cinco milhões de linhas e talvez centenas de colunas, o treinamento incluiria milhares de cálculos usando fórmulas complexas da área de estatística e probabilidade para identificar padrões de comportamento contidos nessa quantidade específica de dados apresentados ao modelo. É exatamente esses padrões que um modelo de machine learning aprende, ou seja, nos casos positivos existe um padrão nos números apresentados nos exames dos pacientes que indicam o diagnóstico. O mesmo ocorre para casos negativos. Alguns modelos de machine learning podem levar dias para serem treinados. Isso na prática significa que o computador usado neste treinamento ficará ligado incessantemente durante alguns dias até esse treinamento terminar e o modelo de machine learning ser criado para uso posterior. Para fins de entendimento é possível esclarecer que um modelo de machine learning treinado para reconhecer uma maçã não reconhecerá um tomate e nem uma bola de tênis. São modelos treinados para aprender a reconhecer objetos bem específicos, portanto, em projetos mais complexos, talvez vários modelos de machine learning diferentes terão que ser desenvolvidos e treinados. Uma série de detalhes podem influenciar na capacidade de aprendizado de um modelo de machine learning.

**4.4 VALIDAÇÃO DO PROCESSO**

No exemplo citado, cinco milhões de linhas existem na planilha correspondendo a cinco milhões de pessoas com diagnósticos positivos ou negativos para uma doença específica. Destes casos reais, quatro milhões foram usados para a etapa anterior de treinamento, portanto, existem um milhão de linhas, ou casos reais, que o modelo treinado não conhece. Desta forma, chega o momento de apresentar esse um milhão de casos desconhecidos para o modelo de machine learning e verificar se este realmente aprendeu e é confiável. Nesta etapa é possível também mensurar com diversas métricas a eficiência do modelo de machine learning treinado. Por exemplo, é possível constatar que em 97% dos casos o algoritmo aprendeu e, portanto, foi capaz de fazer o diagnóstico correto. É comum essas métricas não serem satisfatórias, exigindo que todas as etapas sejam repetidas. Portanto, se esse treinamento fosse repetido cinco vezes num modelo que precisou de cinco dias de treinamento, seriam vinte e cinco dias dedicados apenas a etapa relacionada ao treinamento do modelo, além do tempo necessário para a coleta e a preparação dos dados para a criação da planilha inicial.

Na ilustração abaixo é possível visualizar uma generalização das etapas descritas relacionadas ao desenvolvimento de um modelo de machine learning.

**Figura 1-** Etapas necessárias para o desenvolvimento de um modelo de Machine Learning



**Fonte:** Adaptado de https://www.datascienceacademy.com.br/

Assim que as etapas descritas forem realizadas, um modelo de machine learning que processou cinco milhões de casos reais estará treinado e pronto para ser usado no diagnóstico de uma doença específica. Uma consideração válida de ser citada é que aquela planilha com cinco milhões de linhas e centenas de colunas não é mais necessária para o funcionamento do modelo de machine learning após a conclusão das quatro etapas descritas porque sua função era treinar e validar o modelo, por isso, ela será descartada. De forma prática isso significa que outros softwares, aplicativos ou sites terão que ser desenvolvidos para que o modelo de machine learning seja funcional. Essa fase é conhecida como o deploy do modelo. Em outras palavras, por exemplo, um site restrito apenas aos médicos será criado para permitir que os dados dos exames médicos de um novo paciente sejam digitados para que o modelo de machine learning realize o diagnóstico. Outra possibilidade para evitar a digitação manual desses dados e minimizar erros, seria desenvolver um software que coleta os dados gerados após diversos exames realizados diretamente dos sistemas da clínica e automaticamente enviá-los para o modelo de machine learning realizar o diagnóstico.

Por meio da descrição destas etapas é possível verificar como que o processo de desenvolvimento de um modelo de machine learning possui uma série de detalhes significativos e complexos, que exigem uma qualidade informacional envolvida desde o momento de se delimitar o problema a ser solucionado até a definição das características que compõem esse problema, estas que irão definir desde quais dados devem ser coletados para alimentar e treinar o modelo de machine learning até a definição das métricas adequadas para a solução.

A seguir será apresentada uma conceitualização teórica por meio de estudos correlatos de como a ideia de literacia digital, assim como a questão da dimensionalidade podem contribuir para a formação de um diálogo entre a área de computação e dados com a ciência da informação para que assim possa nascer o entendimento de como que a ciência da informação por meio de seus fundamentos pode contribuir com o desenvolvimento de modelos de machine learning que usam aprendizado supervisionado.

**5 LITERACIA DIGITAL E DIMENSIONALIDADE DO PROBLEMA**

A Federação Internacional de Associações e Instituições Bibliotecárias, IFLA em inglês, em sua declaração sobre bibliotecas e inteligência artificial, destacou a importância da ciência da informação buscar uma compreensão de como os modelos de machine learning funcionam. Nesta declaração o IFLA orienta que a ‘alfabetização em inteligência artificial’ normalmente inicia por entender o básico de como a inteligência artificial e machine learning funcionam, o que eles podem e não podem fazer (IFLA, 2020). Este estudo busca contribuir para a orientação do IFLA quando se propõe a refletir na posição da ciência da informação no desenvolvimento de modelos de machine learning. Contudo, há uma importante questão: até que ponto a ciência da informação deve ir em busca dessa ‘alfabetização em inteligência artificial'? Os conceitos sobre literacia digital fornecem uma orientação neste sentido.

