UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO
EM ENGENHARIA ELÉTRICA

Ciência de Dados no Diagnóstico de Intoxicação por Agrotóxicos em Trabalhadores Rurais usando Teoria de Conjuntos Aproximados

Jaqueline Corrêa Silva de Carvalho

Julho de 2024
Itajubá - MG

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO
EM ENGENHARIA ELÉTRICA

Ciência de Dados no Diagnóstico de Intoxicação por Agrotóxicos em Trabalhadores Rurais usando Teoria de Conjuntos Aproximados

Jaqueline Corrêa Silva de Carvalho

Tese submetida ao Programa de Pós-Graduação em Engenharia Elétrica como parte dos requisitos para obtenção do título de **Doutor em Ciências em Engenharia Elétrica.**

Área de concentração: Microeletrônica

Orientador: Prof. Dr. Tales Cleber Pimenta
Coorientadora: Profa. Dra. Alessandra Cristina Pupin Silvério

Julho de 2024
Itajubá - MG

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO
EM ENGENHARIA ELÉTRICA

Ciência de Dados no Diagnóstico de Intoxicação por Agrotóxicos em Trabalhadores Rurais usando Teoria de Conjuntos Aproximados

Jaqueline Corrêa Silva de Carvalho

Banca Examinadora:

Prof. Dr. Tales Cleber Pimenta, UNIFEI (Orientador)
Profa. Dra. Alessandra C. Pupin Silvério, UNIFENAS (Coorientadora)
Prof. Dr. Luiz Eduardo da Silva, UNIFAL
Prof. Dr. Erivelton Antonio dos Santos, UFLA
Prof. Dr. Gabriel Antonio Fanelli de Souza, UNIFEI
Prof. Dr. Bruno Tardiole Kuehne, UNIFEI

Julho de 2024
Itajubá - MG

À Jesus que sempre me diz: "Tenhais paz em mim. No mundo, tereis tribulações. Mas, tende coragem! Eu venci o mundo!"

Agradecimentos

A realização desta tese de doutorado foi uma experiência desafiadora e gratificante, que não teria sido possível sem a providência de Deus, o apoio e a colaboração de várias pessoas e instituições, às quais devo expressar minha profunda gratidão.

Foi pela providência de Deus que conheci o Professor Tales Cleber Pimenta, que me motivou a iniciar o processo de ingresso no doutorado. No decorrer das disciplinas cursadas, ele se tornou meu orientador, oferecendo orientação inestimável e muita paciência. Considero também como providência de Deus minha co-orientadora, Professora Alessandra Cristina Pupin Silvério, que gentilmente disponibilizou os dados de sua pesquisa e ofereceu suporte crucial para o desenvolvimento deste trabalho. A eles, expresso minha mais sincera gratidão.

Ao meu marido, Marcos Alberto de Carvalho, agradeço pelo amor, incentivo, paciência e disposição em me auxiliar em todas as atividades realizadas durante esse período. Aos meus filhos, João Paulo, Germano e Letícia, exemplos para mim de determinação e conquista, agradeço pela compreensão e pelo incentivo constante nos meus estudos. Aos meus pais Geraldo (in memória) e Marilene, agradeço por serem um exemplo de fé e coragem para enfrentar os desafios.

Gostaria de expressar minha gratidão à UNIFEI, que proporcionou o ambiente acadêmico necessário para a realização desta pesquisa. Agradeço de forma especial ao corpo docente, ao coordenador e aos funcionários da pós-graduação pelo suporte contínuo, que foram essenciais para a concretização desta pesquisa.

Sem o apoio e a colaboração de cada um de vocês, este trabalho não teria sido possível. Muito obrigado!

Resumo

Num projeto de Ciência de Dados, é essencial determinar a relevância dos dados e identificar padrões que contribuam para a tomada de decisões com base no conhecimento específico do domínio. Além disso, uma definição clara das metodologias e a criação de documentação para orientar o desenvolvimento de um projeto desde o início até a conclusão são elementos essenciais. Este estudo apresenta um modelo de Ciência de Dados projetado para orientar o processo, abrangendo desde a coleta de dados até o treinamento com o objetivo de facilitar a descoberta de conhecimento. Motivado por deficiências em metodologias existentes de Ciência de Dados, especialmente a falta de orientação prática passo a passo sobre como preparar os dados para alcançar a fase de produção. Chamado de "Ciclo de Refinamento de Dados com a Teoria de Conjuntos Aproximados (CRD–TCA)", o modelo proposto foi desenvolvido com base nas necessidades emergentes de um projeto de Ciência de Dados com o objetivo de auxiliar profissionais de saúde no diagnóstico de intoxicação por agrotóxicos entre trabalhadores rurais. O conjunto de dados usado neste projeto resultou de uma pesquisa científica na qual foram coletadas 1027 amostras, contendo dados relacionados a biomarcadores de toxicidade e análises clínicas. Alcançamos uma precisão de 99,61% com apenas 27 regras para determinar o diagnóstico. Os resultados otimizaram práticas de saúde e melhoraram a qualidade de vida em áreas rurais. Os resultados do projeto demonstraram o sucesso do modelo proposto.

Abstract

In a Data Science project, it is essential to determine the relevance of the data and identify patterns that contribute to decision–making based on domain–specific knowledge. Furthermore, a clear definition of methodologies and creation of documentation to guide a project's development from inception to completion are essential elements. This study presents a Data Science model designed to guide the process, covering data collection through training with the aim of facilitating knowledge discovery. Motivated by deficiencies in existing Data Science methodologies, particularly the lack of practical step–by–step guidance on how to prepare data to reach the production phase. Named "Data Refinement Cycle with Rough Set Theory (DRC–RST)", the proposed model was developed based on the emerging needs of a Data Sciense project aimed at assisting healthcare professionals in diagnosing pesticide poisoning among rural workers. The dataset used in this project resulted from scientific research in which 1027 samples were collected, containing data related to toxicity biomarkers and clinical analyses. We achieved an accuracy of 99.61% with only 27 rules for determining the diagnosis. The results optimized healthcare practices and improved quality of life in rural areas. The project outcomes demonstrated the success of the proposed model.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A utilização de agrotóxicos no Brasil aumentou exponencialmente nas últimas décadas. Segundo o Ministério da Agricultura, Pecuária e Abastecimento, foram aprovados entre 2000 e 2015 uma média de 130 agrotóxicos por ano. Esta média subiu para 419 agrotóxicos por ano de 2016 a 2020 e, entre 2021 e 2022, esta média subiu para 602 agrotóxicos por ano [1]. Isso se deve principalmente ao controle de pragas como insetos, plantas daninhas, entre outros, e para o aumento da produtividade. O Brasil, hoje, é o quarto maior produtor de alimentos do mundo e o terceiro maior exportador, atrás apenas da Europa e dos Estados Unidos [2]. Apesar da necessidade do uso de agrotóxicos, duras realidades precisam ser enfrentadas, como o crescimento nos quadros de intoxicação provenientes da exposição a essas substâncias, principalmente em trabalhadores rurais [3] [4].

Do ponto de vista clínico, a intoxicação por agrotóxico pode se apresentar de duas formas clássicas: aguda e crônica. A intoxicação aguda acontece quando a exposição ao agente tóxico é excessiva e por curto período de tempo. Os danos agudos são bem descritos e podem ser diagnosticados caso os exames necessários sejam prontamente realizados. A intoxicação crônica caracteriza-se pelo surgimento tardio, por exposição prolongada durante meses ou anos a pequenas ou médias quantidades de agente tóxico. Os danos crônicos podem se apresentar na forma de doenças de pele, carcinogêneses, desregulações endócrinas, neurotoxicidades, efeitos na reprodução humana e no sistema imunológico, nefrotóxicos, citotóxicos, ocasionando paralisias e neoplasias [4][5].

De acordo com o último levantamento oficial do Sistema Nacional de Informações Tóxico-Farmacológicas (SINITOX), entre 2007 e 2017, foram notificados cerca de 40.000 casos de intoxicação aguda por agrotóxicos. Hoje, o SINITOX informa que não possui dados estatísticos reais em virtude da diminuição da participação dos Centros de Informação e Assistência Toxicológica (CIATs) nestes levantamentos [6].

Segundo a Organização Internacional do Trabalho (OIT), os agrotóxicos causam anualmente 170 mil intoxicações agudas e crônicas que evoluem para óbito. Esta estimativa sugere que os trabalhadores rurais têm duas vezes mais risco de enfermidades quando comparados a outros setores, como mineração e construção civil [7]!

De acordo com Peña-Fernández et al.[8], a intoxicação por agrotóxicos em trabalhadores rurais é uma questão de grande preocupação e resulta em significativas perdas sociais e econômicas, sendo um problema mundial. Isso se deve à falta de testes padronizados para diagnóstico biológico e à escassez de profissionais de saúde treinados para lidar com tais casos.

Existe um esforço da comunidade científica em coletar dados provenientes de registros médicos, visando realizar análises desses dados, a fim de auxiliar na atuação de profissionais da saúde, que podem ser especialistas em intoxicação por agrotóxicos. Essas análises geralmente são realizadas por cientistas de dados.

Um projeto de Ciência de Dados é centrado em dados que geralmente estão inseridos em um contexto específico. Muitos problemas, como análise financeira, marketing e análise de saúde, têm se beneficiado de projetos de Ciência de Dados [9].

A Ciência de Dados é um processo de busca por padrões ocultos e desconhecidos em conjuntos de dados que guardam um histórico de informações, procurando verificar relacionamento entre os dados e descobrir conhecimento que auxilie na tomada de decisões futuras. A maioria dos projetos de Ciência de Dados pode envolver uma ou mais abordagens da área de Inteligência Artificial para a descoberta de conhecimento de dados. A colaboração de especialistas do domínio dos dados em todo o processo é de suma importância para atingir as metas propostas [10].

Apesar do notável crescimento no campo da Ciência de Dados, a execução bem-sucedida de projetos nesse domínio continua a apresentar desafios significativos. De acordo com informações de Saltz et al.[11], a maioria dos projetos de Ciência de Dados não consegue chegar à fase de produção.

Os modelos de Ciência de Dados frequentemente seguem estruturas cíclicas conhecidas como ciclo de vida dos dados, que servem como guia para os cientistas de dados ao longo do processo de descoberta de conhecimento de dados, *Knowledge Discovery Data* (KDD). Um desafio recorrente neste contexto é a falta de interpretabilidade em modelos complexos, bem como a presença de dados de baixa qualidade ou ruidosos, que podem comprometer a eficácia e confiabilidade dos modelos [10].

Bie et al. [12] descobriram que os métodos de aprendizado de máquina desempenham um papel predominante na caixa de ferramentas de um cientista de dados. Esses métodos ganharam destaque nas últimas duas décadas, abrangendo desde técnicas relativamente simples até abordagens complexas, como o aprendizado profundo. No entanto, é essencial enfatizar que tais métodos frequentemente pressupõem a disponibilidade de volumes substanciais de dados de alta qualidade, o que, na prática, apresenta desafios adicionais.

A Teoria dos Conjuntos Aproximados (TCA), proposta por Pawlak [13] e revisada por Acharjya e Abraham [14], representa uma ferramenta matemática valiosa para lidar com um tipo específico de incerteza e imprecisão. A TCA, seja adotada sozinha ou em conjunto com outros modelos de aprendizado de máquina, demonstrou sua eficácia na resolução de problemas de aprendizado de máquina do mundo real.

Portanto, esse trabalho propõe um modelo de Ciência de Dados, com base na Teoria de Conjuntos Aproximados como metodologia de Inteligência Artificial aplicada em um conjunto de dados de saúde. Esses dados são resultantes de uma pesquisa científica realizada com 1027 agricultores, onde foram coletadas informações como tempo de uso de agrotóxicos, tipo de agrotóxico utilizado, sintomas clínicos e resultados de exames laboratoriais, entre outros dados necessários para que o especialista fornecesse o diagnóstico de intoxicação por agrotóxicos.

1.1 Descrição do problema

A Toxicologia é a ciência que investiga os efeitos nocivos de substâncias químicas no organismo, sob condições de exposição. No Brasil, a Toxicologia é reconhecida como uma das 57 áreas de atuação médica, conforme resolução Nº 2.149, publicada em 2016 pelo Conselho Federal de Medicina. Mesmo diante do reconhecimento, muitas formações acadêmicas ainda não oferecem de forma relevante, o conteúdo de Toxicologia. Alguns conhecimentos de Toxicologia, muitas vezes, são diluidos em outras disciplinas como Clínica Médica e Emergência Médica. Diante desse cenário, o profissional da saúde apresenta certa insegurança diante de quadros de intoxicação, principalmente das intoxicações crônicas [15].

As notificações referentes a intoxicação em trabalhadores agrícolas, se tornam cada vez mais crescentes, com o aumento do uso de agrotóxicos, necessitando de atendimento e acompanhamento preventivo especializado.

A Ciencia de Dados é a ciência que investiga os dados e procura por padrões que forneçam informações que possam resolver problemas, tal como, a ineficiência na avaliação diagnóstica por intoxicação, porém, segundo a conferência *Transform* 2019, 87% dos projetos de Ciência de Dados nunca chegam a produção [11] [16].Torna se fundamental a elaboração e/ou adoção de metodologias que guiem o desenvolvimento de um projeto de Ciência de Dados, do início ao fim. Existem propostas de vários modelos que acompanham o ciclo de vida dos dados de um projeto de Ciência de Dados, porém de uma maneira geral, os modelos carecem de maior detalhamento e formalismo [17][18][19].

Assim, o objetivo inicial de auxiliar no diagnóstico de um problema de saúde pública evoluiu para um novo propósito: desenvolver um modelo prático de Ciência de Dados capaz de lidar com qualquer problema de aprendizado de máquina supervisionado. A motivação para a proposta do modelo reside na necessidade premente de abordagens práticas e direcionadas para a preparação de dados, preenchendo uma lacuna crucial na gestão de projetos de Ciência de Dados. Este modelo não apenas visa abordar as deficiências identificadas, mas também aumentar a eficácia e transparência em projetos de Ciência de Dados, fornecendo uma abordagem sistemática e transparente ao longo do ciclo de vida do projeto. O diagnóstico de intoxicação por agrotóxicos em trabalhadores rurais tornou-se um exemplo usado para demonstrar este modelo.

1.2 Objetivos

Esse trabalho possui dois objetivos, o primeiro objetivo é desenvolver um novo modelo de Ciência de Dados que utiliza a Teoria de Conjuntos Aproximados como ferramenta de aprendizado de máquina, dado que grande parte dos projetos em Ciência de Dados não são finalizados. Este modelo é denominado "Ciclo de Refinamento de Dados com Teoria de Conjuntos Aproximados (CRD—TCA)". Ele se destaca pela capacidade de analisar e monitorar meticulosamente cada informação presente no conjunto de dados por meio de formulários padronizados, possibilitando a categorização de dados com múltiplos valores ou valores ausentes. Além disso, o

CRD-–TCA propõe a documentação detalhada de cada etapa do processo, incluindo todas as sessões de treinamento realizadas, com o objetivo de simplificar a verificação do progresso e o acompanhamento das tarefas. Essa abordagem promove uma metodologia mais transparente e rastreável, facilitando a análise retrospectiva das sessões de treinamento documentadas.

O segundo objetivo é aplicar o modelo num projeto real de Ciência de Dados, como uma prova de conceito (PoC) e demonstração do modelo proposto. Para demonstrar a aplicação e a eficácia do modelo CRD—TCA, este trabalho apresenta o projeto de Ciência de Dados "Plantando e Colhendo Saúde (PCS)". O objetivo deste projeto é auxiliar no diagnóstico de intoxicação por agrotóxicos em trabalhadores rurais, utilizando o modelo CRD–TCA. O projeto PCS não só contribuirá significativamente para os profissionais de saúde, fornecendo uma ferramenta para diagnóstico mais preciso, mas também servirá como um cenário prático para a validação do novo modelo de Ciência de Dados desenvolvido. A aplicação do CRD-–TCA no contexto do projeto PCS permitirá avaliar sua eficácia e utilidade em um ambiente real, demonstrando seu potencial de impacto na área da saúde.

Para atingir os objetivos mencionados, foram definidos os seguintes objetivos específicos:

- Conhecer de forma abrangente o problema de Diagnóstico de Intoxicação por Agrotóxicos, escolhido para a aplicação do modelo que será proposto.
- Apresentar a Ciência de Dados e analisar os modelos de ciência de dados existentes.
- Demonstrar os conceitos básicos da Teoria de Conjuntos Aproximados e expor, sob o enfoque desta teoria, os Sistemas de Representação de Conhecimento.
- Descrever a proposta original do modelo CRD–TCA, demonstrando sua aplicação no projeto PCS.
- Validar a utilização do modelo CRD–TCA, apresentando os resultados do projeto PCS.

1.3 Hipóteses

Um projeto de Ciência de Dados que possui fontes de dados cientificamente comprovadas, um modelo de ciência de dados adequado, com a formatação de documentos necessários e usando a Teoria de Conjuntos Aproximados na etapa de aprendizado de máquina, vai garantir, a simplicidade e a conclusão desse projeto, conseguindo descobrir dados relevantes, no conjunto dos dados, e estabelecer regras que representem conhecimento.

1.4 Estrutura do trabalho

O trabalho está dividido em cinco capítulos. O primeiro capítulo oferece uma visão geral, objetivos e hipóteses da pesquisa. O capítulo 2 apresenta uma fundamentação teórica em agrotóxicos, incluindo um estudo de caso com 1027 indivíduos, explora a Ciência de Dados e

a metodologia de aprendizado de máquina supervisionada - Teoria de Conjuntos Aproximados. O capítulo 3 apresenta o modelo proposto. O capítulo 4 discute os resultados experimentais, enquanto o capítulo 5 destaca as conclusões finais da pesquisa.

2 Fundamentação teórica

2.1 Agrotóxicos

2.1.1 Introdução

Durante a segunda guerra mundial foram utilizadas substâncias químicas, de ação biocida, como armas. Com o final da segunda guerra mundial, alguns desses estudos, se voltaram para a eliminação de agentes causadores de danos à vegetação [20]. O termo agrotóxico tem origem do grego: agros (campo) + tokicon (veneno). Inclui todos os produtos de natureza tóxica usados na agricultura para o manejo de pragas [21]. Os agrotóxicos muitas vezes são chamados de pesticidas, defensivos agrícolas, biocidas e outras nomenclaturas. São considerados um marco no progresso agrícola, formulados cientificamente e utilizados para garantir o aumento de produtividade através do controle de pragas [22]. As intoxicações por agrotóxicos, representam um sério problema de saúde pública no mundo. No Brasil é ainda mais alarmante, pois o país é um dos maiores consumidores de agrotóxicos do mundo [20].

2.1.2 Classificação dos Agrotóxicos

Um dos critérios para a classificação dos agrotóxicos são os alvos preferenciais sobre os quais atuam, sendo os principais grupos:

2.1.2.1 Inseticidas

Os inseticidas são compostos por substancias químicas sintéticas para o combate de insetos, larvas e formigas, podendo resultar na morte do inseto ou prevenir comportamentos considerados destrutivos [22]. Os inseticidas pertencem a grupos químicos de acordo com as substâncias ativas presentes em sua formulação:

- Organoclorados: representam um grupo relativamente grande de inseticidas. São compostos a base de carbono e radicais de cloro. Característica Tóxica: São persistentes ao ambiente e apresentam forte tendência de bioacumulação nos tecidos gordurosos dos organismos. Estudos mostram a possível relação entre a exposição à agrotóxicos organoclorados e o surgimento de alguns tipos de câncer, como os cânceres de mama, estomago, pulmão, pâncreas e próstata. Têm sido progressivamente restringidos ou mesmo proibido [22].

- Organofosforados: apresentam em sua estrutura molecular o fósforo como átomo central, podendo variar suas ligações com átomos de oxigênio, enxofre, alquil, amino, tioalquil, fenil ou outros grupos substituintes, sendo compostos bem caracterizados a partir da natureza dos átomos circundantes ao átomo central de fósforo. Característica Tóxica: Possuem

alta toxicidade, podendo contaminar os organismos através de ambientes contaminados, como água, ar, solo e alimentos. Podem ser absorvidos pelas vias dérmica, respiratória e oral, causando problemas respiratórios como asma e doença pulmonar, digestivos, hepáticos, desregulações hormonais, alterações no sistema nervoso central, mal de Parkinson, alterações no neurodesenvolvimento pré e pós natal, doenças hematológicas, doenças cardiovasculares, diabetes, e diversos tipos de câncer estão associados ao uso de agrotóxicos organofosforados [4].

- Carbamatos: Os carbamatos podem ser usados como inseticidas e fungicidas. São derivados do ácido carbâmico. Atualmente são usados em grande escala e são conhecidos mais de cinquenta tipos de carbamatos. Característica Tóxica: Possuem baixa toxicidade em longo prazo e baixa ação residual [22].

- Piretróides: Inseticidas de origem vegetal, obtido a partir do piretro que é extraído da trituração de flores, principalmente, das espécies Chysanthemum cineraiaefolium e espécies relacionadas com o Chysanthemum coccineum. Atualmente, é feito a inclusão de átomos de nitrogênio, enxofre e átomos halogênios às piretrinas. Possui ação eficiente sobre uma variedade de insetos. Característica Tóxica: Possuem menor toxicidade aos mamíferos quando comparado aos organoclorados, organofosforados e carbamatos. Porém, sabe-se que piretrinas e piretroides são substancias que causam alergias respiratórias nos mamíferos e podem exercer efeitos neurotóxicos e cardiotóxicos nos vertebrados [22].

- Neonicotinoide: são compostos derivados da nicotina e agem em grande variedade de insetos. Característica Tóxica: são potenciais carcinogénios humanos; podendo ser tóxicos também para as aves, peixes, e muitos neonicotinoides são tóxicos para as abelhas [23].

2.1.2.2 Fungicidas

Os fungicidas são usados para combater fungos em geral. Os fungicidas se dividem de acordo com substancia ativa em estrobirulinas, que inibem a respiração mitocondrial de fungos e triazóis, que provocam a ruptura da parede celular dos fungos. Quase nada é conhecido sobre a ecotoxicidade dos fungicidas estrobirulinas e pouco se sabe a respeito da ecotoxicidade dos fungicidas triazóis [22].

2.1.2.3 Herbicidas

Os herbicidas combatem o desenvolvimento de ervas daninhas. O princípio ativo mais usado para as formulações de herbicidas é o glifosato do grupo químico glicina substituída.

- Glicina Substituída: O glifosato [N–(fosfonometil)glicina, $C_3H_8NO_5P$] é um organofosfato que se subdivide em três tipos glifosatoisopropilamônio, glifosato-sesquisódio, e glifosato-trimesium. Características Tóxica: O ingrediente ativo glifosato segunda a

Organização Mundial da Saúde em 2015, pode causar câncer em animais tratados em laboratório. Podendo ser causador de alterações na estrutura do DNA e nas estruturas cromossômicas das células humanas. Estudos realizados por pesquisadores brasileiros apontam o glifosato como desregulador endócrino em células hepáticas humanas e induz a proliferação de células humanas de câncer de mama e danos ao sistema gastrointestinal, rins e fígado [24].

2.1.3 Anticolinesterásicos

A fisostigmina, é um alcaloide obtido da fava de Calabar, a semente madura seca de Physostigma venenosum, uma planta da África Ocidental tropical. Da fava de Calabar extraiu-se um alcaloide puro denominado fisostigmina. O primeiro uso terapêutico da droga foi em 1877, no tratamento do glaucoma.

Em 1952, essa substancia destacou-se quanto ao seu potencial inseticida. Antes da Segunda Guerra Mundial e durante ela, os esforços da equipe de pesquisa da fisostigmina foram dirigidos para o desenvolvimento de compostos para guerra química.

Compostos químicos a base de fisiostigmina e derivados são denominados anticolinesterásicos. Isso, porque, são substâncias que bloqueiam a ação da enzima Acetilcolinesterase (AChE). Essa enzima é responsável por hidrolisar o neurotransmissor acetilcolina (ACh) nas sinapses colinérgicas. Nestas sinapses a ACh atua transmitindo a mensagem de um neurônio a outro, sendo importante para a manutenção de inúmeras funções fisiológicas humanas. Com a inibição da (AChE), ocorre um acúmulo de acetilcolina na fenda sináptica, potencializando os efeitos Parassimpáticos. O efeito farmacológico se deve a ação da acetilcolina, durante mais tempo nos seus receptores [25][26].

Os Organofosforados e carbamatos também são classificados como anticolinesterásicos.

2.1.4 Intoxicação

Quando o indivíduo entra em contato com o elemento químico, contido no agrotóxico, esse pode ou não, ser absorvido pelo organismo (seja via oral, cutânea ou pulmonar) podendo causar somente um efeito local ou dar origem ao processo de biotransformação e intoxicação. A intoxicação corresponde a uma gama de ações e reações do organismo, desde a exposição ao agente tóxico, até sua excreção por bioativação e detoxificação, onde os compostos tóxicos são quebrados ou transformados em outros compostos, menos ou mais tóxicos que o original, podendo ser armazenados no sangue, fígado, rins ou outros órgãos, ou preparados para eliminação urinária conforme a Figura 2.1 [4].

No processo de biotransformação, os agrotóxicos sofrem ação de enzimas produzidas pelo organismo como as A–esterase, colinesterases (ch) e outras enzimas. A atividade de algumas dessas enzimas é utilizada como bioindicadores de efeito tóxico a médio, curto e longo prazo, podendo ser medidas através de exames realizados em laboratório [4].

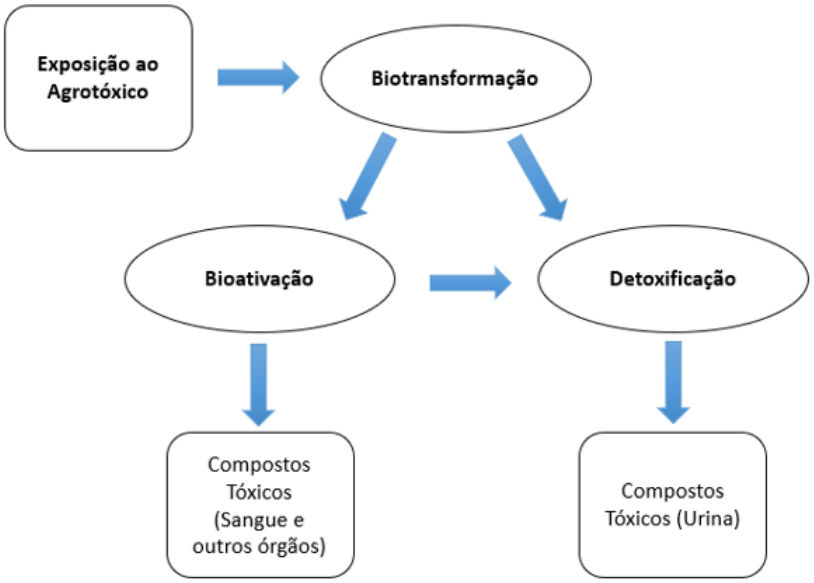

Figura 2.1 – Processo de biotransformação [4].

Contudo somente a análise dos bioindicadores não é suficiente para o monitoramento de intoxicação de pessoas expostas aos agrotóxicos, pois podem apresentar algumas limitações de acordo com cada indivíduo. Assim se faz necessário realizar um acompanhamento quanto a idade, sexo, estado nutricional, presença de patologias e tipos de agrotóxicos utilizados que podem alterar a atividade das enzimas a serem analisadas [4].

Estudos com bioindicadores específicos, com células provenientes da mucosa bucal e amostras de sangue periférico, demonstraram também que a presença de alguns compostos químicos no organismo, principalmente os da classe dos organofosforados, podem aumentar o aparecimento de mutações do DNA das células. Estas mutações podem dar origem aos processos farmacocinéticos e intoxicação. Os processos farmacocinéticos como absorção, distribuição, biotransformação e eliminação, determinam a concentração no sítio de ação e o grau da intoxicação [4].

2.1.5 Estudo de Caso

Uma avaliação clínica em trabalhadores rurais associada aos bioindicadores de toxicidade foi realizada na regional de saúde de Alfenas em Minas Gerais -– Brasil, constituída de 26 municípios. O estado de Minas Gerais é o maior estado produtor de café do Brasil, respondendo por 50% da produção nacional. Segundo dados do IBGE, em 2010 essa regional possuía uma população de 66266 moradores rurais em idade produtiva. O tamanho da amostra calculada foi de 1038 pessoas escolhidas aleatoriamente. Porém, ao final, o estudo foi realizado com uma população de 1027 indivíduos que se prontificaram a colaborar [4]. Todos foram submetidos a triagem clínica e análise dos bioindicadores de toxicidade como:

- Determinação da atividade das enzimas colinesterases (Ch), em sangue coletado em tubo

com heparina.

- Determinações da atividade das enzimas AST, ALT e γ-GT, para avaliação hepática e determinação da creatinina sérica para avaliação renal. A amostra de soro foi coletada em sistema de vácuo sendo a avaliação realizada por meio de kit da marca Labtest [4].

Dentro das 1027 amostras, como visto na Figura 2.2, foram selecionados três grupos visando uma análise comparativa e para uma avaliação mais específica, com exames de disfunções urinária (DAPs urinário) e genotoxicidade em amostras da mucosa bucal e sangue periférico (Ensaio do Citoma, Ensaio do Cometa e Análise Polimórfica):

- 94 indivíduos expostos à mistura complexa de agrotóxicos, contendo compostos organofosforados.
- 94 indivíduos expostos a mistura complexa de agrotóxicos, sem compostos organofosforados.
- 50 indivíduos não expostos de forma ocupacional aos agrotóxicos, moradores da zona urbana [4].

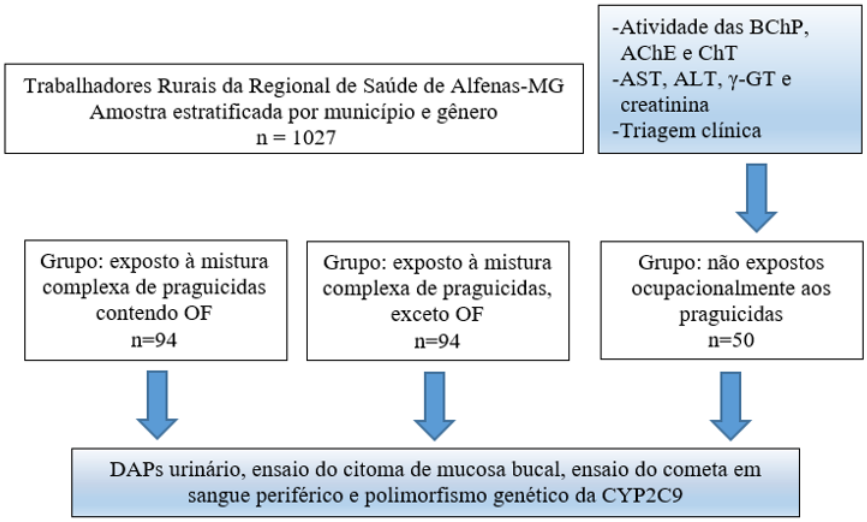

Figura 2.2 – Distribuição de amostras [4].

O instrumento utilizado para levantamento de dados epidemiológicos e clínicos dos trabalhadores rurais, foi obtido junto a Universidade Estadual de Campinas – UNICAMP e modificado pelos pesquisadores no qual foi realizado o processo de Refinamento por Juízes e também um teste piloto com 50 trabalhadores rurais do município de Alfenas–MG. Mediante estas intervenções foram realizadas mudanças e inserido questões neste instrumento para melhor avaliação clínica e coleta dos dados de exposição. A aplicação do instrumento e avaliação clínica foram realizados por médicos e uma equipe de aplicadores treinados para tal finalidade [4].

2.1.5.1 Resultados do Estudo de Caso na Amostra Total (n = 1027)

A Figura 2.3 apresenta a seleção de indivíduos por cidade. Em nenhum dos 26 municípios visitados, era de conhecimento das equipes de saúde da zona rural, a importância clínica do monitoramento dos trabalhadores rurais que têm contato diário com agrotóxicos por meio das dosagens dos bioindicadores das enzimas colinesterases (BChP, AChE e ChT). De acordo com o questionário, a agricultura familiar foi predominante em 74,3% do total de trabalhadores rurais entrevistados. Como apresentado na Figura 2.4, a falta da utilização adequada de equipamento de proteção individual EPIs e o contato intenso e prolongado com agrotóxicos foi constatado, sendo que muitos dos trabalhadores relataram ter mais do que 20 anos de exposição.

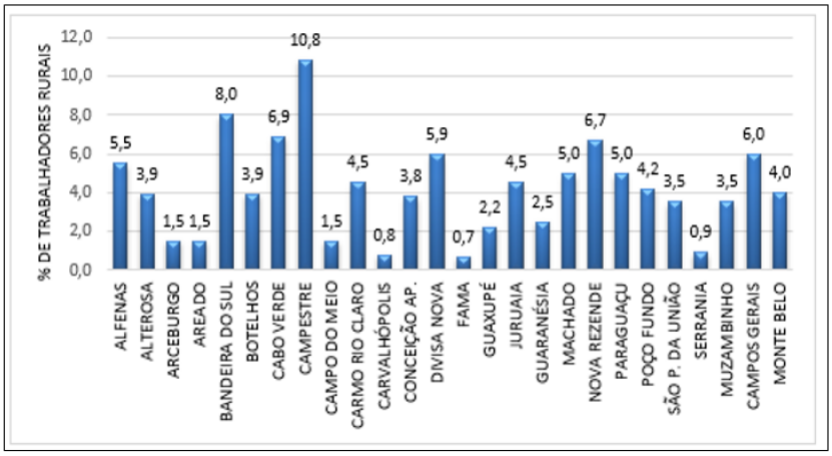

Figura 2.3 – Amostras de trabalhadores rurais por cidade [4].

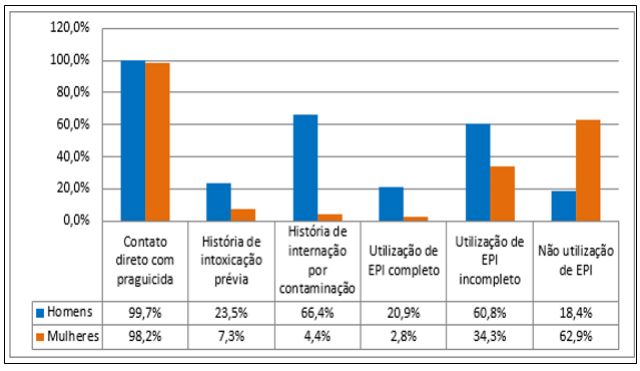

Figura 2.4 – Histograma das análises das amostras [4].

A Tabela 1 apresenta a atividade das enzimas Ch, avaliação hepática e renal na população estudada.

A Tabela 2 apresenta a avaliação clínica nos 1027 indivíduos, trabalhadores da zona rural, por meio de respostas ao questionário no qual foram submetidos.

Tabela 1 – Atividade das enzimas Ch [4].

Resultado Alterado	População (n=1027)
AChE	33,6%
BChP	5,7%
Função Hepática (AST e ALT)	15%
Função Renal (Creatinina)	3,5%

Tabela 2 – Avaliação clínica [4].

Sistema Alterado	População (n=1027)
Cardiovascular	45,0%
Sistema Nervoso Central	75,5%
Digestivo	54,6%
Respiratório	54,6%
Auditivo	52,0%
Membranas/Pele e Mucosa	37,6%
Urinário	21,8%

2.1.5.2 Resultados do Estudo de Caso por Grupos

Como já apresentado, foi selecionado três grupos da população dos 1027 indivíduos para uma avaliação de comparação entre esses grupos. Nessa seleção procurou-se escolher para os mesmos grupos, indivíduos que apresentavam várias características em comum, destacando o uso ou não de agrotóxicos organofosforados.

A Tabela 3 mostra as porcentagens de alteração em cada sistema do organismo. Observa-se alterações significativas no sistema nervoso central, respiratório, auditivo e nas membranas, pele e mucosas, nos grupos expostos a agrotóxicos.

Tabela 3 – Sistemas alterados por organismos [4].

Sistemas Alterados	Expostos a agrotóxicos contendo organofosforados (n=94)	Expostos a agrotóxicos exceto organofosforados (n=94)	Não expostos ocupacionalmente aos agrotóxicos (n=50)
Cardiovascular	36,5%	31,2%	28,2%
Sistema Nervoso Central	80,0%	77,4%	57,1%
Digestivo	59,3%	57,0%	40,5%
Respiratório	56,5%	46,2%	19,0%
Auditivo	38,8%	39,6%	7,1%
Membranas, pele e mucosas	31,3%	36,7%	0%
Urinário	20,9%	16,1%	7,1%

A avaliação usando bioindicadores foi realizada nos três grupos de comparação e os resultados são apresentados na Tabela 4.

Tabela 4 – Resultados de atividade das Ch e concentrações de dietilfosfato (DETP) e dietilditiofosfato (DEDTP) nos três grupos analisados [4].

Bioindicadores (% Alterada)	**Expostos a agrotóxicos contendo organofosforados (n=94)**	**Expostos a agrotóxicos exceto organofosforados (n=94)**	**Não expostos ocupacionalmente aos agrotóxicos (n=50)**
Atividade da Colinesterase Plasmática (ChP)	12,8%	-	-
Atividade da Colinesterase Eritrocitária (ChE)	63,8%	-	-
Atividade da Colinesterase Total (ChT)	14,8%	-	-
DETP	92,6%	26,5%	1,1%
DEDTP	91,5%	25,5%	1,1%

Características da atividade das enzimas colinesterases:
BChP:

- Apresenta meia-vida de uma semana
- Bioindicador que reflete exposições recentes e a dose absorvida dos OF.

AChE:

- Bioindicador de efeito e exposições mais prolongadas.
- A meia–vida da AChE depende da meia–vida da hemácia que é de cerca de 120 dias, reflete melhor as situações de exposição crônica e efeito dos anticolinesterásicos.
- Vários estudos relataram relação significativa entre a exposição crônica e inibição da AChE em populações de trabalhadores expostos.

Os efeitos tóxicos também podem ser avaliados através do aumento dos níveis de DAPs urinários. Este aumento foi verificado em mais de 90% do grupo exposto à compostos organofosforados, conforme mostra a tabela 4 nas concentrações de dietilfosfato – DETP e dietilditiofosfato – DEDTP. O estudo comparativo das análises hepática e renal não apresentaram diferenças significativas, nos voluntários dos grupos.

Como os agrotóxicos organofosforados são descritos na literatura como responsáveis pelo aparecimento de diversos tipos de tumores, foi realizado entre os grupos uma análise de células da mucosa bucal e os resultados estão apresentados na Tabela 5.

Esta análise apresentou significativas alterações celulares em trabalhadores expostos aos agrotóxicos organofosforados, comparado ao grupo dos trabalhadores expostos a outros agrotóxicos. Os resultados dos testes do ensaio do cometa para verificar alterações celulares, mostraram que a exposição aos agrotóxicos seja eles organofosforados ou não, aumentam o nível de danos ao DNA da célula, indicando maior suscetibilidade ao desenvolvimento de tumores. A presença de genótipos específicos na avaliação polimórfica de genótipos celulares também se mostrou alterada para os grupos expostos a agrotóxicos organofosforados ou não [4].

Tabela 5 – Resultados do teste de ensaio do citoma das células de mucosa bucal para os 3 grupos avaliados [4].

Presença de alterações celulares	**Expostos a agrotóxicos contendo organofosforados (n=94)**	**Expostos a agrotóxicos exceto organofosforados (n=94)**	**Não expostos ocupacionalmente aos agrotóxicos (n=50)**
Broto Nuclear	1	0	0
Binucleada	5	4	1
Cromatina Condensada	1	0	0
Cariolítica	300	201	110

2.1.5.3 Conclusão do Estudo de Caso

A avaliação clínica associada a determinação da atividade dos bioindicadores de dose interna (Ch), de efeito (DAPs urinários) e de genotoxicidade são capazes de avaliar a exposição a agrotóxicos por trabalhadores rurais e sugere que essa exposição é altamente perigosa [4].

O estudo resultou em um conjunto de dados digitalizados, contendo 1027 amostras da população geral avaliada. Esse acervo digital constitui um valioso histórico científico de dados sobre intoxicação por agrotóxicos, facilitando pesquisas futuras e permitindo análises detalhadas e contínuas.

2.2 Ciência de Dados

2.2.1 O que é Ciência de Dados

A Ciência de Dados é um campo interdisciplinar, podendo ser aplicada em diversos domínios. Tem como objetivo converter dados linguísticos em dados inteligentes que ofereçam recursos capazes de responder a questões sociais, empresariais e científicas. O princípio da Ciência de Dados é extrair informações e conhecimento dos dados [27].

A Ciência de Dados se torna útil quando apresenta um relevante número de exemplos de dados e quando possui padrões muito complexos incapazes de serem extraídos manualmente. Considerando um número relevante e um nível de complexidade de acordo com a facilidade desses dados serem processados por um especialista humano. Quando o ser humano começa a trabalhar com mais de três tipos de dados diferentes, já começa a ter dificuldade nas interações entre eles. Normalmente a Ciência de Dados é aplicada em contextos onde se procura por padrões entre dezenas ou até milhares de dados, sendo esses padrões úteis somente se for fornecido uma visão do problema que permite alguma aplicabilidade [10].

2.2.2 Habilidades

Como é uma ciência interdisciplinar, se torna difícil para uma única pessoa ter domínio de diversas habilidades, assim é importante estar ciente da contribuição de cada habilidade para um projeto de ciência de dados, conforme mostra a Figura 2.5. Os dados estão no centro de

todos os projetos de ciência de dados. No entanto, o fato de ter acesso aos dados não significa direito sobre eles, é preciso usar os dados de forma ética. Na maioria das jurisdições, existe uma legislação de proteção de dados. Esse aspecto reveste-se de tamanha gravidade que permite a abertura de processos judiciais, em conformidade com a Lei Geral de Proteção de Dados de 2018, que regula as atividades de tratamento de dados pessoais [28].

A visualização acompanha todas as fases do processo de Ciência de Dados, podendo ou não, fazer uso de ferramentas estatísticas que podem auxiliar na compreensão do processo de comunicação. O conhecimento do domínio dos dados é muito importante para identificar uma solução otimizada. Quando o cientista de dados não possui esta habilidade é necessário que se envolva com um especialista do domínio para que possa compreender o conhecimento relevante sobre o assunto [10][29][30].

Figura 2.5 – Habilidades para Ciência de Dados [10].

O cientista de dados deve possuir habilidades de manipular conjuntos de dados digitais, sendo a forma de armazenamento mais utilizada os modelos de banco de dados relacional e orientado a objetos. Ferramentas para descoberta de conhecimento de dados e de aprendizado de máquina, são imprescindíveis para a busca da solução [10][29].

2.2.3 Origem

Para compreender a origem do termo Ciência de Dados é importante a relação entre Ciência de Dados e Estatística. A Estatística é a ciência que estuda a coleta e análise de dados. Entre 1780 e 1820, aproximadamente, diversos estatísticos estavam inventando gráficos e lançando as bases para a visualização e análise exploratória de dados [10][30].

Em 1970 com o surgimento do modelo de dados relacional, ocorreu uma revolução no

armazenamento de dados digitais. Os dados passaram a ser armazenados de forma organizada, permitindo a consulta fácil e rápida aos bancos de dados. O modelo de banco de dados relacional armazena dados em tabelas com a estrutura de uma instancia por linha e um atributo por coluna. Este modelo permitiu o crescimento de bancos de dados tanto em volume de dados quanto em amplitude de dados. Assim a análise dos dados armazenados se tornou mais complexa abrindo espaço para a Ciência de Dados [10].

O termo Ciência de Dados foi descrito por Hayashi em 1998 da seguinte forma: "A Ciência de Dados não é apenas um conceito sintético para unificar estatísticas, análise de dados e seus relacionados métodos, mas também inclui seus resultados" [30].

Daí em diante vários trabalhos passaram a contrastar a análise de dados e a estatística matemática. Cleveland escreveu: "Uma visão muito limitada da Ciência de Dados é que ela é praticada por estatísticos. A visão ampla é que a Ciência de Dados é praticada por estatísticos e analistas de assunto, confundindo exatamente quem é e quem não é estatístico" [30].

Atualmente a Ciência de Dados vai bem mais além dessas origens, além das estatísticas [30]. Estima-se que no mundo todo, são gerados e armazenados em torno de 5 exabytes de dados, todos os dias. O conjunto desses dados recebeu o nome de big data, que é definido pelos 3 Vs: um grande volume de dados, uma grande variedade de tipos de dados e a grande velocidade na qual os dados são processados [10].

2.2.4 Descoberta de Conhecimento de Dados

A descoberta de conhecimento de dados, *Knowledge Discovery Data* (KDD), envolve todo o processo de descoberta de conhecimento desde o dado bruto, limpeza de dados, transformação de dados, descoberta de padrões e apresentação do conhecimento. Também conhecida como Mineração de Dados no mundo comercial [31]. A Figura 2.6 mostra a hierarquia das atividades de Ciência de Dados, desde a coleta de dados, compreensão e exploração dos dados, descoberta de padrões, criação de modelos usando aprendizado de máquina e suporte a decisão [10].

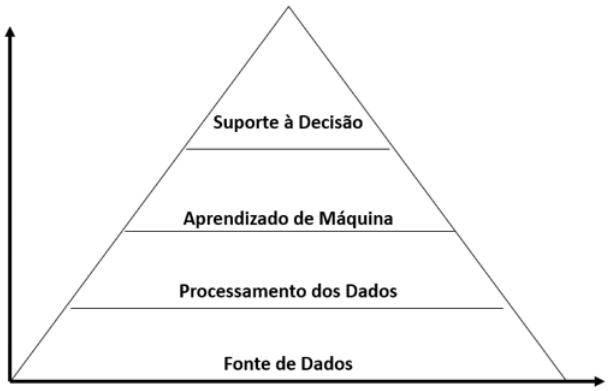

Figura 2.6 – Hierarquia da Ciência de Dados [10].

2.2.4.1 Dados

A fonte do KDD na Ciência de Dados como o próprio nome diz são os dados. Os termos atributo, variável, recurso e informação também são utilizados para designar um dado. Um dado é uma abstração de uma entidade do mundo real. Entidades concretas como pessoa, carro ou abstratos como data, hora, formas de pagamento [10]. Cada entidade é descrita por um conjunto de dados ou atributos. Por exemplo, uma data pode ter dia, mês e ano. Uma pessoa pode conter nome, sexo, idade.

Em um projeto de Ciência de Dados é importante, com base no conhecimento do domínio, extrair dados que possam ser insignificantes. Dados extras podem desviar e aumentar esforços computacionais. A decisão de quais dados, são ou não relevantes, é um grande desafio no projeto de Ciência de Dados e às vezes pode se resumir em um processo de tentativa e erro de forma iterativa, onde a cada iteração, verificam se os resultados para subconjuntos de dados diferentes [10].

2.2.4.2 Ciclo de Vida dos Dados

Um ciclo de vida dos dados delineia a atividade de um cientista de dados durante um projeto de Ciência de Dados, apresentando o caminho dos dados durante o projeto. Existem diversos tipos de modelo de ciclo de vida dos dados, que geralmente são direcionados para áreas específicas do conhecimento. Alguns dos modelos mais reconhecidos incluem o *Agile Data Science Lifecycle* [32], *Cross–Industry Standard Process for Data Mining* (CRISP-–DM) [33], *Microsoft Team Data Science Process* (Microsoft—TDSP) [34], e *Domino Data Science Lifecycle* [17]. Serão destacados os modelos *Agile Data Science Lifecycle* e *Cross—Industry Standard Process for Data Mining* (CRISP-–DM) por serem os mais utilizados.

O modelo Ciclo de Vida dos Dados para Ciência da Informação (CVD–CI) [35], embora não seja tão conhecido, também é destacado devido à importância que atribui a fatores como segurança e qualidade dos dados durante as etapas do ciclo.

- ***Agile Data Science Lifecycle*:** É uma metodologia que aplica os princípios e práticas ágeis no desenvolvimento de projetos de ciência de dados. Originária do desenvolvimento de software, a abordagem ágil enfatiza a adaptabilidade, a colaboração, a entrega contínua e iterativa de valor ao cliente, e a capacidade de responder rapidamente a mudanças. No contexto da ciência de dados, isso significa desenvolver e melhorar modelos de dados, pipelines de processamento e aplicações analíticas de forma incremental e com feedback contínuo, conforme ilustra a Figura 2.7.

  O *Agile Data Science* funciona através de iterações curtas, conhecidas como *sprints*, que duram geralmente de duas a quatro semanas, onde cada *sprint* resulta em uma versão utilizável do produto, mesmo que seja apenas um protótipo inicial. As necessidades dos usuários são capturadas como histórias de usuário, que orientam o desenvolvimento e

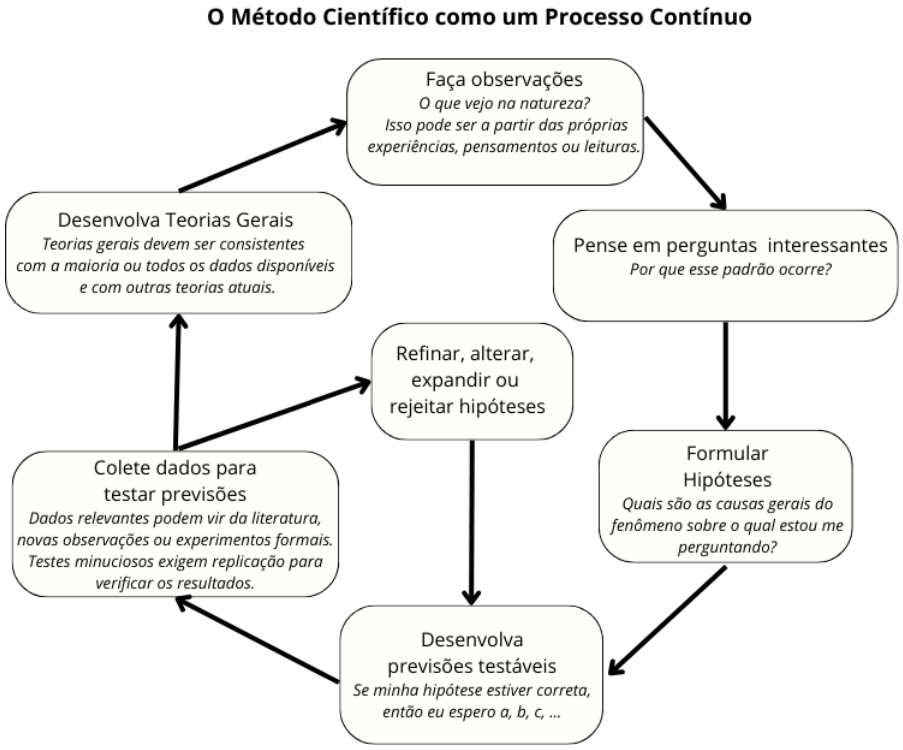

Figura 2.7 – O método científico *Agile Data Science*, como processo contínuo [36].

priorizam as tarefas. Cientistas de dados colaboram constantemente, permitindo ajustes rápidos com base no feedback frequente das partes interessadas, garantindo que o produto final atenda às suas necessidades. Ao final de cada sprint, uma versão funcional é entregue, permitindo que os usuários vejam progressos regulares e ajustem requisitos conforme necessário.

Embora a metodologia *Agile* seja amplamente utilizada no desenvolvimento de software, ela não é necessariamente suficiente para atender às necessidades de um cientista de dados. Isso porque, apesar das semelhanças superficiais entre projetos de software e ciência de dados, como a base em dados, matemática e código, a ciência de dados possui um processo inerentemente mais ambíguo e não linear. A ciência de dados exige um processo altamente exploratório e iterativo, que muitas vezes inclui longos períodos de aquisição e limpeza de dados, análise exploratória, engenharia de características, treinamento e avaliação de modelos. Este ciclo de vida natural da ciência de dados frequentemente demanda mudanças de direção no meio da investigação, algo que não se encaixa perfeitamente nas práticas ágeis de engenharia de software. Portanto, para mesclar filosofias ágeis com ciência de dados de maneira eficaz, é crucial reconhecer essas diferenças [36].

- ***Cross-–Industry Standard Process for Data Mining* (CRISP-–DM):** O CRISP–DM baseia-se em tentativas de definir metodologias de descoberta de conhecimento por um consórcio de empresas fornecedoras e consumidores potenciais de *Data Mining* [33]. O ciclo de vida da metodologia CRISP–DM, é dividido em seis fases, conforme ilustrado na

Figura 2.8. A sequência das fases não é rígida; as setas indicam apenas as dependências mais importantes e frequentes entre as fases. No entanto, em um projeto específico, a fase ou tarefa a ser realizada em seguida depende do resultado de cada fase anterior.

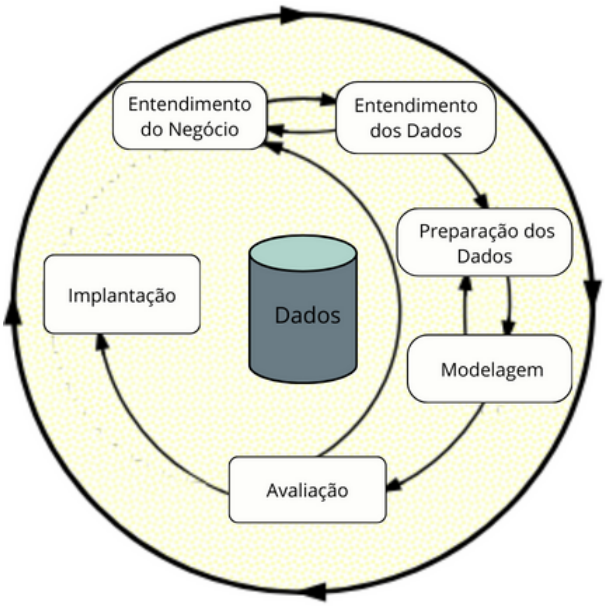

Figura 2.8 – Fases do modelo CRISP–DM [33].

Estas são as fases do ciclo de vida da metodologia CRISP–DM:

1. Entendimento do Negócio: Define objetivos e requisitos do projeto a partir de uma perspectiva de negócios, convertendo-os em um problema de mineração de dados e um plano preliminar.

2. Entendimento dos Dados: Coleta inicial de dados e atividades para entender os dados, identificar problemas de qualidade e descobrir insights iniciais.

3. Preparação dos Dados: Constrói o conjunto final de dados a partir dos dados brutos, incluindo seleção, limpeza, construção de novos atributos e transformação dos dados.

4. Modelagem: Seleciona e aplica técnicas de modelagem, ajustando parâmetros para valores ótimos, e identifica problemas de dados ou ideias para novos dados.

5. Avaliação: Avalia detalhadamente os modelos para garantir que atendem aos objetivos de negócios e verifica se algum aspecto importante foi negligenciado.

6. Implantação: Organiza e apresenta o conhecimento obtido para que o cliente possa utilizá-lo, variando de relatórios a processos repetíveis de mineração de dados.

- **Ciclo de Vida dos Dados para Ciência da Informação (CVD–CI):** Sant'Ana [35] propõe o uso do Ciclo de Vida dos Dados para a Ciência da Informação para uma análise estruturada dos dados, contendo quatro fases: Coleta, Armazenamento, Recuperação e Descarte. Nessas quatro fases estão sempre presentes fatores como: Privacidade,

Integração, Qualidade, Direitos Autorais, Disseminação e Preservação, conforme mostra a Figura 2.9 [35] .

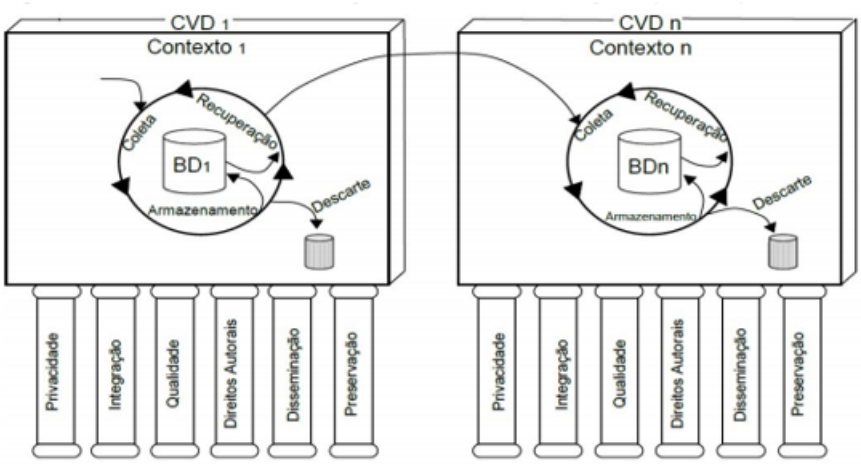

Figura 2.9 – Ciclo de Vida dos Dados para Ciência da Informação (CVD–CI) [35].

Estas são as fases do ciclo de vida da metodologia CVD–CI:

1. Coleta: Nessa etapa é importante definir bem o escopo da necessidade da informação, quais serão os dados necessários e onde serão coletados. A coleta pode ter como fonte um processo contínuo de fornecimento de dados ou uma fonte pontual com declaração nítida de início e de término, podendo se repetir em uma nova coleta.

Na fase da coleta dos dados, podem ser avaliados os fatores como:

– Privacidade: Os dados coletados apresentam algum risco de quebra de privacidade de pessoas ou instituições, que possam comprometer em um passivo futuro relacionado aos dados obtidos?

– Integração: Identificar na coleta dos dados, dados unívocos, que podem representar um conjunto de dados e que possam fazer a integração, entre dados espalhados em diversos tipos de armazenamento de forma a manter a integridade das informações.

– Qualidade: Considerar questões como a procedência dos dados e garantias de integridade física e lógica, visando qualidade e confiabilidade na informação.

– Direito Autoral: Consultar informações sobre o responsável pela fonte dos dados com relação aos direitos autorais e o direito de acesso aos dados e respeitar a competência jurídica que provê a legalidade desse acesso. Importante que todas essas informações sejam registradas junto aos dados.

– Disseminação: Prever a disseminação dos dados, desde a fase da coleta, de forma que dados que serão necessários para alcançar um maior acesso já façam parte do planejamento.

– Preservação: A conservação dos dados originalmente coletados, mesmo que tenham sofrido algum acréscimo de informação para melhor acurácia ou precisão [35].

2. Armazenamento: Tendo definido quais dados serão armazenados é preciso definir um modelo lógico e uma estrutura física para a persistência dos dados. O processo de organização e armazenamento de dados envolve seis passos. No primeiro passo, define-se um nome e especificações para cada dado coletado, incluindo tipo, tamanho, formato e especificações semânticas. No segundo passo, os dados são organizados em subconjuntos de acordo com o contexto, formando entidades como "Cliente" com atributos como CPF, nome, telefone e endereço. O terceiro passo envolve a definição de autorizações de acesso para entidades e dados específicos. No quarto passo, define-se o modelo físico de armazenamento, considerando o uso de um Sistema Gerenciador de Banco de Dados (SGBD) ou sistemas abertos como planilhas, onde um padrão de formatação semântica é essencial. O quinto passo sugere o uso do formato CSV para sistemas abertos, com a primeira linha representando o cabeçalho e as subsequentes contendo os dados separados por vírgulas. O sexto passo aborda a escolha do local de armazenamento dos dados, destacando a tendência atual de usar provedores de serviços de informação.

Na fase de armazenamento, podem ser avaliados os fatores:

– Privacidade: Está fortemente associada ao passo 3, identificando regras e pessoas que terão acesso aos dados. No passo 4, caso a opção seja um SGBD, a privacidade poderá ser alcançada com a disponibilização de mais recursos. Com relação ao local onde será armazenado os dados, descrito no passo 6, também influenciará na privacidade. Caso esteja a base dos dados armazenada localmente e desconectada de uma rede, estará mais segura que armazenada em um servidor.

– Integração: A integração vai depender fortemente dos passos 4 e 5 definindo como o usuário terá acesso e interação com os dados. Os passos 1 e 2 também terão efeito na interação do usuário com os dados, pois depende da relação que um determinado dado tem sobre o outro.

– Qualidade: As definições de armazenamento são de grande importância na garantia da qualidade dos dados. Quando um SGBD é utilizado, este auxilia a manter a integridade lógica e física dos dados.

– Direito Autoral: Ao armazenar os dados, é necessário também armazenar a fonte dos dados, visando garantir a segurança institucional de quem responde pelo armazenamento.

– Disseminação: É necessário planejar meios de acesso futuro dos dados de forma a não corromper a privacidade.

– Preservação: Ainda pensando no uso futuro, é preciso prever uma atualização dos dados sem perder a integridade lógica e uma atualização tecnológica sem perder a integridade lógica e física [35].

3. Recuperação: Esta etapa foca em tornar os dados disponíveis para acesso e uso, garantindo que sejam corretamente usados, tratados e interpretados. Quanto aos fatores

envolvidos nessa fase, é importante considerar que: a privacidade deve ser lembrada ao disponibilizar os dados para recuperação; a integração das entidades é essencial para permitir análises abrangentes dos dados; os recursos de acesso aos dados, sejam interfaces ou geração de arquivos, devem garantir a qualidade dos dados. Além disso, é crucial deixar claros os direitos autorais e as permissões de uso dos dados, e definir estratégias que facilitem a disseminação, para que os dados necessários sejam encontrados facilmente. Outra preocupação na fase de recuperação é a preservação da integridade dos dados, especialmente diante das frequentes mudanças tecnológicas [35].

4. Descarte: Em qualquer ponto deste ciclo podem ser identificados dados que não são necessários, levando a um processo de limpeza ou desativação de parte dos dados. Ou pode acontecer uma questão de privacidade, onde quem tem os direitos autorais dos dados reclame o descarte de determinados dados. No momento da realização do descarte, manter a integração das relações entre os dados é primordial para não degenerar a base de dados como um todo e colocar a qualidade e a disseminação dos dados em risco. Mesmo quando ocorre descarte de um dado, a preservação deve ser buscada, assim esse processo de descarte deve ser registrado com detalhes, pois podem surgir necessidades não previstas, que precisem dos dados eliminados [35].

Martinez et al. [37] apresentaram dados empíricos obtidos a partir de uma pesquisa com 237 profissionais de Ciência de Dados, destacando a predominância do Ciclo de Vida da Ciência de Dados Ágil sobre a metodologia tradicional CRISP–DM. No entanto, apenas 25% dos participantes afirmaram seguir uma metodologia específica, ressaltando a falta de um modelo claramente definido para a gestão de projetos de Ciência de Dados, o que tem sido identificado como um dos principais desafios nesta área.

2.2.4.3 Aprendizado de Máquina

Dentro da área de Inteligência Artificial, uma parte importante, mas não única é o aprendizado de máquina. Quando você tem um problema que precisa prever resultados com base em dados recebidos, provavelmente tem-se um problema de aprendizado de máquina [38].

O aprendizado de máquina apresenta bons resultados para o reconhecimento de padrões. Esses resultados são baseados no aprendizado com grande número de exemplos que recebem o nome de conjunto de treinamento. Para ilustrar, suponha que queira prever o resultado de uma corrida de cavalos. Poderíamos fornecer à máquina dados sobre o ranking mundial de cada animal, ranking mundial do jóquei, características do animal como raça, peso, etc., seguidos dos resultados das corridas anteriores. Diante de um grande volume de exemplos, esperamos que a máquina aprenda os padrões que indiquem a probabilidade de vitória. E com o passar do tempo, podem ser acrescentados novos exemplos que ainda não constavam no conjunto de treinamento inicial melhorando a precisão do resultado. O aprendizado de máquina pode ser classificado em três tipos: Supervisionado, Não Supervisionado e por Reforço [39].

- **Aprendizado de Máquina Supervisionado**: No aprendizado supervisionado os dados são classificados por um especialista [38]. Os exemplos são descritos por uma lista de dados e pelo rótulo da classe associada. O objetivo é construir um classificador baseado em exemplos, afim de que a máquina consiga classificar novos exemplos ainda não rotulados [39][40]. A Tabela 6 apresenta alguns exemplos:

Tabela 6 – Exemplos de conjuntos de dados supervisionados.

Lista de dados	**Classe**
Idade, sexo, doenças	Seguro de vida
Imagem, cor	Semáforo de trânsito
Salário, idade, formação	Perfil de investidor financeiro

Dentre os algoritmos mais conhecidos para resolver problemas na aprendizagem de máquina supervisionada destaca–se: *Naive Bayes*, Árvore de Decisão, Suporte a Máquina de Vetor (SVM), Regressão linear, Teoria de Conjuntos Aproximados e Redes Neurais Artificiais [38][39][41][42][43][44][45][46].

- **Aprendizado de Máquina Não Supervisionado**: Obter dados rotulados nem sempre é possível. Nesse caso, a máquina precisa descobrir padrões repetidos nos exemplos que permitam a criação de uma classe ou rótulo. Na aprendizagem de máquina não supervisionada deseja se encontrar uma representação informativa dos dados, agrupando as informações de forma relevante [38][39]. A Tabela 7, apresenta alguns exemplos.

Tabela 7 – Exemplos de conjuntos de dados não supervisionados.

Lista de dados	**Representação informativa**
Registros de compras	Perfil do Consumidor
Movimentações bancarias	Anormalidades de movimentação
Registros de compras	Associação entre produtos

Dentre os algoritmos mais conhecidos para resolver problemas na aprendizagem de máquina não supervisionada destaca-se: Algoritmos de Clusterização (*K–means clustering*, *Mean–Shift*, DBSCAN, Redes neurais), Algoritmos de Redução de Dimensão (*Principal Component Analysis*, *Singular Value Decomposition*, *Latent Dirichlet allocation*, *Latent Semantic Analysis*, Redes Neurais) e Algoritmos de Aprendizagem de Regras de Associação (*Apriori*, *Euclat*, *FP–growth*, Redes Neurais) [38][39][46].

- **Aprendizado Por Reforço**: O aprendizado por reforço recebe esse nome do conceito de "aprendizagem por reforço" da psicologia, onde se dá uma recompensa ou punição de acordo com a tomada de decisão. No algoritmo acontece o mesmo, diante da repetição de experimentos, espera se que a máquina seja capaz de associar as ações que geram maior recompensa para cada decisão tomada e evite ações que geram punições. Essa ideia é muito utilizada em jogos, automatização financeira, carros autônomos e outras aplicações. Os algoritmos mais conhecidos são: *Q–Learning*, SARSA, DQN, A3C, *Genetic algorithm*, Redes Neurais [38][39][46].

Em todos os três tipos de aprendizado de máquina, as redes neurais podem ser utilizadas com sucesso, por isso muitos se referem a aprendizado de máquina — *machine learning* como aprendizado por redes neurais profundas — *deep learning*.

Contudo fica a pergunta: Qual método utilizar? A resposta irá depender do problema analisado.

O importante em se tratar de aprendizado de máquina é ter sempre em mente que: "A máquina não faz o que você quer: faz o que você ordena!" [39]

2.3 Teoria de Conjuntos Aproximados

2.3.1 Origem

Teoria de Conjuntos Aproximados (TCA), é uma ferramenta matemática usada para tratar um tipo de incerteza e imprecisão, tendo sido proposta por Pawlak [47][48][49][50][51] e revisada em [14]. A TCA é amplamente estendida a muitas aplicações da vida real em cuidados de saúde, manufatura, finanças, engenharia e outros [52][53]. De uma maneira informal, como mostra a Figura 2.10, estes conjuntos são aproximações de um dado conjunto, ou seja, envoltórios superiores e inferiores de um dado conjunto, gerando uma região em que se encontra a linha limite do conjunto procurado.

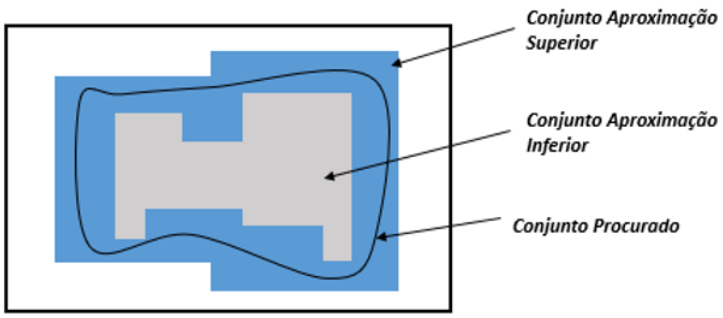

Figura 2.10 – Aproximação superior e inferior de um dado conjunto.

Essa teoria tem sido utilizada em Inteligência Artificial com ênfase nas áreas de [54]:

- diagnóstico;
- representação de conhecimento incerto;
- aprendizado indutivo e redução do número de atributos do conjunto de treinamento;
- descoberta de conhecimento em base de dados;
- controle, planejamento e administração.

2.3.2 Conjuntos Aproximados

Seja U um conjunto universo e $A = (U, R)$ um par ordenado que representa um espaço aproximado, onde R é uma relação de equivalência sobre U, chamada de relação de indiscernibilidade. Dados $x, y \in R$, então x e y são indiscerníveis em A, assim as classes de equivalência definidas por x e y são iguais.

As classes de equivalência induzidas por R em U denominam–se conjuntos elementares, notada por U/R, vista como $\tilde{R} = U/R = \{E_1, \ldots, E_n\}$, onde $E_i, 1 \leq i \leq n$, é um conjunto elementar de A. Sendo assim, o espaço aproximado $A = (U, R)$ pode ser notado como $A = (U, \tilde{R})$. O conjunto vazio $\emptyset$ será sempre um conjunto elementar para qualquer espaço aproximado A. Um conjunto definível em A é toda união finita de conjuntos elementares.

Exemplo 2.1: Dado um espaço aproximado $A = (U, R)$, sendo $U = \{x_1, x_2, x_3, x_4, x_5\}$ e R uma relação de equivalência sobre U de forma que $U/R = \{\{x_1, x_4\}, \{x_2, x_5\}, \{x_3\}\}$. São conjuntos elementares em A.

$$D_1 = \{x_1, x_4\}, \quad D_2 = \{x_2, x_5\}, \quad D_3 = \{x_3\}, \quad D_4 = \emptyset$$

Dado $A = (U, R)$ um espaço aproximado e dado X um subconjunto qualquer de objetos de U, com o objetivo de verificar quão bem o conjunto de descrições $\text{des}([x]_R)$, $x \in U$, reflete as funções de pertinência de objetos a X, são definidos os seguintes conceitos:

1. A aproximação inferior de X em A, notada por $A_{\text{inf}}(X)$, é caracterizada pela união dos conjuntos elementares de A que estão contidos totalmente em X, em símbolos:

$$A_{\text{inf}}(X) = \{x \in U \mid [x]_R \subseteq X\}$$

2. A aproximação superior de X em A, notada por $A_{\text{sup}}(X)$, é caracterizada como a união de todos os conjuntos que possuem intersecção não vazia com X, em símbolos:

$$A_{\text{sup}}(X) = \{x \in U \mid [x]_R \cap X \neq \emptyset\}$$

Exemplo 2.2: Seja A= (U,R) um espaço aproximado, onde U é um conjunto universo, $X \subseteq U$, e R a relação de equivalência em U. As Figuras 2.11 e 2.12 mostram a aproximação inferior e superior de X, respectivamente.

2.3.2.1 Regiões Identificadas em um Conjunto Aproximado

Seja $A = (U, R)$ um espaço aproximado e $X \subseteq U$. Podem ser facilmente encontradas as seguintes regiões, relacionando X em U:

1. Região positiva de X em A, $\text{pos}_A(X)$, é definida pelas classes de equivalência de U que

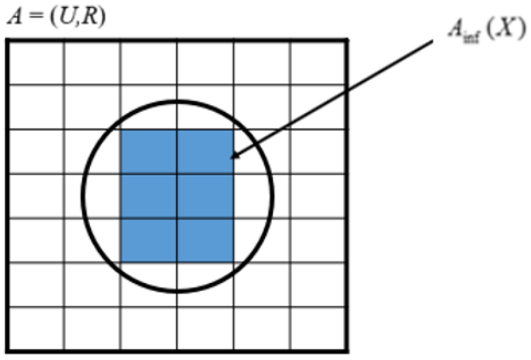

Figura 2.11 – Aproximação inferior de $X \subseteq U$ em $A = (U, R)$.

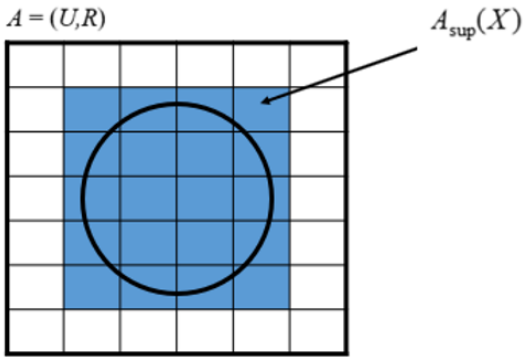

Figura 2.12 – Aproximação superior de $X \subseteq U$ em $A = (U, R)$.

estão integralmente no conjunto X, equivale à aproximação inferior de X em A, i.e., $\text{pos}_A(X) = \text{A}_{\text{inf}}(X)$.

2. Região negativa de X em A, $\text{neg}_A(X)$, é definida pelas classes de equivalência de U que não contêm elemento em X. Os conjuntos elementares que não fazem parte da aproximação superior, formam a região negativa, i.e., $\text{neg}_A(X) = U - \text{A}_{\text{sup}}(X)$.

3. Região duvidosa de X em A, $\text{duv}_A(X)$, é aquela região nebulosa, onde X pode ou não estar presente, porém seus elementos pertencem à aproximação superior e não constam na aproximação inferior. Há uma incerteza de um elemento dessa região, com relação ao conjunto X. Notação,

$$\text{duv}_A(X) = \text{A}_{\text{sup}}(X) - \text{A}_{\text{inf}}(X)$$

Sendo A conhecido, pode-se simplificar escrevendo $\text{pos}(X)$, $\text{neg}(X)$ e $\text{duv}(X)$ em substituição a $\text{pos}_A(X)$, $\text{neg}_A(X)$ e $\text{duv}_A(X)$.

Exemplo 2.3: Seja $A = (U, R)$ um espaço aproximado. Considere $U = \{x_1, x_2, x_3, \ldots, x_{12}\}$ e a relação de equivalência ou indiscernibilidade como $U/R = \{\{x_1, x_2, x_3\}, \{x_4\}, \{x_5, x_6, x_7\}, \{x_8, x_9, x_{10}\}, \{x_{11}, x_{12}\}\}$. E $X = \{x_2, x_3, x_4, x_5, x_{12}\}$.

Pode-se observar a aproximação inferior, superior e as regiões positiva, negativa e duvidosa:

$$A_{inf}(X) = \{x_4\} \quad \text{e} \quad A_{sup}(X) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_{11}, x_{12}\}$$
$$pos(X) = \{x_4\}, \quad neg(X) = \{x_8, x_9, x_{10}\}, \quad \text{e} \quad duv(X) = \{x_1, x_2, x_3, x_5, x_6, x_7, x_{11}, x_{12}\}.$$

As regiões positiva, negativa e duvidosa de X, conforme o Exemplo 2.3, estão ilustradas nas Figuras 2.13, 2.14 e 2.15.

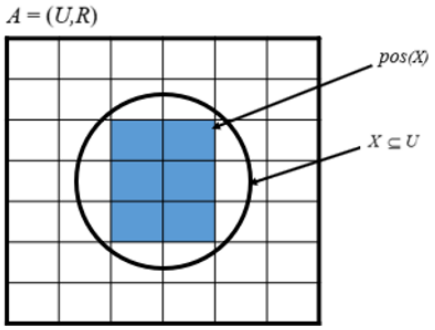

Figura 2.13 – A região positiva de $X \subseteq U$.

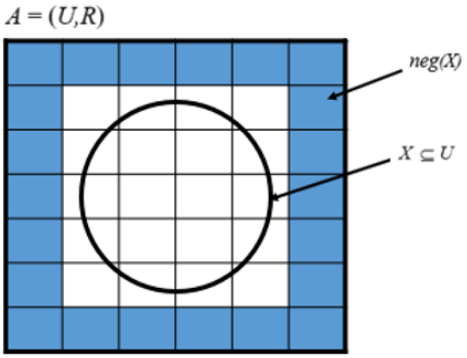

Figura 2.14 – A região negativa de $X \subseteq U$.

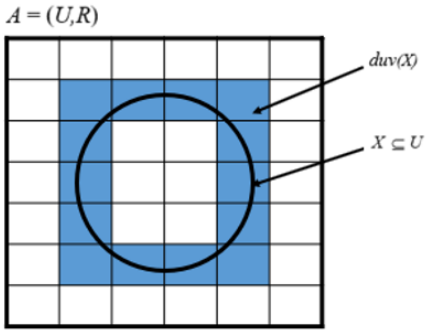

Figura 2.15 – A região duvidosa de $X \subseteq U$.

Na Figura 2.16, todos os conjuntos são aproximadamente iguais. Nesse universo, qualquer conjunto representa o conjunto original X. Por isso são chamados de conjuntos aproximados.

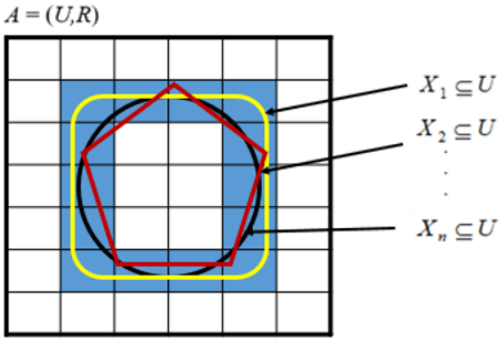

Figura 2.16 – Os conjuntos $\{X_1, X_2, \ldots, X_n\}$, todos tendo a mesma A_{inf} e A_{sup}, definem um conjunto aproximado X no espaço aproximado $A = (U, R)$.

2.3.3 Sistemas de Representação de Conhecimento

Sistemas de Representação de Conhecimento (SRC), devem ser capazes de representar, manipular e comunicar dados, definidos como a 4–upla $S = (U, Q, V, \rho)$, onde [55][56][57][58]:

- U: é o universo do problema e é composto por conjuntos de objetos. Estes conjuntos de objetos são chamados de exemplos quando se trata de Aprendizado de Máquina.
- Q: é o conjunto finito de atributos, $Q = D \cup C$, onde D é o conjunto de atributos de decisão e C o conjunto de atributos de condição.
- $V = \bigcup_{q \in Q} V_q$: sendo V o domínio do atributo q quando q pertence ao conjunto Q.
- ρ: é a descrição do domínio de $U \times Q$, ou seja, $\rho : U \times Q \rightarrow V$.

Seguem algumas observações:

1. Quanto maior o número de descrições diferentes em $|V_q|$ para cada atributo, maior o número de descrições possíveis em S.
2. Como os objetos ou exemplos são formados por vários atributos, a função ρ será definida como uma sequência de valores de atributos.

Pode-se representar um SRC por uma tabela, onde a descrição de cada objeto é representada pelas linhas da tabela e cada atributo é representado pelas colunas da tabela. Normalmente, o último atributo da tabela representa a classe do objeto.

Exemplo 2.4: Considere $S = (U, Q, V, \rho)$ um SRC sobre informações de imóveis de uma imobiliária, fornecido pela Tabela 8, sendo que o atributo a representa o imóvel, b informa se a construção é rebocada ou de tijolo a vista, c representa a idade da construção e d informa a opinião da imobiliária sobre o imóvel. Neste exemplo os atributos a, b, c e d são codificados da seguinte forma:

- $a = (1$: casa de 1 quarto, 2 : casa de 2 quartos, 3 : casa de 3 quartos, 4 : casa de 4 quartos)
- $b = (R$: Rebocada, T : Tijolo a vista)
- $c = (N$: Construção nova para construções com menos de três anos, $V1$: Construção velha e com bom estado de conservação, $V2$: Construção velha e com mal estado de conservação)
- $d = (B$: Preço baixo, R : Preço razoável, A : Preço alto)

Note que, neste caso, tem-se $U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$ e suponha $V_a = \{1, 2, 3, 4\}$, $V_b = \{R, T\}$, $V_c = \{N, V1, V2\}$, e $V_d = \{B, R, A\}$.

Tabela 8 – Exemplo de um SRC.

	a	**b**	**c**	**d**
x_1	3	R	N	A
x_2	2	R	N	A
x_3	3	R	V1	R
x_4	1	R	V1	B
x_5	2	R	V1	R
x_6	2	T	V2	B
x_7	3	R	V2	B
x_8	3	R	N	A
x_9	1	R	N	A
x_{10}	2	T	N	A

Entre as descrições fornecidas pelo sistema encontram-se:

- $\rho(x_4, a) = 1$
- $\rho(x_7, c) = V2$
- $\rho(x_{10}, b) = T$
- $\rho(x_1, d) = A$

É possível observar, na Tabela 8, as seguintes relações de equivalência: se a construção é nova, o preço sempre será alto; se a construção é velha, o preço é baixo; se a construção é velha e está em bom estado de conservação, o preço pode ser razoável ou baixo. Note também que o atributo b não influencia no preço do imóvel.

Observe que, para o sistema do Exemplo 2.4, não existe nenhum objeto $x \in U$ tal que $\rho(x, a) = 4$. Assim, o sistema é dito não-maximal, ou seja, existem descrições previstas que não são utilizadas pelo sistema na descrição dos objetos. Diz-se que o sistema é maximal se os objetos de um sistema utilizam todas as descrições previstas.

Um SRC pode representar eficientemente tabelas de decisão, onde um conjunto de atributos forma o conjunto de condições e um atributo representa a decisão. Os SRCs formados pela

tabela de decisão têm sido usados com sucesso para extrair conhecimento na forma de regras da tabela de decisão, facilitando a tomada de decisão.

Considerando o Exemplo 2.4, distinguimos os seguintes conjuntos de atributos de condição e decisão:

- $U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$ é o conjunto de objetos.
- $C = \{a, b, c\}$ é o conjunto das condições.
- $D = \{d\}$ é o conjunto das decisões.

2.3.4 Dependência entre Atributos de Condição e Decisão

A TCA permite a eliminação de atributos supérfluos. A eliminação desses possibilita a minimização do conjunto de atributos preservando o seu poder discriminatório [59]:

> Análise de dependência é usada para determinar se um grupo de atributos (condições) pode caracterizar os valores de um outro atributo (decisão). A dependência entre as condições e decisão indica o quanto a decisão pode ser baseada nos valores dos atributos das condições. (...) Em dados onde o atributo de decisão pode ser unicamente caracterizado em termos de condições, o grau de dependência é igual a um e a dependência é considerada totalmente funcional. O grau de dependência próximo de um, por exemplo, que caracteriza uma situação onde a maioria das decisões pode ser determinada pelas condições, de maneira não ambígua, estabelece as bases para hipóteses que evidenciam relações significativas entre condições e decisão. A dependência próxima de zero sugere uma relação fraca ou mesmo não existente. [1]

Sejam duas famílias de atributos, sendo um conjunto de atributos de condição e um conjunto de atributos de decisão em S [60]. Representam:

$$D' = \{D'_1, D'_2, \ldots, D'_n\} \quad \text{a família de conjuntos elementares da relação } \tilde{D}, \text{i.e., } D' = U/\tilde{D}.$$

e

$$C' = \{C'_1, C'_2, \ldots, C'_m\} \quad \text{a família de conjuntos elementares da relação } \tilde{C}, \text{i.e., } C' = U/\tilde{C}.$$

A região positiva de D', no espaço aproximado $A = (U, \tilde{C})$ é dada pela equação (2.1):

$$pos(C, D) = \bigcup_{i=1}^{n} \{A_{C-\inf}(D'_i) \mid D'_i \in D'\} \tag{2.1}$$

Para encontrar pela notação a região positiva de D', deve-se definir a aproximação inferior de cada elemento de D', no espaço aproximado induzido pela relação $\tilde{C}$ em U. A região positiva

[1] Cópia da tradução feita em [58]

de D' será a região de U que é visível usando $\tilde{C}$, i.e., todo objeto em pos(C, D) pode estar unicamente em uma das classes de D' baseando-se somente nos valores de atributos de C.

Exemplo 2.5: Seja um espaço aproximado $A = (U, \tilde{C})$, onde os conjuntos C' e D' são dados por $U/\tilde{D} = D' = \{D'_1, D'_2, D'_3\}$ e $U/\varepsilon = C' = \{C'_1, C'_2, \ldots, C'_9, C'_{10}\}$, conforme mostra a Figura 2.17. Assim:

$$\text{pos}(C, D) = A_{C-\text{inf}}(D'_1) \cup A_{C-\text{inf}}(D'_2) \cup A_{C-\text{inf}}(D'_3)$$

$$= (C'_1 \cup C'_4 \cup C'_5) \cup (C'_9) \cup (C'_3 \cup C'_7) = C'_1 \cup C'_3 \cup C'_4 \cup C'_5 \cup C'_7 \cup C'_9$$

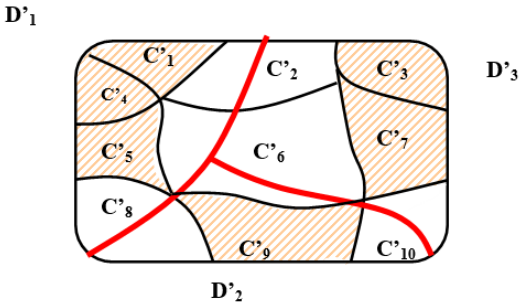

Figura 2.17 – C' e D' correspondem aos conjuntos induzidos pelos atributos de condição e decisão.

Percebe-se que os elementos pertencem a:

- C'_1 ou C'_4 ou C'_5 garante que pertencem a D'_1.
- C'_9 garante que pertencem a D'_2.
- C'_3 ou C'_7 garante que pertencem a D'_3.

O grau de dependência é a relação de dependência de um conjunto de atributos de decisão sobre um conjunto de atributos de condição. O grau de dependência em um espaço aproximado $A = (U, \tilde{C})$, é calculado conforme a equação (2.2):

$$k(C, D) = \frac{|\text{pos}(C, D)|}{|U|} \tag{2.2}$$

O resultado de $k(C, D)$ permite valores no intervalo $[0, 1]$, como na referência [60]. Se:

- $k(C, D) = 1$, a dependência é completa (ou funcional).
- $0 < k(C, D) < 1$, a dependência é parcial.
- $k(C, D) = 0$, os conjuntos de atributos C e D são independentes.

Exemplo 2.6: Seja um SRC fornecido pela Tabela 9, que descreve uma avaliação de exames de imunização após trinta dias de tomar uma determinada vacina, em termos de um conjunto de atributos. O objetivo é analisar a relação entre o atributo imunidade e os demais atributos, o atributo imunidade foi considerado como o atributo de decisão D e os demais atributos, como o conjunto de condições C, ou seja:

C = {obesidade, presença de anticorpos, doenças autoimunes, sexo, tipo sanguíneo, concentração de vitamina D} e D = {imunidade}.

Tabela 9 – SRC que descreve 11 objetos (exames) usando 7 atributos (6 condições e 1 decisão).

U	Obesidade	Presença de Anticorpos	Doenças Autoimune	Sexo	Tipo Sanguíneo	Vitamina D	Imunidade
e_1	baixa	baixa	sim	F	A	média	baixa
e_2	média	baixa	sim	F	A	média	média
e_3	média	média	não	M	AB	alta	média
e_4	média	alta	não	M	O	alta	alta
e_5	média	média	não	M	O	alta	média
e_6	média	média	não	M	A	alta	média
e_7	média	baixa	sim	F	O	média	média
e_8	média	alta	não	M	O	alta	alta
e_9	alta	média	não	M	A	alta	alta
e_{10}	alta	alta	não	M	A	alta	alta
e_{11}	alta	baixa	sim	F	O	média	média

No espaço aproximado $A = (U, \tilde{C})$ do Exemplo 2.6, podem ser identificados os conjuntos:

- $U/\tilde{C} = C' = \{C'_1, C'_2, ..., C'_{10}, C'_{11}\} =$
  $\{\{e_1\}, \{e_2\}, \{e_3\}, \{e_4\}, \{e_5\}, \{e_6\}, \{e_7\}, \{e_8\}, \{e_9\}, \{e_{10}\}, \{e_{11}\}\}$
- $U/\tilde{C} = D' = \{D'_1, D'_2, D'_3\} =$
  $\{\{e_1\}, \{e_2, e_3, e_5, e_6, e_7, e_{11}\}, \{e_4, e_8, e_9, e_{10}\}\}$

Sendo assim,

$$\text{pos}(C, D) = \bigcup_{1<i<3} \{A_{C-\text{inf}}(D'_i) \mid D'_i \in D'\} = A_{C-\text{inf}}(D'_1) \cup A_{C-\text{inf}}(D'_2) \cup A_{C-\text{inf}}(D'_3) =$$

$$= (C_1) \cup (C_2 \cup C_3 \cup C_5 \cup C_6 \cup C_7 \cup C_{11}) \cup (C_4 \cup C_8 \cup C_9 \cup C_{10}) = U$$

Portanto,

$$k(C, D) = \frac{|\text{pos}(C, D)|}{|U|} = \frac{11}{11} = 1$$

A dependência completa mostra que o atributo imunidade de um exame é afetado pela interação de todos ou de alguns dos atributos do conjunto C. De acordo com Ziarko [44], ainda existe uma questão em aberto, que seria qual das combinações de atributos precisamente

afetam imunidade e quais são irrelevantes. Os Exemplos 2.7 e 2.8 demonstram combinações de atributos irrelevantes e relevantes respectivamente da Tabela 9 em relação ao atributo imunidade.

Exemplo 2.7: Considere C = {presença de anticorpos, doenças autoimune} e D = {imunidade} do SRC do Exemplo 2.6, no espaço aproximado $A = (U, \tilde{C})$, representado pela Tabela 10.

Tabela 10 – SRC que descreve 11 objetos (exames) usando 3 atributos (2 condições e 1 decisão).

U	**Presença de Anticorpos**	**Doenças Autoimune**	**Imunidade**
e_1	baixa	sim	baixa
e_2	baixa	sim	média
e_3	média	não	média
e_4	alta	não	alta
e_5	média	não	média
e_6	média	não	média
e_7	baixa	sim	média
e_8	alta	não	alta
e_9	média	não	alta
e_{10}	alta	não	alta
e_{11}	baixa	sim	média

Pode-se simplificar a Tabela 10 com relação somente aos atributos de condição, conforme mostra a Tabela 11.

Tabela 11 – SRC que descreve 11 objetos (exames) usando 2 atributos de condições.

U	**Presença de Anticorpos**	**Doenças Autoimune**
e_1, e_2, e_7, e_{11}	baixa	sim
e_3, e_5, e_6, e_9	média	não
e_4, e_8, e_{10}	alta	não

Como mostra a Figura 2.18 e 2.19, observa-se que:

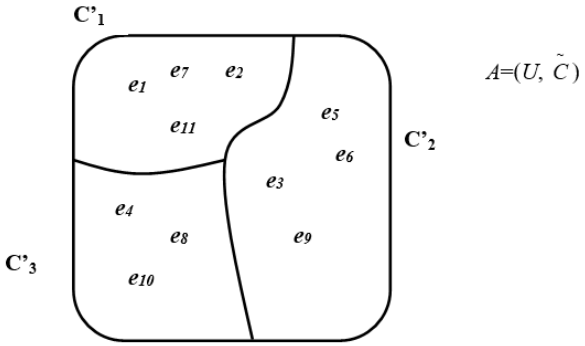

Figura 2.18 – