DOI: 10.53660/CLM-1815-23M33

CAPTO - A Method for Understanding Problem Domains for Data ScienceProjects

CAPTO - Um Método para Entendimento de Domínio de Problema para Projetos em Ciência de Dados

Received: 2023-07-16 | Accepted: 2023-08-18 | Published: 2023-08-21

Luis Enrique Zárate
ORCID: https://orcid.org/0000-0001-7063-1658
Pontifícia Universidade Católica de Minas Gerais, Brasil
E-mail: zarate@pucminas.br
Bruno Petrocchi
ORCID: https://orcid.org/0000-0003-4914-2627
Instituto de Ensino Superior, País
E-mail: brunoptcsa@gmail.com
Carlos Dias Maia
ORCID: https://orcid.org/0009-0009-1327-6499
Pontifícia Universidade Católica de Minas Gerais, Brasil
E-mail: carlosdiasmaia@gmail.com
Caio Felix
ORCID: https://orcid.org/0009-0005-1358-9242
Pontifícia Universidade Católica de Minas Gerais, Brasil
E-mail: caiofelix.reis@gmail.com
Marco Paulo Soares Gomes
ORCID: https://orcid.org/0009-0001-0611-2148
Pontifícia Universidade Católica de Minas Gerais, Brasil
E-mail: marcopaulo@pucminas.br

ABSTRACT

Data Science aims to infer knowledge from facts and evidence expressed from data. This occurs through a knowledge discovery process (KDD), which requires understanding about the application domain. However, in practice, not enough time is spent on understanding this domain, and consequently the extracted knowledge may not be correct or not relevant. Considering that understanding the problem is an essential step in the KDD process, this work proposes the CAPTO method for understanding domains, based on knowledge management models, and together with the available/acquired tacit and explicit knowledge, proposes a strategy for construction of conceptual models to represent the problem domain. This model will contain the main dimensions (perspectives), aspects and attributes that may be relevant to start a data science project. As a case study, it will be applied in the Type 2 Diabetes domain. Results show the effectiveness of the method. The conceptual model, obtained through the CAPTO method, can be used as an initial step for the conceptual selection of attributes.

Keywords: Data science; Knowledge capture; Knowledge Discovery in Databases;

RESUMO

A Ciência de Dados objetiva inferir conhecimento a partir de fatos e evidências expressas a partir de dados. Isso ocorre por meio de um processo de descoberta de conhecimento (KDD), que requer entendimento sobre o domínio da aplicação. No entanto, na prática, não é gasto tempo suficiente para entender o domínio e, consequentemente, o conhecimento extraído pode não estar correto ou não ser relevante. Considerando que a compreensão do problema é uma etapa essencial no processo KDD, este trabalho propõe o método CAPTO para compreensão de domínios, baseado em modelos de gestão do conhecimento, e juntamente com o conhecimento tácito e explícito disponível/adquirido, propõe uma estratégia para construção de modelos conceituais para representar o domínio do problema. Este modelo contêm as principais dimensões (perspectivas), aspectos e atributos que podem ser relevantes para iniciar um projeto de ciência de dados. Como estudo de caso, será aplicado no domínio da Diabetes tipo 2. Os resultados mostram a eficácia do método. O modelo conceitual, obtido por meio do método CAPTO, pode ser utilizado

Palavras-chave: Ciência dos dados; Captura de conhecimento; Descoberta de Conhecimento em Bases de Dados

INTRODUÇÃO

A Ciência de dados é uma área interdisciplinar que integra método científico, estatística, algoritmos, além de análise de dados estruturados e não estruturados, para a solução nos mais variados contextos e domínios de problemas. Seu principal objetivo é inferir conhecimento cientifico a partir de fatos e evidências expressos a partir de dados. Em geral, a área busca conhecer aspectos dos fenômenos do mundo real, e para isso, deve pressupor a identificação dos principais atributos envolvidos no domínio, e a disponibilidade de dados representativos.

Intrinsecamente, qualquer projeto em ciência de dados busca conhecimento e informação aplicando tipicamente um processo de Descoberta de Conhecimento (*Knowledge Discovery in Data Base – KDD*) como proposto por Fayyad et al., (1996). Com esse objetivo, diversas metodologias tem sido propostas (Martinez, et. al. 2021), embora todas convergem para as seis etapas essenciais sugeridas por Fayyad (seleção, pré-processamento, transformação, mineração de dados, validação, e interpretação). Em relação à etapa de seleção, Fayyad ressalta a necessidade de entendimento do domínio da aplicação e da importância do conhecimento a priori acerca do domínio.

Na prática profissional, é possível observar que responsáveis por projetos em ciência de dados muitas vezes se apresam por encontrar padrões e construir modelos por meio da aplicação das ferramentas disponíveis no mercado. Porém, um processo *KDD*, executado cuidadosa e criteriosamente requer sempre um tempo maior. Experiências bem sucedidas em projetos *KDD*, mostram que os maiores custos de um projeto (tempo e recursos) normalmente são gastos na etapa de entendimento do domínio, pré-processamento, e preparação dos dados, etapas essenciais para a correta descoberta de conhecimento por meio da aplicação de algoritmos de aprendizado de máquina. Porém, na prática, não é gasto tempo suficiente na etapa de

entendimento do domínio de problema, e por consequência, o conhecimento extraído pode não estar correto ou não ser relevante.

È importante também ressaltar que com a massificação de cursos *on-line* em ciência de dados, com foco unicamente tecnológico (aplicação de pacotes de algoritmos), é possível observar uma dissintonia e afastamento dos objetivos primários da área, o de alcançar um conhecimento correto, útil e relevante, acerca de um domínio de problema, por meio de modelos baseados em dados, como resultado da aplicação dos algoritmos de aprendizado de máquina. Isso tem levado a muitos iniciantes da área, de acreditar que a ciência de dados é capaz de fazer "alquimia de dados", de transformar qualquer conjunto de dados em conhecimento correto, útil e não óbvio. Em Guyon et al. (2018), foi realizada uma análise das soluções apresentadas pelos participantes durante os desafios em AutoML. Os autores identificaram que o pré-processamento não foi alvo dos participantes. Segundo as análises, os participantes mais bem colocados não aplicaram um processo para seleção de atributos, e 2/3 dos participantes ignoraram atributos irrelevantes.

O entendimento acerca de um domínio de problema deveria começar pela aquisição de conhecimento, e de uma base de dados representativa acerca desse domínio. Segundo Nonaka (1991), o entendimento acerca de um domínio de problema pode vir de duas fontes de conhecimento: 1º) do conhecimento tácito, vindo da experiência do especialista de domínio, normalmente redundante, subjetivo, e não formalizado (Polanyi 1967); e 2º) do conhecimento explícito, sistematizado, ou adquirido pelo estudo científico disponível acerca do domínio. De acordo com Nonaka (1991), ambos conhecimentos não se separam e são mutuamente complementares. Porém, o conhecimento tácito é de difícil captura mas poderia incorporar "*know-how*", para melhor entendimento do domínio, o que pode contribuir na identificação de padrões válidos, úteis e não óbvios em projetos de ciência de dados.

Durante a etapa de entendimento do problema, o cientista de dados, sob a orientação do especialista de domínio, deveria primeiro compreender o domínio de problema e caracterizá-lo utilizando modelos de ontologia ou mapas conceituais (Hong & Han, 2002, Cao, 2010). O objetivo dessa etapa seria identificar os atributos que podem enriquecer o conjunto de dados utilizado para o projeto de ciência de dados, podendo tornar o conhecimento extraído representativo e relevante.

A experiência mostra que o conhecimento descoberto e não óbvio, muitas vezes é resultado de atributos considerados inicialmente como pouco relevantes. A descoberta de conhecimento sobre um conjunto de dados baseado unicamente em informações fatuais pode levar a conhecimento irrelevante e óbvio, e limitar o conhecimento extraído para unicamente uma análise descritiva dos dados considerados.

A ideia do envolvimento do especialista de domínio (quem é uma fonte valiosa de conhecimento tácito relevante) para entendimento do domínio em processos de descoberta de conhecimento, não é uma recomendação atual, porém, muitas vezes é negligenciada. Ainda não há método para incorporação do conhecimento tácito, ainda na fase inicial de um processo *KDD*. Originalmente, Cao (2010) apontou a preocupação para o desenvolvimento de processos de descoberta de conhecimento corretamente executados, aderentes à realidade e necessidades das empresas. Com o intuito de atender a essa preocupação, o autor propôs o conceito de Mineração de Dados Orientada ao Domínio (D3M - Domain-Driven Data Mining) que tem por objetivo manter latentes os conhecimentos e demandas das organizações durante toda a condução de um processo de descoberta de conhecimento.

Nos trabalhos de Oliveira et. al. (2013), Meirelles e Zárate (2015), Silva et. al. (2018), Ribeiro e Zárate (2019), e Araújo et. al. (2022) os autores têm incorporado como parte das suas metodologias de descoberta de conhecimento, uma etapa de entendimento de domínio do problema, propondo modelos conceituais, com o objetivo de utilizar esses modelos como um guia para selecionar conceitualmente os principais atributos relacionados a um domínio de estudo. Por exemplo em Ribeiro e Zárate (2019) e Araújo et. al. (2022), os autores consideraram fontes de dados de alta dimensionalidade, chegando a 3.500 e 10.000 atributos disponíveis para construção dos seus modelos, respectivamente. A manipulação dessa quantidade de dados poderia tornar-se inviável computacionalmente, o que demandou um entendimento acerca do problema com a participação de especialistas de domínio. Isso permitiu a redução conceitual da dimensionalidade, tornando viável a proposta dos modelos de predição. De acordo com os autores, a etapa de entendimento foi relevante para aumentar a representatividade e assertividade dos modelos propostos, porém, os autores não descrevem um procedimento formal para captura do conhecimento tácito e explicito para construção de um modelo conceitual que represente um domínio de problema em estudo.

Considerando o pressuposto que o entendimento acerca do domínio de problema é uma etapa essencial que pode auxiliar na seleção conceitual de atributos sobre, o qual será aplicado um processo de descoberta de conhecimento, neste trabalho é proposto um método denominado CAPTO para captura do conhecimento tácito de domínio, baseado em modelos de gestão do conhecimento, e junto com o conhecimento explícito disponível/adquirido, propõe uma estratégia de construção de modelos conceituais para representação de domínios de problemas. Esse modelo conterá as principais dimensões (perspectivas), aspectos e atributos que podem ser relevantes para dar início a um projeto em ciência de dados.

O modelo conceitual, obtido pelo método CAPTO, pode ser utilizado como uma etapa inicial para seleção conceitual de atributos. Note-se que este procedimento pode diminuir o tempo gasto no processo de descoberta, tipicamente incremental, aplicado na busca por

conhecimento em bases de dados. Além disso, o entendimento formal acerca de um domínio de problema pode auxiliar na construção de "Data-Lakes" orientando os engenheiros de dados na construção desses repositórios, anteriores a sua disponibilidade em projetos futuros de ciência de dados.

FUNDAMENTAÇÃO - MODELOS PARA GESTÃO DO CONHECIMENTO

Em Kuriakose et al. (2010) uma revisão morfológica de Modelos para Gestão do Conhecimento disponíveis na literatura foi apresentado. De acordo com Heisig (2009), o objetivo principal de um Modelo de Gestão do Conhecimento (Modelo GC) é auxiliar uma organização a melhorar seus processos a partir de todo o conhecimento potencial-existente. Os Modelos GC permitem coletivamente identificar as diferentes partes constituintes de um "enigma" (por exemplo, um domínio de problema, num projeto em ciência de dados) de forma a fascilitar uma compreensão mais aprofunda dessas partes e das suas relações.

A seguir serão brevemente explicados os principais modelos GC utilizados na proposta do método CAPTO.

1) Modelo von Krogh e Roos (von Krogh e Roos, 1995). De acordo com o modelo, o conhecimento normalmente ocorre nas mentes dos membros de uma sociedade/comunidade/ grupo de pessoas, e/ou, nas ligações entre eles. O "grupo" é um sistema cognitivo, tendo a responsabilidade de iniciar ou criar modelos realistas acerca de um domínio. De acordo com os autores, o conhecimento é "corporificado", ou seja, "tudo o que é conhecido é conhecido por alguém". A partir disso, o método CAPTO utiliza deste modelo para identificar distintas perspectivas do conhecimento tácito dos especialistas.

2) Modelo Nonaka e Takeuchi (Nonaka e Takeuchi, 1995). Os autores consideram a quatro modos diferentes de conversão do conhecimento:

- Socialização: É uma etapa através da qual os membros do "grupo", compartilham conhecimento tácito de forma a aumentar o fluxo de conhecimento. O entendimento acerca de um domínio de problema é alcançado por meio da transformação do conhecimento pessoal dos indivíduos, por meio do diálogo, discurso, compartilhamento, e da narrativa. Esta fase busca encontrar os elementos constituintes sob um domínio de problema. Para o método CAPTO, esta etapa é muito relevante pois permite realizar um "*brainstorm*" para identificar diferentes perspectivas relacionados ao domínio de problema em estudo.
- Exteriorização: Nesta etapa o conhecimento adquirido pelo "grupo" deve ser convertido de tácito para explícito por meio da formalização ou sistematização. O compartilhamento de informação e conhecimentos por meio de uma representação formal, desempenha um importante papel, pois permitem ao conhecimento tácito ser internalizado. No método CAPTO a explicitação será por meio de mapas cognitivos.

• Combinação: Nesta etapa o conhecimento explícito, formalizado na etapa anterior, deve ser integrado com os mapas cognitivos construídos pelos grupos envolvidos, por meio de um modelo unificado de conhecimentos. Nesta etapa o método CAPTO propõe a definição de um mapa conceitual unificado para representação do domínio de problema.

• Internalização: Nesta etapa o conhecimento é convertido de explícito para tácito. A internalização é uma maneira responsiva de consolidar o conhecimento adquirido por meio da interação do modelo unificado, com o senso comum já estabelecido dentro de um grupo. Note que a internalização possibilita um processo dinâmico que antecede novamente à etapa de Socialização, produzindo assim o "espiral do Conhecimento" (Nonaka e Takeuch, 1995). O método CAPTO não considera a etapa de internalização, porém pode ser aplicada para fazer frente a mudanças como parte da evolução dinâmica do conhecimento.

3) Modelo Sense-Making, Choo (2002). Proporciona uma base teórica para construção do conhecimento. Os três componentes principais nos quais o modelo de Choo se sustenta são brevemente descritos a seguir:

• Dar sentido: O objetivo principal de qualquer grupo organizacional é se adaptar e prosperar em um ambiente dinâmico e altamente complexo. Neste estágio, deve-se dar sentido à informação que flui a partir do ambiente. As prioridades devem ser identificadas e usadas para filtrar as informações recebidas. Os membros dos grupos devem construir interpretações comuns a partir das suas experiências anteriores. O método CAPTO aplica este estágio para identificar dimensões e aspectos fortes sob o domínio de problema em estudo.

• Criação de conhecimento: De acordo ao modelo de Choo (2002), uma organização pode facilmente gerenciar e processar dados, e gerar novas informações para criar novos conhecimentos. Para este estágio, o método CAPTO entende que o desenvolvimento de um modelo, baseado em ciência de dados, pode contribuir com o modelo de Choo, por meio da descoberta de novo conhecimento.

4) Modelo Wiigs. Karl Wiig propôs um modelo de gestão do conhecimento para tornar o conhecimento mais útil e valioso (Wiig, 1993). O modelo amplia a visão de inteligência e a considera como a capacidade de qualquer pessoa de pensar, raciocinar, entender e agir. Do ponto de vista organizacional, todo funcionário pode exibir um comportamento inteligente. Esse modelo permite que os indivíduos adotem uma atitude mais detalhista ou refinada acerca do seu conhecimento. O método CAPTO utiliza do modelo Wiigs para propor a formação de grupos ecléticos, com participação de especialista de domínio, cientista de dados, colaboradores, pessoas envolvidas direta ou indiretamente no domínio de problema.

MÉTODO - CAPTO

O método CAPTO proposto neste trabalho, tem por objetivo a captura de conhecimento tácito, extraído de especialistas de domínio, e junto com conhecimento explícito estruturado/sistematizado, a partir de distintas fontes como literatura científica, relatórios técnicos, dicicionários das bases de dados disponíveis, propor a construção de modelos conceituais que sirvam de guia para a seleção conceitual de atributos a serem utilizados na etapa inicial do entendimento do problema em projetos em ciência de dados.

O modelo conceitual é estruturado em dimensões (diferentes perspectivas acerca do domínio de problema); em aspectos, para cada dimensão (pontos de vista relevantes para descrever o domínio de problema); e atributos (variáveis vinculadas a cada aspecto do domínio, associadas com as variáveis das fontes de dados disponíveis para o estudo). Na fase de vinculação dos atributos aos aspectos do modelo conceitual, espera-se que o engenheiro de dados tenha efetiva participação, desde que este é o responsável de fornecimento desses dados.

Como mencionado, a metodologia CAPTO (Figura 1) é fundamentada em modelos para gestão do conhecimento. O Modelo Espiral do Conhecimento proposto por Nonaka e Takeuchi (1995) foi utilizado como estrutura principal para auxiliar na conversão do conhecimento tácito em conhecimento explícito, por meio das fases de socialização, exteriorização e combinação. O modelo Krogh e Roos foi utilizado para fundamentar a contribuição dos indivíduos, envolvidos nos grupos de trabalho, para captura do conhecimento tácito.

O modelo Wiigs foi considerado para sugerir a formação dos grupos de trabalho, os quais devem ser ecléticos, com participação de especialistas de domínio. O grupo têm por objetivo apontar distintas direções de entendimento (perspectivas como dimensões do modelo). A formação dos grupos de trabalho também é reforçada pelo modelo ICAS (Chan, 2001), por meio do qual recomenda-se a inclusão nos grupos de trabalho de indivíduos que façam parte da hierarquia organizacional em torno do domínio de problema. Notemos que o ICAS contribui na execução de projetos em ciência de dados aderentes aos interesses das organizações, daí trazendo retorno eficaz para essa, observação apontada por Cao (2010). Para finalizar, o método CAPTO utiliza do modelo Sense-Making para selecionar fortes evidências (dimensões e aspectos) e descartar as fracas. A seguir apresentamos as etapas do método CAPTO.

Etapa 1: Esta etapa tem como inicio a formação de grupos de trabalho para socialização seguindo os princípios dos modelos Krogh e Roos, Wiigs, e ICAS. Esta etapa busca a identificação dos elementos constituintes acerca de um domínio de problema por meio da identificação de dimensões (perspectivas) e aspectos associados ao domínio. Para isto, pode ser aplicado uma análise divergente/convergente (Jones, 1998) de forma a analisar distintas direções do problema, e seus principais aspectos relacionados com o domínio sob estudo.

Figura 1 – Método CAPTO

Fonte: elaborada pelos autores

De acordo com Jones (1998), o pensamento divergente consiste em direcionar nossa mente em diferentes direções de um problema, procurando novas perspectivas, aspectos e evidências; enquanto que o pensamento convergente consiste em direcionar a nossa atenção, focalizando nossa mente sobre um simples aspecto do problema. Ambas são necessárias para efetivo entendimento do problemas. A divergência abre a mente para criar alternativas e a convergência peneira alternativas fracas, fortalecendo as fortes.

Note que, nesta etapa são considerados conhecimentos tácitos e explícitos, que de acordo com Nonaka, (1991), ambos conhecimentos são complementares. Esta etapa é fundamentada na fase de socialização do modelo Nonaka e Takeuchi. A seguir, o procedimento a ser adotado.

Procedimento:
1. Criar grupos de socialização com membros vinculados ao domínio de problema de forma direta (especialistas de domínio e cientistas de dados) e indireta (gestores, colaboradores, e usuários).
2. Os grupos devem ter na sua constituição especialistas de domínio, evitando mais de um “especialista” no mesmo grupo. Esta restrição é para extrair o máximo do conhecimento tácito de um especialista.
3. Sugere-se definir vários grupos contendo pelo menos um especialista de domínio.
4. Os membros de cada grupo devem compartilhar informação por meio de experiências diretas com o domínio de problema, observações, diálogos informais, reuniões, etc.
5. Aplicar uma análise divergente/convergente e identificar dimensões (perspectivas) e seus aspectos associados ao domínio de problema em estudo.

Etapa 2: Esta etapa tem por objetivo a explicitação dos conhecimentos tácito e explícito capturado na etapa anterior. A explicitação pode ser por meio de registros documentais, diagramas de causa-efeito, regras de negócio entendidas acerca do domínio, ou construção de modelos ou mapas cognitivos (Hong e Han, 2002; Costa e Krucken, 2004).

Durante esta etapa, os grupos podem construir mapas cognitivos para o domínio sob estudo. Para tal elaboração, sugere-se considerar o dicionário de dados das fontes de dados disponíveis, de forma a reduzir a diversidade de termos, o que obrigaria a aplicação de procedimentos de convergência semântica de termos. Esta etapa é fundamentada na fase de exteriorização do modelo Nonaka e Takeuchi.

Procedimento:
1. Construir mapas cognitivos considerando as principais dimensões e aspectos acerca do domínio de problema. Para esta fase, não é necessário utilizar uma representação específica para construção dos mapas. O relevante é identificar as diversas dimensões (perspectivas acerca do domínio de problema) e seus principais aspecto, associados a cada dimensão.

Etapa 3: Esta etapa tem por objetivo a troca de conhecimentos entre os grupos de trabalho, e a combinação dos modelos/mapas cognitivos explicitados na etapa anterior. A meta é a unificação dos modelos de captura do conhecimentos acerca do domínio de problema. Note que nesta fase do método, o modelo conceitual unificado é composta de dimensões e aspectos. Esta etapa é fundamentada na fase de combinação do modelo Nonaka e Takeuchi.

Procedimento:
1. Combinar os modelos/mapas cognitivos propostos com o objetivo de construir um modelo conceitual unificado. Sugere-se começar unificando as dimensões e depois os aspectos vinculados a estas. Note que deve ser procurado nomear as dimensões e os aspectos com nomes únicos e adequados, que representem seu significado e fascilite o correto entendimento.

Etapa 4: A partir do modelo conceitual unificado, o objetivo desta etapa é indicar possíveis atributos que podem caracterizem os diversos aspectos do modelo proposto. Cada atributo deve ser avaliado conceitualmente em relação a sua relevância com o domínio de problema. Esta etapa deve ser executada em conjunto pelo cientista de dados, engenheiro de dados e especialista de domínio.

Procedimento:
1. Para cada aspecto do modelo conceitual unificado indicar potenciais atributos que podem caracterizar o domínio de problema sob estudo. Esta etapa deve ser executada pelo cientista de dados, especialista de domínio e engenheiro de dados.

Etapa 5: Esta etapa têm por objetivo a congruência do modelo conceitual unificado, buscando a harmonização entre a expectativa com a disponibilidade dos dados. Para esta etapa recomenda-se a participação do engenheiro de dados, para avaliar a disponibilidade dos dados e

analisar a inviabilidade do projeto, pela falta dos mesmos. Dimensões e aspectos não sustentados por fontes de dados devem ser listados para aquisição futura, ou para apontar restrições ao modelo a ser obtido por um processo KDD, ou ainda cancelar o projeto.

Procedimento:
1. Com auxílio do engenheiro de dados, procurar associar atributo ou conjunto de atributos disponíveis para cada aspecto indicado no modelo conceitual. O engenheiro de dados deve utilizar como fontes as bases de dados internas e externas disponíveis.

APLICAÇÃO DO MÉTODO

Para aplicação do método CAPTO foi considerado como domínio de problema, a descrição do perfil de indivíduos que apresentam a doença crônica <Diabetes, Tipo 2>. O objetivo é descrever as características desse tipo de indivíduos com o propósito de aplicar um processo de descoberta de conhecimento.

A fonte de dados utilizada corresponde a um estudo acerca do estado de saúde, estilos de vida, doenças crônicas e de saúde bucal realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE, 2019), o PNS-2013 (Pesquisa Nacional em Saúde). O estudo considera além das informações do domicílio, as características gerais dos moradores, o nível de educação, trabalho e rendimentos. A base contêm informação se o indivíduo possui alguma deficiência, se possui plano de saúde, e/ou utiliza os serviços públicos de saúde. Características específicas da saúde para os indivíduos com 60 anos ou mais, e de crianças menores de 2 anos de idade são também considerados. Informações de acidentes de trânsito ou de trabalho, estilos de vida (hábitos de alimentação, prática de atividade física, uso de bebidas alcoólicas e fumo) e de doenças crônicas são também registradas. Se mulher, questões acerca do estado de saúde, atendimentos pré-natal. Informações de violência, saúde bocal, doenças transmissíveis, atividade sexual, relações e condições de trabalho são também consideradas.

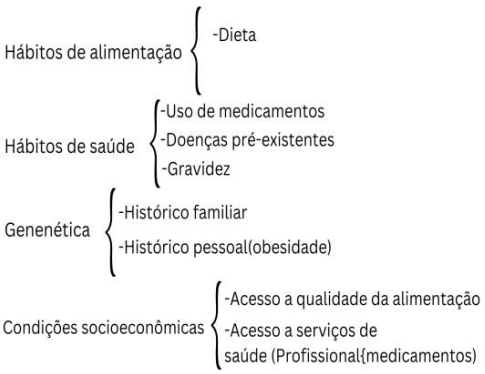
O estudo considerou 94000 domicilios, 205.546 indivíduos, e 942 atributos (questões respondidas). É importante ressaltar que a base de dados pode ser considerada de alta dimensionalidade, requerendo um processo de seleção de atributos para posterior aplicação de um processo de descoberta de conhecimento, daí a relevância do método CAPTO para seleção conceitual desses atributos.

Seguindo os passos do método proposto, a partir da formação de 3 grupos de trabalho (formados por um especialista de domínio, um colaborador e um cientistas de dados), como resultado da fase de socialização, foram identificados por meio da captura do conhecimento tácito/explícito, dimensões e aspectos relevantes ao domínio (Etapa 1). O primeiro resultado dos grupos de trabalho correspondeu à produção de mapas cognitivos, os quais representam as diferentes dimensões e aspectos associados ao domínio de estudo (Etapa 2).

Os procedimentos adotados nas duas primeiras etapas foram: 1) redução da diversidade de termos, por meio da convergência semântica dos mesmos. Para minimizar a divergência, todos os grupos consideraram o mesmo dicionário de dados, a partir das fontes de dados disponíveis para o PNS-2013, IBGE (2019); 2) Foi evitado rotular uma dimensão com termos muito abrangentes. Por exemplo, 'Estilo de vida', o qual pode englobar tanto hábitos alimentares como hábitos de saúde; 3) Foi procurado utilizar termos de significado único vinculados ao domínio, rotulando as dimensões com ideias específicas. Por exemplo, termos específicos como 'Hábitos de saúde' e 'Hábitos alimentares'. O objetivo é evitar discrepâncias de significados; 4) Previamente, foram definidos os termos utilizados para guiar a construção dos mapas cognitivos. Por exemplo, os seguintes significados foram estabelecidos: drogas: substância que afeta a consciência humana; dieta: alimentos e bebidas de um grupo social.

Como resultado de aplicar as Etapas 1 e 2, os modelos elaborados pelos três grupos de trabalho, visando descrever dimensões e aspectos associados ao domínio <Diabetes, Tipo 2>, são mostrados nas Figuras 2-4.

Figura 2 - Mapa conceitual - Grupo de trabalho1



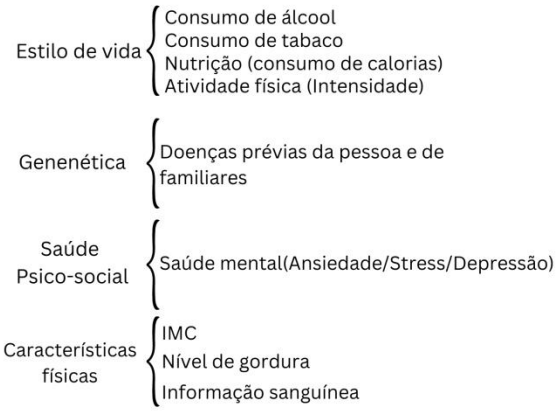
Fonte: elaborada pelos autores

Figura 3 - Mapa conceitual - Grupo de trabalho2

-Excesso de peso
-Apneia do sono
-Sexo (gravidez)
-Idade
-História familiar
-Stress emocional
-Estado socio-econômico
-Sedentarismo
-Uso de medicamentos

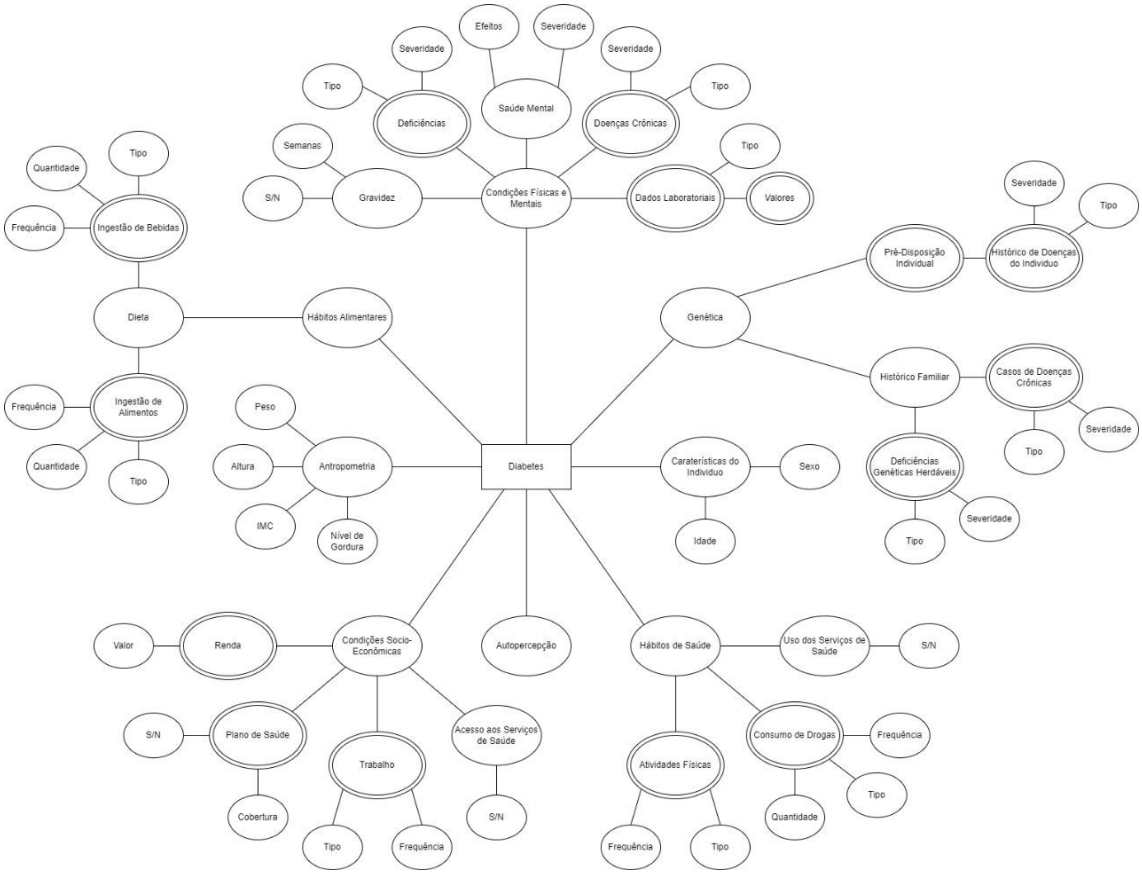
Fonte: elaborada pelos autores

Figura 4 - Mapa conceitual - Grupo de trabalho 3



Fonte: elaborada pelos autores

Figura 5. Modelo conceitual unificado para o domínio de Diabetes, Tipo 2



Fonte: elaborada pelos autores

A Etapa 3 do método CAPTO objetiva construir um modelo conceitual unificado, obtido pela socialização dos grupos de trabalho. Como resultado desse processo, a Figura 5 mostra um modelo conceitual unificado para o domínio <Diabetes, Tipo 2>. A partir desse modelo, foi procurado associar atributos para cada aspecto (Etapa 4). Os atributos propostos

devem ser analisados em relação a sua representatividade com o domínio de problema. Por exemplo, a dimensão <Hábitos alimentares>, para o aspecto <Dieta de ingestão de bebidas>, os atributos associados foram <Tipo de bebida, Quantidade, Frequência>. Após aplicação da Etapa 5, por meio da qual, é realizada a vinculação dos aspectos e atributos indicados com os atributos disponíveis nas fontes de dados, a Tabela 1, sintetiza as dimensões, aspectos e atributos considerados relevantes para o domínio <Diabetes, Tipo 2>.

Tabela 1: Dimensões, Aspectos e Atributos vinculados para o domínio Diabetes, Tipo 2

Descrição do Mapa Conceitual – Domínio de problema: Diabetes Tipo 2		
Dimensão: Hábitos de saúde		
Aspectos (conhecimento explícito e estudo científico vinculado)	**Atributos associados ao aspecto**	**Atributos vinculados com as fontes de dados**
Uso dos serviços de saúde: O acompanhamento do paciente por profissionais de saúde, e a frequência de exámens clínicos, se mostraram relacionados com o uso de medicamentos, e com a necessidade de busca de emergência/ internação, por indivíduos com diabetes (Freitas et al 2018).	- Motivo de saúde que requereu o uso dos serviços de saúde - Diagnóstico médico - Última consulta - Uso de medicamentos	Módulo J - Utilização dos serviços de saúde e Módulo Q – Doenças Crônicas: J4a, J7, J11a, J14, J15a, Q32a, Q33b, Q34c, Q38a3 e Q39a; Fonte: Base de dads PNS (BD-PNS)
Consumo de Drogas: O uso de álcool (Sarit Polsky e Halis K. Akturk), fumo e drogas influenciam negativamente a imunidade dos indivíduos (Pastor et al 2020).	- Quais drogas são consumidas - Qual a frequência de consumo de drogas - Qual a quantidade de drogas consumidas	Módulo P – Estilos de Vida: P27, P28a, P29, P50, P54, P56, P67 e P67a; Fonte: BD-PNS
Atividade Física: Atividades físicas estão intrinsecamente ligadas uma vez que tornam seu corpo mais sensível à insulina (o hormônio que permite que as células do seu corpo usem o açúcar do sangue como energia), o que ajuda a prevenir e controlar a diabetes, (LaMonte et al. 2005).	- Quais atividades físicas são realizadas - Com qual frequência são realizadas atividades	Módulo P – Estilos de Vida: P34, P35, P37 e P36; Fonte: BD-PNS
Dimensão: Hábitos alimentares		
Dieta - Ingestão de alimentos: A alimentação está ligada à diabetes como citado por Reni Aparecida Barsaglini e Ana Maria Canesqui (Barsaglini e Canesqui, 2010).	- Quais alimentos são consumidos - Qual a frequência de consumo de alimentos - Qual a quantidade de alimentos consumidos	Módulo P – Estilos de Vida: P6a até P26a; Fonte: BD-PNS
Dieta - Ingestão de bebidas: A alta ingestão de álcool aumenta o risco de diabetes (Polsky e Akturk, 2017)	- Quais bebidas são consumidas - Qual a frequência de consumo de bebidas - Qual a quantidade de bebidas consumidas	Módulo P – Estilos de Vida: P6b até P24a; Fonte: BD-PNS

Dimensão: Condições Físicas e Mentais		
Deficiências: As deficiências podem ter impacto direto nas atividades físicas que uma pessoa realiza e atividades físicas tem impacto direto na diabetes (LaMonte el al, 2005).	- Informações sobre deficiências que possa possuir	Módulo G – Pessoas com Deficiências; Fonte: BD-PNS
Gravidez: Grávidas têm maior suscetibilidade a diabetes (McCance 2011).	- Status (condição S/N) - Influência da diabetes na gravidez	Módulo P – Estilos de Vida e Módulo Q – Doenças Crônicas: P5 e Q30b; Fonte: BD-PNS
Saúde Mental: Doenças mentais como a depressão são fatores de risco para a diabetes (Fráguas et al 2009).	- Possui alguma doença mental - Qual o efeito na vida cotidiana caso possua uma doença mental	Módulo J - Utilização dos serviços de saúde e Módulo Q – Doenças Crônicas: J7, Q92, Q110a e Q115; Fonte: BD-PNS
Percepção: Por meio da percepção individual, as pessoas relatam dores ou sintomas, que podem estar ligados a diabetes (Lee et al 2020).	- Autoavaliação - Percepção de sintomas de Diabetes	Módulo N – Percepção do estado de saúde: N1, N1a e N14; Fonte: BD-PNS
Doenças crônicas: Algumas doenças crônicas, como doenças renais, estão intimamente ligadas a diabetes **(**Koye et al 2018**):**	- Possui alguma doença crônica (especialmente a própria diabetes) - Efeitos da diabetes	Módulo Q – Doenças Crônicas: Q30a e Q55a; Fonte: BD-PNS
Dados laboratoriais: Dados laboratoriais podem prever diabetes previamente (Gagliardino et al 2017)	- Quais exames foram requisitados a respeito de diabetes	Módulo Q – Doenças Crônicas: Q29a, Q47a e Q51a; Fonte: BD-PNS
Dimensão: Condições Sócio-econômicas		
Renda: A renda tem relação com a diabetes diretamente, uma vez que ela reflete ela é fator importante para diversos pontos, como acesso a saúde, dieta, entre outros (Bird 2015).	- Rendimento de trabalhos - Rendimento de outras fontes	Módulo E – Características de trabalho das pessoas 14 anos ou mais de idade e Módulo F – Rendimentos de outras fontes: E16, E18 e F1a até F14a; Fonte: BD-PNS
Acesso a serviços de saúde: Acesso a serviços de saúde é importante para diagnósticos (Zhang et al 2012).	- Acesso a farmácias e profissionais de saúde - Dificuldades de acesso a farmácias e profissionais de saúde	Módulo Q – Doenças Crônicas: Q33a, Q34d, Q37a, Q38a4, Q38a6, Q40a, Q43 e Q50; Fonte: BD-PNS
Plano de saúde: Planos de saúde refletem nos acessos à saúde do indivíduo (IMCCU 2002).	- Informações sobre o plano de saúde caso possua	Módulo I – Cobertura de Plano de Saúde; Fonte: BD-PNS
Trabalho: O trabalho é relacionado com atividades físicas que é intimamente ligado a ocorrência de diabetes (Hu et al 2003).	- Informações sobre o trabalho caso possua - Consequências do trabalho sobre a saúde - Interferências de doenças crônicas no trabalho	Módulo E – Características de trabalho das pessoas 14 anos ou mais de idade e Módulo M – Características do trabalho e apoio social: E12, E14a, E17, E19, M5d e M6; Fonte: BD-PNS
Dimensão: Características do indivíduo		
Sexo: Sexo esta ligado a diabetes, que afeta mais os homens em nosso país (Gale e Gillespie 2001).	- Sexo	Módulo C – Características gerais dos moradores: C6; Fonte: BD-PNS
Idade: A idade afeta principalmente a população mais velha (Laakso e Pyörälä,	- Idade	Módulo C – Características gerais dos moradores: C7 e C8;

1985).		Fonte: BD-PNS
Dimensão: Genética		
Predisposição individual: Os fatores de risco para desenvolver o diabetes são: sedentarismo, história familiar de diabetes em parentes de 1º grau (Ali 2013).	- Recomendações para reduzir os efeitos da diabetes	Módulo Q – Doenças Crônicas: Q46a; Fonte: BD-PNS
Histórico familiar – Casos de doenças crônicas: A diabetes está diretamente ligada à hereditariedade.	- Tipo - Severidade	Dados indisponíveis na Fontte: BD-PNS
Dimensão: Antropometria		
Peso e altura: Utilizados para calcular o IMC, fazendo referência a características como a obesidade, que é um fator causador da diabetes (Leong e Wilding, 1999).	- Peso (Múltiplas medições) - Altura (Múltiplas medições)	Módulo P – Estilos de Vida e Módulo W – Antropometria: P1a, P4a, W00201, W00202, W00101 e W00102; Fonte: BD-PNS

Fonte: elaborada pelos autores

VALIDAÇÃO DO MÉTODO

De forma a validar o método CAPTO, este foi aplicado sobre a base de dados da PNS-2013. Como mencionado, a PNS visa obter informações detalhadas sobre a saúde e o estilo de vida da população brasileira, de forma a identificar as principais doenças e fatores de risco para a saúde, além de avaliar o acesso e a qualidade dos serviços de saúde no país.

Para o processo de validação foi considerado o domínio <Diabetes, Tipo 2> e construído modelos supervisionados de classificação por árvore de decisão. Os modelos são ajustados para classificar (descrever) os indivíduo nas classes 1: Indivíduos com Diabetes, e 2: Indivíduo não diabético. Para construção do conjunto de dados foi considerado somente os indivíduos diagnosticados clinicamente com a doença. Mulheres diagnosticadas com diabetes por gravidez foram desconsiderados deste processo. No total foram identificados 3636 instâncias de indivíduos diabéticos. Para contrapor informações aos modelos de classificação, a mesma quantidade de indivíduos foram inseridos ao conjunto de dados, totalizando 7272 instâncias.

O processo de validação consistiu na obtenção de modelos de classificação para três conjuntos de dados: o primeiro conjunto (DataSet1) considera a base de dados completa, sem seleção conceitual de atributos, nem tratamento de dados ausentes. O segundo conjunto (DataSet2) considera casos completo por meio de um pré-processamento de eliminação de atributos com grande quantidade de dados ausentes. Foi aplicado também a eliminação de casos (registros) com presença de dados faltantes. O terceiro conjunto (DataSet3) para o qual foi aplicado o método CAPTO para seleção conceitual de atributos, como indicado na Tabela 1. Para este último conjunto de dados, foi também considerado casos completos.

Os três conjuntos de dados foram submetidos ao modelo baseado em Árvore de decisão. A construção das árvores considerou o Ganho de Informação para construção das árvores, e como critério de parada, para expansão da árvore, foi considerado até 5 instâncias por nó. A

Tabela 2 mostra os resultados dos modelos de classificação (medida F1) para diversos tamanhos do conjunto de treinamento e validação. A escolha aleatória das instâncias para treinamento, para todos os modelos, receberam a mesma semente.

Tabela 2: Validação do método CAPTO. Resultados por árvore de decisão - medida F1.

		Tamanho do Conjunto de Treinamento				
	Dimensão	90%	70%	50%	30%	10%
DataSet1	7272 x 942	1: 100 2: indeterm.	1: 99,9 2: indeterm.	1: 99,7 2: indeterm.	1: 100 2: indeterm.	1: 99,7 2: indeterm.
DataSet2	5291 x 50	1: 80,7 2: 72,6	1: 68,9 2: 23,4	1: 71,0 2: 74,8	1: 60,5 2: 57,8	1: 48,5 2: 54,7
DataSet3	5291 x 31	1: 72,6 2: 71,5	1: 77,9 2: 69,3	1: 71,0 2: 73,8	1: 66,3 2: 69,7	1: 70,4 2: 75,5

Pelos resultados da Tabela 2 é possível observar que o pior desempenho foi obtido para a base de dados DataSet1. O modelo não foi capaz de predizer nenhuma instância da classe 2: Indivíduo não diabético. O conjunto DataSet2, o qual considerou somente casos completos apresentou um desempenho em degradação a medida que o tamanho do conjunto de treinamento foi reduzido. O conjunto DataSet3, no qual foi aplicado o método CAPTO, foi o mais estável frente à redução do tamanho do conjunto de treinamento, mostrando o melhor resultado. As simulações foram implementadas no ambiente Knime.

CONSIDERAÇÕES FINAIS

Neste trabalho o método CAPTO, para entendimento e representação conceitual de domínios de problemas, para o desenvolvimento de projetos em ciência de dados, é proposto. O método, baseado em modelos para gestão do conhecimento, permite capturar o conhecimento tácito do especialista de domínio, e junto com o conhecimento explícito disponível, propõe a construção de modelos conceituais para representação de um domínio de problema. Esses modelos contêm as principais perspectivas (dimensões), aspectos e atributos relevantes para dar início a um projeto em ciência de dados.

Como mencionado anteriormente, o entendimento do domínio de problema é normalmente negligenciado, e isso pode levar à descoberta de padrões óbvios, não relevantes, e inclusive errôneos. É importante ressaltar que a falta de atributos para representar o modelo conceitual, é um indicativo que a descoberta de conhecimento pode ter restrições, ou que inclusive o projeto pode ser abortado.

Tradicionalmente processos de descoberta de conhecimento a partir de dados é realizada por meio de um processo incremental, onde retorno as etapas anteriores do processo KDD, por exemplo, na etapa de enriquecimento do conjunto de dados, é uma ação comum. Porém, o retorno às etapas anteriores, com o objetivo de melhorar o desempenho dos algoritmos é um

processo custoso. Daí, iniciar pelo entendimento do domínio de problema, representado por um modelo conceitual, ajuda a não ter a necessidade de retornar às etapas iniciais, pois a seleção conceitual já foi realizada considerando todas as variáveis disponíveis na base de dados.

Note que todo modelo conceitual, acerca de um domínio de problema é sempre uma aproximação da realidade, daí a descoberta de conhecimento corresponderá a uma interpretação ou visão acerca do domínio de problema em estudo.

É importante ressaltar que a modelagem conceitual pode ser relevante na construção de ”Data-lakes”. Esses repositórios poderiam identificar antecipadamente os dados que devem fazer parte destes, antes de iniciar um processo de ciência de dados.

REFERÊNCIAS

ALI, O. Genetics of type 2 diabetes. **World J Diabetes**, 4(4), p. 114-23, 2013. doi: 10.4239/wjd.v4.i4.114.

ARAÚJO, A. S.; SILVA, A. R.; ZÁRATE, L .E. Extreme precipitation prediction based on neural network model – A case study for southeastern Brazil, **Journal of Hydrology**, V. 606, 127454 2022. doi: 10.1016/j.jhydrol.2022.127454.

BARSAGLINI, R. A.; CANESQUI, A. M. A alimentação e a dieta alimentar no gerenciamento da condição crônica do diabetes. **Saude soc.** 19 (4), 2010. doi: 10.1590/S0104-12902010000400018.

BIRD, Y.; LEMSTRA, M.; ROGERS, M.; MORAROS, J.; The relationship between socioeconomic status/income and prevalence of diabetes and associated conditions: A cross-sectional population-based study in Saskatchewan, Canada. **Int J Equity Health**, 14:93, 2015. doi: 10.1186/s12939-015-0237-0.

CAO, L. Domain Driven Data Mining: Challenges and Prospects. In: IEEE Transactions on Knowledge and Data Engineering, v. 22, n. 6, p. 755-769, 2010. Disponível em: <http://www.computer.org/csdl/trans/tk/2010/06/ttk2010060755-abs.html>

CHAN, S.; Complex adaptive systems. In: ESD. 83 Research seminar in engineering systems. Cambridge, MA, USA: MIT, p. 1-9, 2001. Disponível em: < https://web.mit.edu/esd.83/www/notebook/Complex%20Adaptive%20Systems.pdf>

CHOO, Ch. W. **Information management for the intelligent organization: the art of scanning the environment**. Information Today (Ed), Inc., ISBN 1573871257, 2002. 325 p.

COSTA, M. D., KRUCKEN, L. Aplicações de mapeamento do conhecimento para a competitividade empresarial. In: KM BRASIL 2004 - Gestão do Conhecimento na Política Industrial Brasileira, São Paulo, 2004. Disponível em: < https://cmapspublic3.ihmc.us/rid=1237746139625_1042789295_8469/mapas%2Bdo%2Bconhecimento%2Bcosta%2Bkrucken.pdf>

de OLIVEIRA, F. A.; NOBRE, C.; ZÁRATE, L. E. Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction

index – Case study of PETR4, Petrobras, Brazil, **Expert Systems with Applications**, v. 40, n. 18, p. 7596-7606, 2018. doi: 10.1016/j.eswa.2013.06.071.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. **Communication ACM**, New York, NY, USA, 39(11), p. 27-34, 1996. doi: 10.1145/240455.240464.

FRÁGUAS, R.; SOARES, S. M.; BRONSTEIN, M. D. Depressão e diabetes mellitus. Arch. **Clin. Psychiatry** (São Paulo), 3, s. 3, 2009. doi: 10.1590/S0101-60832009000900005

FREITAS, P. S.; MATTA, S. R.; MENDES, L. V. P.; LUIZA, V. L.; CAMPOS, M. R. Uso de serviços de saúde e de medicamentos por portadores de Hipertensão e Diabetes no Município do Rio de Janeiro, Brasil. **Ciênc. saúde colet**., v. 23, n. 7, 2018. doi: 10.1590/1413-81232018237.21602016

GAGLIARDINO, J. J.; ELGART, J. F.; BOURGEOIS, M.; ETCHEGOYEN, G.; FANTUZZI, G.; RÉ, M.; RICART, J. P.; GARCÍA, S.; GIAMPIERI, C.; GONZÁLEZ, L.; SUÁREZ-CRIVARO, F.; KRONSBEIN, P.; ANGELINI, J. M.; MARTÍNEZ, C.; MARTÍNEZ, J.; RICART, A.; SPINEDI, E.; Diabetes primary prevention program: New insights from data analysis of recruitment period. **Diabetes Metab Res Rev**., v. 34, n. 1, 2018. doi: 10.1002/dmrr.2943.

GALE, E.; GILLESPIE, K. Diabetes and gender. **Diabetologia,** v. 44, p. 3–15, 2001. doi: 10.1007/s001250051573

GUYON, I.; SUN-HOSOYA, L.; BOULLÉ, M.; ESCALANTE, H. J.; ESCALERA, S.; LIU, Z.; JAJETIC, D.; RAY, B.; SAEED, M.; SEBAG, M.; STATNIKOV, A.; TU, W-W; VIEGAS, E. Analysis of the AutoMl Challenge series 2015-2018. Frank Hutter; Lars Kotthoff; Joaquin Vanschoren (eds). AutoML: Methods, Systems, Challenges, Springer Verlag, In: press, The Springer Series on Challenges in Machine Learning. 2019. doi: 10.1007/978-3-030-05318-5_10

HEISIG, P. Harmonisation of knowledge management: comparing 160 KM frameworks around the globe. **Journal of knowledge management**, v. 13, n. 4, p. 4-31, 2009. doi: 10.1108/13673270910971798

HARRISON, T. A.; HINDORFF, L. A.; KIM, H.; WINES, R. C. M.; BOWEN, D. J.; MCGRATH, B .B.; EDWARDS, K. L. (2003) Family history of diabetes as a potential public health tool, **American Journal of Preventive Medicine**, v. 24, n. 2, p. 152-159. https://doi.org/10.1016/S0749-3797(02)00588-3.

HONG, T., HAN, I. Knowledge-based data mining of news information on internet using cognitive maps and neural networks. **Expert Systems with Applications**, v. 23, p. 1-8, 2002. doi: 10.1016/S0957-4174(02)00022-2

HU, G.; QIAO, Q.; SILVENTOINEN, K.; ERIKSSON, J. G.; JOUSILAHTI, P.; LINDSTRÖM, J.; VALLE, T. T.; NISSINEN, A.; TUOMILEHTO, J. Occupational, commuting, and leisure-time physical activity in relation to risk for Type 2 diabetes in middle-aged Finnish men and women. **Diabetologia,** v. 46, n. 3, p. 322-9, 2003. doi: 10.1007/s00125-003-1031-x

IBGE, 2019. Instituto Brasileiro de Geografia e Estatística. Pesquisa nacional de saúde: 2019: percepção do estado de saúde, estilos de vida, doenças crônicas e saúde bucal: Brasil e grandes regiões. IBGE, Coordenação de Trabalho e Rendimento. Rio de Janeiro: **IBGE**; 2020. 113p. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/saude/29540-2013-pesquisa-nacional-de-saude.html>

IMCCU, (2002). **Institute of Medicine (US) Committee on the Consequences of Uninsurance. Care Without Coverage: Too Little, Too Late.** Washington (DC): National Academies Press (US); 2002. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK220639/ doi: 10.17226/10367>

JONE, M. D. **The Thinker´s Toolkit": fourteen powerful techniques for problem solving**. Time Business, Random House. 1998, 384 p.

KURIAKOSE, K. K.; RAJ, B.; MURTY, S. A. V. S.; SWAMINATHAN, P. Knowledge Management Maturity Models – A Morphological Analysis. **Journal of Knowledge Management Practice**, v. 11, n. 3, p. 1-10, 2010.

LEONG, K. S.; WILDING, J. P.; Obesity and diabetes, **Best Practice & Research Clinical Endocrinology & Metabolism**, v. 13, n. 2, 1999. Doi: 10.1053/beem.1999.0017

KOYE, D. N.; MAGLIANO, D. J.; NELSON, R. G.; PAVKOV, M. E. The Global Epidemiology of Diabetes and Kidney Disease. **Adv Chronic Kidney Dis**, 25(2), p. 121-132, 2018. doi: 10.1053/j.ackd.2017.10.011.

LAMONTE, M. J.; BLAIR, S. N.; CHURCH, T. S. Physical activity and diabetes prevention. **Journal of Applied Physiology**, v. 99, p. 1205-1213, 2005. doi: 10.1152/japplphysiol.00193.2005

LAMONTE, M. J.; BLAIR, S. N.; CHURCH, T. S. (2005). Physical activity and diabetes prevention. **Journal of Applied Physiology,** v. 99, n. 3, p. 1205-1213, 2005. Doi: 10.1152/japplphysiol.00193.2005

LAAKSO, M.; PYÖRÄLÄ, K. (1985). Age of Onset and Type of Diabetes. **Diabetes Care,** v. 8, n. 2, p. 114–117, 1985.

LEE, J-W.; MOON, J.S.; KANG, D.R.; LEE, S.J.; SON, J-W.; YOUN, Y.J.; AHN, S.G.; AHN, M-S.; KIM, J-Y.; YOO, B-S.; LEE, S-H.; KIM, J.H.; JEONG, M.H.; PARK, J-S.; CHAE, S.C.; HUR, S.H.; CHO, M-C.; RHA, S.W.; CHA, K.S.; CHAE, J.K.; CHOI, D-J.; SEONG, I.W.; OH, S.K.; HWANG, J.Y.; YOON, J. Clinical Impact of Atypical Chest Pain and Diabetes Mellitus in Patients with Acute Myocardial Infarction from Prospective KAMIR-NIH Registry. **Journal of Clinical Medicine, v.** 9, n. 2, p. 505, 2020. Doi: 10.3390/jcm9020505

MARTINEZ, I.; VILES, E.; OLAIZOLA, I. G. Data Science Methodologies: Current Challenges and Future Approaches. **Big Data Research**. v. 24, 100183, 2021. doi:10.1016/j.bdr.2020.100183.

MCCANCE, D.R. Pregnancy and diabetes. **Best Practice & Research Clinical Endocrinology & Metabolism,** v. 25, n. 6, p. 945-58, 2011. doi: 10.1016/j.beem.2011.07.009.

MEIRELLES, W. C. L.; ZÁRATE, L. E. Data mining in the reduction of the number of places of experiments for plant cultivates, **Computers and Electronics in Agriculture**, v. 113, p. 136-147, 2015. DOI: 10.1016/j.compag.2015.02.006.

NONAKA, I. **The knowledge creating company**. Harvard Business Review, 69, (Nov-Dec), p. 96-104, 1991.

NONAKA, I.; TAKEUCHI, H. **The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation.** Oxford University Press, New York. 1995. Disponível em: <https://id.lib.harvard.edu/alma/990052925990203941/catalog>

PASTOR, A.; CONN, J.; MACISAAC, R. J.; BONOMO, Y. Alcohol and illicit drug use in people with diabetes. **Lancet Diabetes & Endocrinology**, v. 8, n. 3, p. 239-248, 2020. Doi: 10.1016/S2213-8587(19)30410-3

POLANYI, M. The tacit dimension. London: Routledge and Kegan Paul, 1967.

POLSKY, S.; AKTURK, H. K. (2017). Alcohol Consumption, Diabetes Risk, and Cardiovascular Disease Within Diabetes**. Curr Diab Rep.** v. 17, n. 12, p. 136, 2017. Doi: 10.1007/s11892-017-0950-8

RIBEIRO, C. E.; ZÁRATE, L. E. Classifying longevity profiles through longitudinal data mining, **Expert Systems with Applications**, v. 117, p. 75-89, 2019. DOI: 10.1016/j.eswa.2018.09.035

SILVA, P. R.; DIAS, S. M.; BRANDÃO, W. C.; SONG, M. A.; ZÁRATE, L. E. Professional Competence Identification Through Formal Concept Analysis. In Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017), v. 1, p. 123-134, 2018. DOI: 10.5220/0006333401230134

VON KROGH, G.; ROOS, J. **Organizational Epistemology**. New York, NY: St. Martin's Press. 1995. DOI: 10.1007/978-1-349-24034-0.

WIIG, K. M. **Knowledge Management Foundations: Thinking about Thinking: How People and Organizations Create**, Represent and Use Knowledge, 1993, 471 p.

ZHANG, X.; BULLARD, K. M.; GREGG, E. W.; BECKLES, G. L.; WILLIAMS, D. E.; BARKER, L.E.; ALBRIGHT, A.L. Imperatore G. Access to health care and control of ABCs of diabetes. **Diabetes Care**. v. 35, n.7, p. 1566-71, 2012. doi: 10.2337/dc12-0081.