De acordo com Martin e Ashworth (2004), a literacia digital envolve o desenvolvimento de habilidades e o entendimento necessário para se sentir confortável em ambientes informacionais que usam tecnologias da informação e comunicação, TIC. Além disso, segundo Loureiro e Rocha (2012) o acesso, utilização e produção do conhecimento científico e acadêmico no processo de ensino e aprendizagem, é de grande importância na sociedade atual, cada vez mais dependente das tecnologias. Uma vez compreendidos estes conceitos, não se trata dos profissionais da ciência da informação se tornarem programadores, mas sim, adquirirem uma compreensão geral de como os modelos de machine learning funcionam, assim como compreender em que situações aplicá-los. Ou seja, os profissionais da ciência da informação devem ter o conhecimento necessário para se sentirem confortáveis em ambientes que modelos de machine learning estão cada vez mais inseridos (IFLA, 2020; MARTIN & ASHWORTH, 2004).

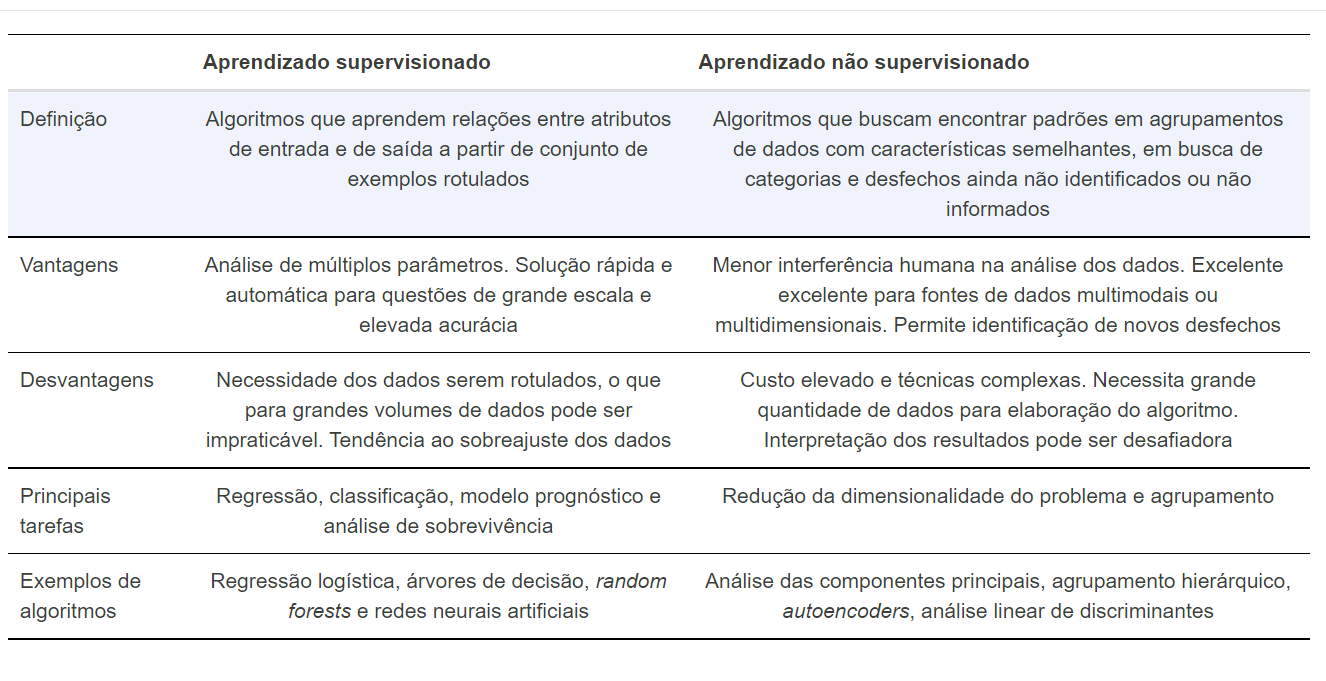
Na pesquisa correlata a este estudo, Rockembach (2021) sobre a relação entre ciência da informação e inteligência artificial no contexto de arquivos e bibliotecas, destaca a necessidade do desenvolvimento de uma literacia em inteligência artificial, com o objetivo de adquirir habilidades necessárias para colaborar em projetos interdisciplinares de inteligência artificial, por profissionais que não possuem uma formação específica nesta área, especialmente nas humanidades. O autor afirma que do mesmo modo que a literacia em informação possibilita identificar, encontrar, avaliar, aplicar e reconhecer a informação e suas fontes, uma literacia em inteligência artificial pode auxiliar o profissional da informação a adquirir habilidades necessárias para contribuir em projetos de inteligência artificial de forma interdisciplinar. O autor considera existir uma relação entre ciência da informação e inteligência artificial quando destaca que questões da organização do conhecimento, das classificações, taxonomias e ontologias, estão diretamente relacionadas às aplicações específicas de machine learning. (ROCKENBACH, 2021).

Os autores Silva e Nathanhson (2018) realizaram uma análise da produção científica em inteligência artificial na área da ciência da informação no Brasil que justifica a importância da literacia digital nas reflexões sobre a posição da ciência da informação no desenvolvimento de modelos de machine learning. Os autores identificaram uma necessidade de aprofundamento teórico e aplicado no campo da inteligência artificial, sugerindo inclusive, uma maior produção e comunicação científicas. Embora a análise da produção científica realizada por esses autores buscou apresentar os principais conceitos e aplicações da inteligência artificial, não foi localizado o termo “machine learning" no trabalho. O termo “aprendizagem automática”, uma referência ao termo atual “aprendizado de máquina”, é usado, porém, nenhum detalhe é apresentado. Esta consideração demonstra a importância da literacia digital incluir especificamente modelos de machine learning, conforme a orientação do IFLA (2020).

Silva e Nathanhson (2018) informam sobre o resultado das buscas realizadas nas bases de dados em que três artigos não abordavam a temática da inteligência artificial, cuja recuperação se deu apenas pela citação do termo, normalmente nos resumos. Isso indica que ainda há um cenário de estudos científicos a serem recuperados em uma busca porque apenas citaram algo relacionado à inteligência artificial, mas de fato, não abordam com profundidade o que as expectativas dos pesquisadores esperavam encontrar. A realização deste estudo além das orientações indicadas pela IFLA buscam se aprofundar na narrativa de Rockembach (2021) de que a informação é o objeto de estudos da ciência da informação, por isso, surge a necessidade de identificar qual é a posição adequada de se inserir a área neste contexto visando uma contribuição para a inteligência artificial por meio de seus fundamentos. O autor aponta para uma nova percepção sobre a informação como objeto de estudo da ciência da informação e da necessidade de identificar seu posicionamento nos estudos da inteligência artificial (ROCKEMBACH, 2021).

Outro conceito importante para este estudo é a dimensionalidade do problema no contexto do desenvolvimento de modelos de machine learning. Segundo Pressman (2002), independentemente do modelo de desenvolvimento de software, o processo possui três grandes fases: definição, desenvolvimento e manutenção. Mesmo considerando estas três grandes fases no contexto do desenvolvimento de modelos de machine learning, existem algumas diferenças que compõem o aprendizado supervisionado e não supervisionado como nos é apresentado na tabela abaixo baseada no estudo de Paixão et al (2022). Como nos é demonstrado, a intervenção humana ocorre com maior frequência no aprendizado supervisionado e, portanto, se trata do modelo explorado neste estudo, focando justamente nas intersecções com o profissional da informação.

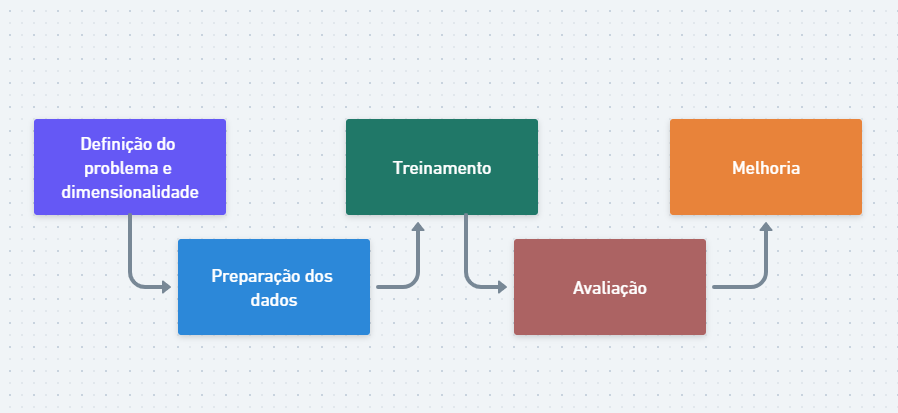
**Tabela 2 –** Comparação entre aprendizado supervisionado e não supervisionado



**Fonte:** (PAIXÃO et al., 2022)

Completando a ilustração apresentada na introdução deste estudo, para posicionar a fase de dimensionalidade do problema dentro de um processo de machine learning, foram consideradas as seguintes fases para uma solução completa de machine learning, incluindo uma etapa de revisão e aperfeiçoamento do projeto como demonstra a figura abaixo:

**Figura 2 –** Etapas necessárias para o desenvolvimento de um modelo de Machine Learning II



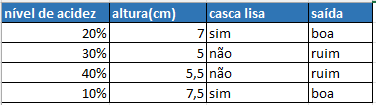
**Fonte:** Elaborado pelos autores (2022)

Primeiramente, temos a etapa de definição do problema, que é a fase da qual ocorrem os levantamentos de requisitos, como entrevista com cliente para modelagem de negócio, modelagem dos recursos envolvidos, documentação em glossário, isto é, o alinhamento entre a expectativa de negócio e o valor a ser gerado.

Logo após esta fase, inicia-se a etapa da dimensionalidade do problema, que é definida por Houaiss (2004) como a definição do número de dimensões de uma grandeza. No caso do machine learning, a grandeza a ser analisada é o problema central, isto é, nesta etapa procura-se identificar as características a partir de modelos da vista de estrutura do negócio e da vista de processos de negócio que melhor definem o contexto do problema (JÚNIOR; CAMPOS, 2008).

A dimensionalidade do problema, portanto, vai delinear todas as entradas relevantes para o problema, que serão, na etapa de treinamento, parâmetros de entrada para os algoritmos. Na tabela abaixo pode-se identificar um exemplo de dimensionalidade e definição para um problema de classificação de laranjas aptas ou não para serem comercializadas:

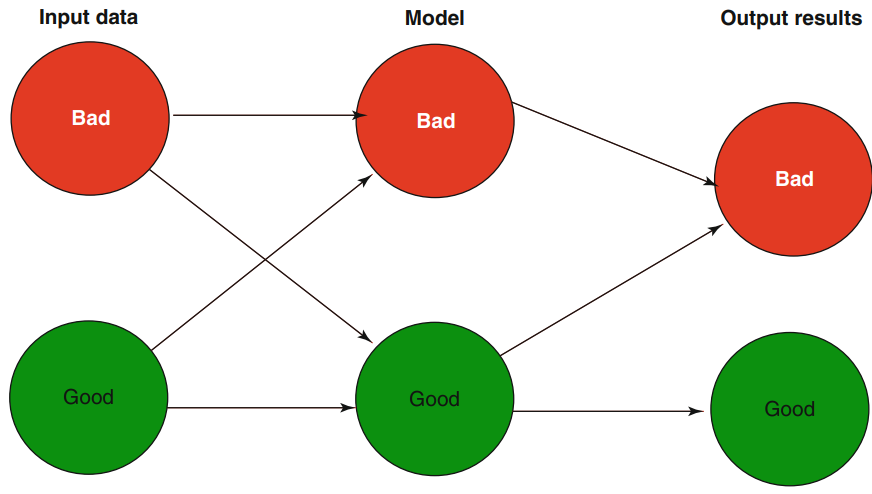
**Tabela 3 –** Exemplo de dimensionalidade de classificação de laranjas



**Fonte:** Elaborado pelos autores (2022)

A fase de definição de dimensionalidade é de fundamental importância para o sucesso do algoritmo pois, como afirma El Naqa; Murphy (2015) “um bom modelo não pode corrigir dados ruins”, e mostra na adaptação do paradigma GIGO (Garbage in, garbage out) na figura abaixo:

**Figura 3-** Paradigma GIGO

.

**Fonte:** Elaborado pelos autores (2022)

A questão de “dados ruins” observada nesta situação aplica-se na escolha incorreta ou não otimizada das classes relevantes para o problema e isto implica também na condição da quantidade de classes eleitas, pois um número reduzido de classes otimiza o processo, como afirmam os autores abaixo:

Métodos eficientes de redução de dimensionalidade são importantes não apenas para viabilizar a visualização de dados em dimensões adequadas para a percepção humana como também em sistemas automáticos de reconhecimento de padrões (MEDEIROS & COSTA, 2008).

Portanto, a etapa de dimensionalidade do problema apresenta-se como uma fase de extrema importância para o processo de construção de um modelo de machine learning, da qual o domínio do problema deve ser explorado ao máximo, norteado por um profissional que tenha o pleno entendimento do cenário apresentado e também leve em consideração as fontes em que os dados foram coletados, atividade esta da qual Medeiros & Costa (2008) entitulam de técnicas de redução de dimensionalidade (RD) objetivando reduzir o espaço de características preservando ao máximo as relações topológicas dos dados.

Retornando ao exemplo das laranjas, a definição da dimensionalidade do problema de classificação de laranjas aptas ou não para serem comercializadas envolveu uma pesquisa em fontes confiáveis, realizada por um profissional da informação, para descobrir dentro deste contexto, que a qualidade da laranja depende da análise do nível de acidez, altura em centímetros e se a casca é lisa ou não. Caberá posteriormente ao algoritmo de machine learning, na fase de treinamento do modelo, encontrar o padrão matemático nesses dados para aprender a classificar posteriormente novas laranjas como de boa ou má qualidade para comercialização, e esta análise só se torna possível graças aos dados adequadamente selecionados e coletados por um profissional da informação, que realizou da maneira correta uma definição de critérios daquele ambiente. Obedecendo a esta linha de raciocínio, o profissional da informação possui a competência necessária para compreender um determinado cenário e identificar as melhores fontes de informação, mostrando-se como um profissional essencial em um ambiente informacional, ao localizar e selecionar as principais fontes para a realização das análises de dados (CONEGLIAN; GONÇALVEZ; SEGUNDO, 2017).

Com a fase de dimensionalidade do problema, o profissional da ciência da informação possui uma fundamental importância na transcrição e transparência relacionada à compreensão do problema, uma vez que, sistemas de informações são os habilitadores do negócio, sendo assim, necessitam manter um alinhamento constante com os reais objetivos relacionados (JUNIOR; CAMPOS, 2008,).

A seleção de fontes de dados e definição de novas fontes diz respeito à tarefa do profissional de, mediante à um problema, analisar as fontes de informação disponíveis e se necessário, solicitar a captura de novas fontes que serão utilizadas em todas as fases seguintes. Para Coneglian; Gonçalves e Segundo (2017) a seleção das fontes é um fator determinante para o sucesso de uma análise, visto que não é possível gerar valor de dados sem coerência, ainda que os algoritmos de análises sejam eficientes por si só.

Ainda na etapa de dimensionalidade do problema, o profissional da informação pode atuar juntamente com uma equipe multidisciplinar para nortear a redução da dimensionalidade, pois, ainda que hajam algoritmos avançados que possam ser utilizados para automatizar esta tarefa, a coordenação e direção da equipe por um profissional da informação agrega valor e direciona a tarefa a ser realizada para um resultado otimizado (EL NAQA; MURPHY, 2015).

Na etapa de preparação dos dados, onde é necessário unificar e homogeneizar os dados oriundos de diversas fontes, que muitas vezes estão armazenados em estruturas e padrões diferentes, a ciência da informação pode colaborar com a definição de glossários e mapeamento de transformações de vocabulários, visando o alinhamento semântico dos dados (JUNIOR; CAMPOS , 2008).

Ainda que a fase do treinamento seja mais técnica, o profissional da informação pode colaborar intrinsecamente com a exposição de forma clara e transparente das métricas necessárias para o problema, visto que, existem algoritmos que apresentam um melhor desempenho com determinados tipos de métricas. Assim, com um objetivo claro e alinhado, os programadores podem dar preferência aos algoritmos que demonstraram uma melhor funcionalidade perante determinada métrica (BRAMER, 2007).

Na etapa de avaliação, o profissional da ciência da informação pode atuar tanto na conferência e alinhamento dos resultados, como também, na exposição destes resultados para a gestão, pois segundo Coneglian, Gonçalves e Segundo (2017) caso a exposição, visualização e toda a narrativa, detalhamento dos dados e informações para a equipe ou para o cliente não seja bem-sucedida, não existindo uma clareza do valor agregado, todos os momentos anteriores perdem a utilidade.

Por fim, na etapa de melhoria, o profissional da informação também se faz presente, apresentando a sua relevância ao lado do profissional da computação, atuando nesta fase em que se valerá de suas competências, contribuindo para o aprimoramento das análises (CONEGLIAN; GONÇALVES & SEGUNDO, 2017).

**6 RESULTADOS**

Por meio dos resultados obtidos com o desenvolvimento deste estudo, é possível verificar que o profissional da ciência da informação está habilitado para se inserir e atuar em uma equipe multiprofissional em todas as fases de desenvolvimento de um modelo de machine learning, devido as suas competências multidisciplinares e capacidade de lidar com todo o fluxo informacional desde sua coleta, transformação, armazenamento e recuperação.

Com base na literatura pesquisada e ao consultar a grade curricular relacionada aos cursos de biblioteconomia apresentada nas quatro universidades públicas brasileiras: Universidade de São Paulo de Ribeirão Preto (USP); a Universidade Estadual Paulista Júlio de Mesquita Filho de Marília (UNESP); a Universidade Federal de São Carlos (UFSCAR) e a Universidade Federal do Rio Grande do Sul (UFRGS) é possível relacionar uma série de disciplinas oferecidas na formação do bibliotecário, o profissional da informação que o tornam apto a colaborar em todas as etapas de desenvolvimento de um modelo de machine learning, destacando-se principalmente na fase de coleta de dados e dimensionalidade do problema, pois seus conhecimentos de técnicas de pesquisas acadêmicas, aplicação e análise de questionários e recuperação da informação podem auxiliar desde a definição de quais dados devem ser coletados até a fase de avaliação e métricas como sugere Ribeiro (2014). Portanto, no que diz respeito às primeiras etapas do desenvolvimento de um modelo de machine learning (dimensionalidade do problema e preparação dos dados), as disciplinas de serviços de referência e fontes de informação relacionam-se com o início do processo, pois envolvem todo o cuidado que um profissional da informação necessita ter ao selecionar as fontes adequadas de pesquisa para a definição das métricas necessárias a serem consideradas, estas que em momentos posteriores realizarão a alimentação do modelo. Se os dados coletados não forem precisos e de qualidade, haverá um viés e uma alteração nos resultados produzidos pelo modelo.

Na terceira etapa de construção de um modelo de machine learning, relacionada ao treinamento do modelo que está relacionada a seleção do algoritmo apropriado, disciplinas como recuperação da informação assim como análise e representação temática da informação dialogam com esta fase.

Finalmente, na quarta e na quinta etapa relacionada ao teste e avaliação do modelo visando correções e aperfeiçoamentos, encaixam-se disciplinas como gestão do conhecimento e da informação que estão intimamente ligadas à elaboração e gerenciamento de projetos, o que demonstra a importância do profissional da informação como um mediador do projeto como um todo.

A tabela abaixo apresenta estas possibilidades de atuação do profissional da informação com as diferentes etapas de desenvolvimento de um modelo de machine learning.

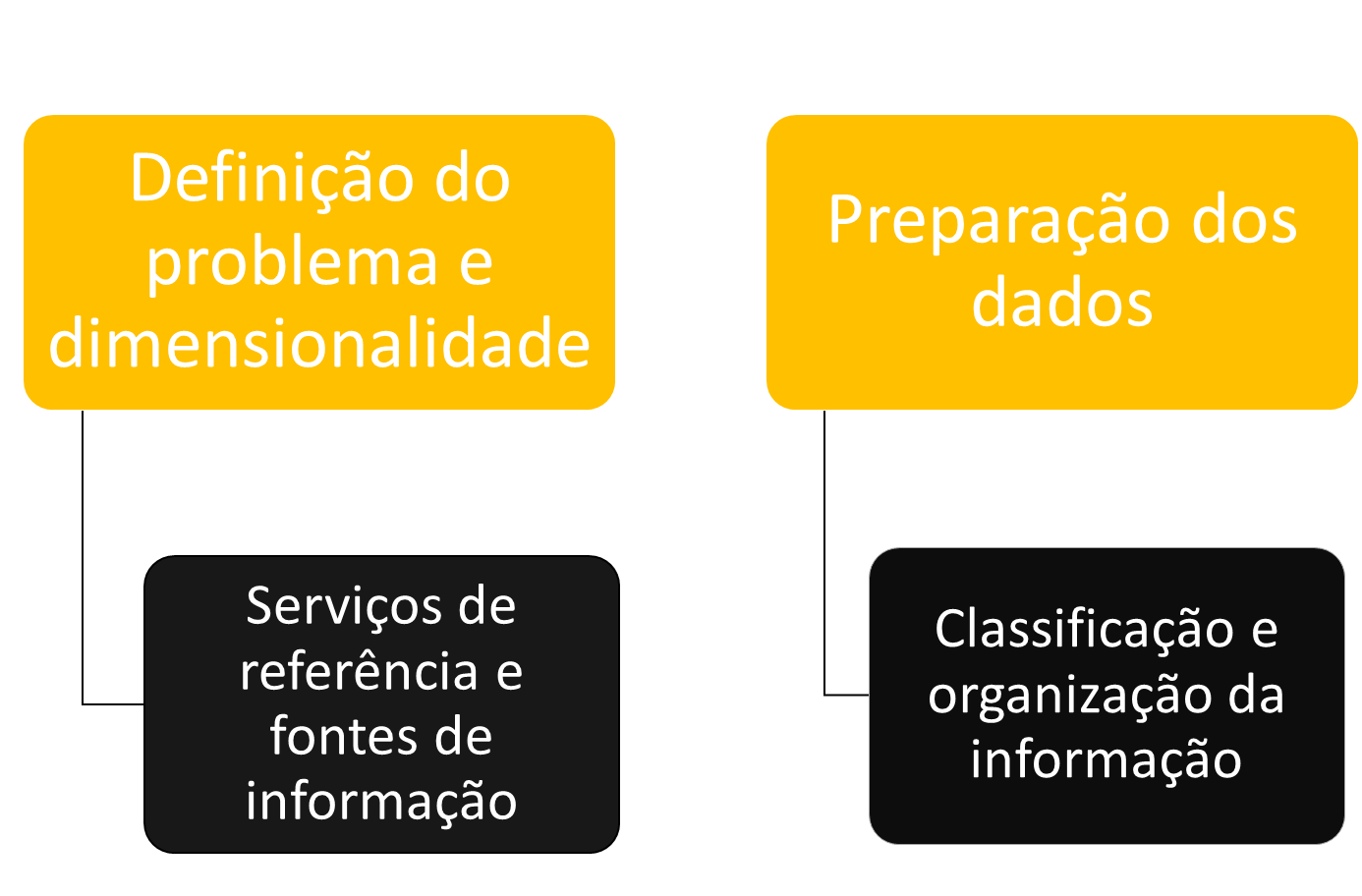
**Tabela 4** – Possíveis atuações do profissional da informação nas etapas de desenvolvimento de um modelo de machine learning

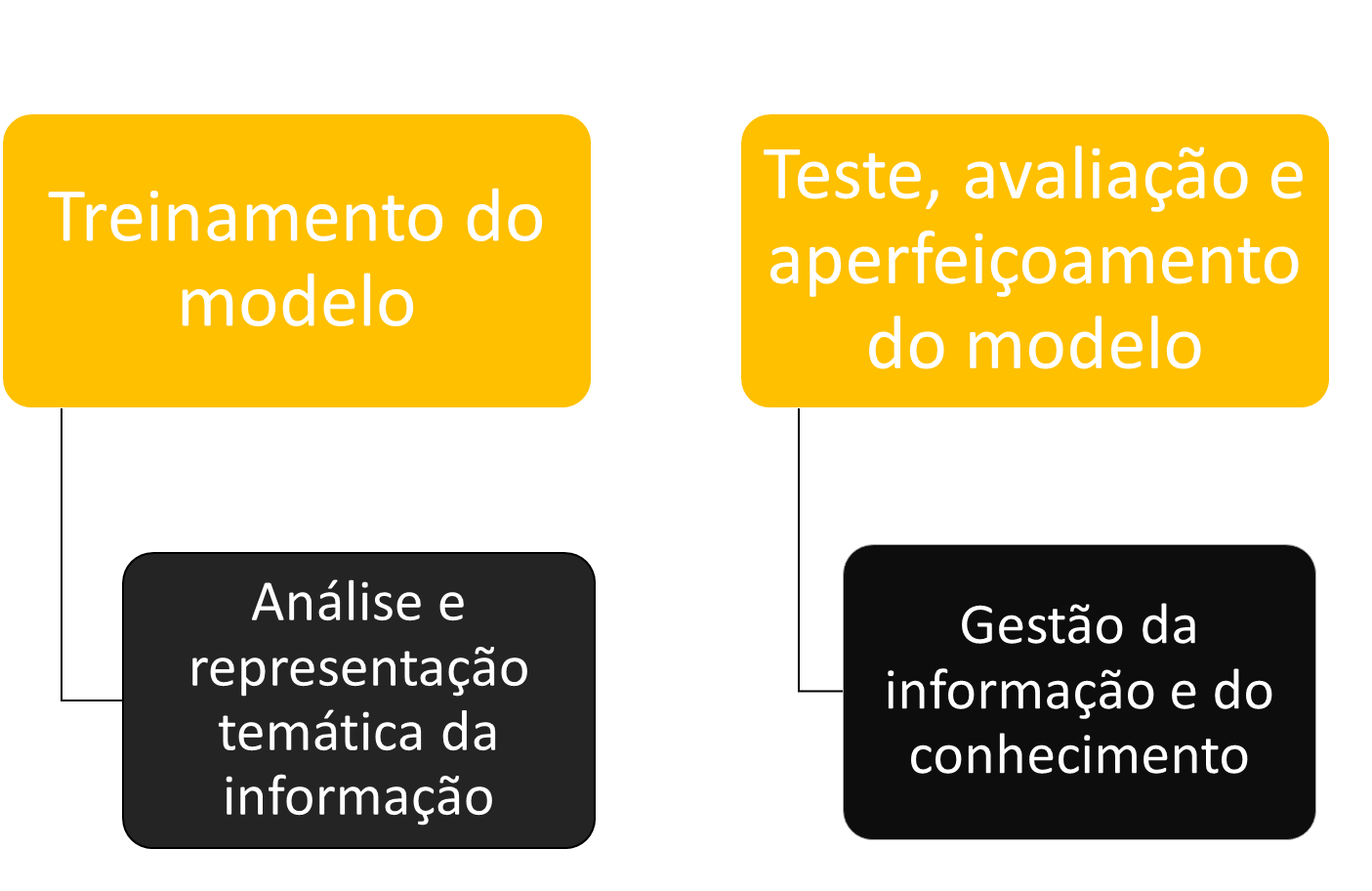
| **Definição do problema e dimensionalidade** | **Preparação dos dados** | **Treinamento** | **Avaliação** | **Melhoria** |
| --- | --- | --- | --- | --- |
| Alinhamento estratégico e de negócio com a gestão | Definição de glossários e vocabulários | Exposição das métricas relevantes para otimização de escolhas de algoritmos | Avaliação de métricas | Escolha de novas métricas |
| Seleção de fontes de dados e definição de novas | Definição de homogeneização dos dados |  | Alinhamento com gestores | Alteração da dimensionalidade |
| Definição de dimensionalidade e redução |  |  |  |  |

**Fonte:** Elaborado pelos autores (2022)

A figura a seguir apresenta uma associação entre as disciplinas oferecidas na grade curricular dos cursos de biblioteconomia nas universidades consultadas com as diferentes etapas de desenvolvimento de um modelo de machine learning.

**Figura 4-** A formação do profissional da informação associada às diferentes etapas de desenvolvimento de um modelo de machine learning





**Fonte:** Elaborado pelos autores (2022)

**7 CONSIDERAÇÕES FINAIS**

Conclui-se que ao analisar as competências relacionadas à formação do bibliotecário, o profissional da informação por meio da consulta de uma literatura especializada, da grade curricular pertencente a quatro universidades públicas brasileiras e observando a sua atuação empírica no mercado de trabalho, é possível inserir este profissional em todas as etapas de desenvolvimento de um modelo de machine learning. Contudo, observou-se que na fase relacionada à dimensionalidade do problema, as competências do profissional da informação, tais como, seleção de fontes de dados e informações adequadas de pesquisa se destacam.

Acredita-se que com a produção deste estudo haja uma contribuição com o processo de literacia digital no que diz respeito às etapas de desenvolvimento de um modelo de machine learning para que assim seja possível ao profissional da informação gerar uma compreensão de como ele pode se inserir neste processo. Sendo assim, por meio das competências relacionadas à sua formação, o profissional da informação poderá não somente enriquecer os modelos de machine learning em si como cada vez mais se posicionar em ambientes da qual é possível contar com a sua contribuição.

Como trabalhos futuros há expectativa de se realizar um levantamento da grade curricular de cursos de biblioteconomia internacionais (americanos ou europeus) visando realizar esta mesma análise para perceber o quão íntimo do processo de desenvolvimento de modelos de machine learning o profissional da informação se encontra.

**REFERÊNCIAS**

BRAMER, M. Principles of data mining. **S.l.: Springer, 2007. v. 180.**

CONEGLIAN, Caio Saraiva; GONÇALVEZ, Paula Regina Ventura Amorim; SEGUNDO, José Eduardo Santarém. O profissional da informação na era do big data. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 22, n. 50, p. 128-143, 2017.

DA SILVA, Roberta Jerônimo. A inteligência artificial no contexto da ciência da informação: uma análise de domínio. 2021.

EL NAQA, Issam; MURPHY, Martin J. What is machine learning?. **In: machine learning in radiation oncology.** Springer, Cham,2015. p. 3-11

EMYGDIO, Jeanne Louize. Artificial Intelligence from the perspective of Information Science:: where are we toward computational reasoning?. **Fronteiras da Representação do Conhecimento; Vol. 1 No. 2 (2021): Fronteiras da Representação do Conhecimento;** 171-193, v. 24, n. 2, p. 193-171.

GROENNER, Luciana Castro et al. Um Estudo Bibliométrico sobre a pesquisa em Inteligência Artificial no Brasil. **Brazilian Journal of Information Science: research trends**, v. 16, p. e02147-e02147, 2022.

GUALDANI, Fabrício Amadeu; GALVÃO, Maria Cristiane Barbosa. Perspectivas da Ciência da informação sobre o prontuário do paciente. **InCID: Revista de Ciência da Informação e Documentação**, v. 11, n. 2, p. 142-161, 2020.

GUIMARÃES, Lucas Marques Sathler; MEIRELES, Magali Rezende Gouvêa; ALMEIDA, Paulo Eduardo Maciel de. Avaliação das etapas de pré-processamento e de treinamento em algoritmos de classificação de textos no contexto da recuperação da informação. **Perspectivas em Ciência da Informação, v**. 24, p. 169-190, 2019.

IFLA. **Statement on Libraries and Artificial Intelligence.** 2020. Disponível em <<https://repository.ifla.org/handle/123456789/1646>>

JORDAN, M. I.; MITCHELL, T. M. **Machine learning: Trends, perspectives, and prospects.** Science, v. 349, n. 6245, p. 255-260, 2015. (texto disponível em nosso drive: **Science-ML-2015.pdf**)

LOUREIRO, Ana; ROCHA, Dina. Literacia digital e literacia da informação-competências de uma era digital. In: **Atas do ticEDUCA2012-II Congresso Internacional TIC e Educação**. Instituto de Educação da Universidade de Lisboa, 2012. p. 2726p.-2738p.

MARTIN, Allan; ASHWORTH, Susan. **Welcome to the Journal of eLiteracy!** 2004

MITCHELL, Tom M. **Does machine learning really work?**. AI magazine, v. 18, n. 3, p. 11-11, 1997. (texto disponível em nosso drive: **machine learning really work.pdf**)

NEVES, Bárbara Coelho. Inteligência artificial e computação cognitiva em unidades de informação: conceitos e experiências. **Logeion: filosofia da informação**, v. 7, n. 1, p. 186-205, 2020.

PAIXÃO, Gabriela Miana de Mattos; SANTOS, Bruno Campos; ARAUJO, Rodrigo Martins de; RIBEIRO, Manoel Horta; MORAES, Jermana Lopes de; RIBEIRO, Antonio L.. Machine Learning na Medicina: Revisão e Aplicabilidade. **Arq. Bras. Cardiol., v. 118, n. 1, p. 95-102,** jan. 2022.

PEIXOTO DE AZEVEDO JUNIOR, DELMIR; DE CAMPOS, RENATO Definição de requisitos de software baseada numa arquitetura de modelagem de negócios Production, vol. 18, núm. 1, enero-abril, 2008, pp. 26-46 **Associação Brasileira de Engenharia de Produção São Paulo, Brasil**

PRODANOV, Cleber Cristiano; DE FREITAS, Ernani Cesar. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico-2ª Edição**. Editora Feevale, 2013.

REIS, Luiz Claudio Rezende et al. Big Data: Um novo campo de atuação para bibliotecários. **Prisma, com**, n. 41, p. 231-250, 2020.

RIBEIRO, C. J. S. **Big Data: os novos desafios para o profissional da informação. Informação & Tecnologia,** João Pessoa, v. 1, n. 1, p. 96-105**,** jan. 2014

ROCKENBACH, Moisés. Ciência da informação e inteligência artificial: um caminho para arquivos e bibliotecas inteligentes. In: **Congresso ISKO Espanha-Portugal (5.: 2021: Lisboa): Organização do conhecimento no horizonte 2030: desenvolvimento sustentável e saúde.** Atas..[recurso eletrônico]. Lisboa: Universidade de Lisboa, 2021.

SANTOS, Hellen Geremias dos et al. **Machine learning para análises preditivas em saúde:** exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. Cadernos de Saúde Pública [online]. v. 35, n. 7. 2019. [Acessado 30 Julho 2022] , e00050818. Disponível em: <https://doi.org/10.1590/0102-311X00050818>. ISSN 1678-4464. https://doi.org/10.1590/0102-311X00050818.

SILVA, Narjara; NATHANHSON, Bruno Macedo. Análise da produção científica em Inteligência Artificial na área da Ciência da Informação no Brasil. **ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO**, v. 19, p. 111-216, 2018.

SOUZA, Marcos de; IZO JÚNIOR, Antonio; SOUZA, Renato Rocha. Modelagem de tópicos: mapeamento científico do GT-8 do ENANCIB. **ENANCIB; ENANCIB 2019**, v. 24, n. 2, 2019.

TSUNODA, Denise Fukumi; DA CONCEIÇÃO MOREIRA, Paulo Sergio; GUIMARÃES, André José Ribeiro. Machine learning e revisão sistemática de literatura automatizada: uma revisão sistemática. **Revista Tecnologia e Sociedade**, v. 16, n. 45, p. 337-354, 2020.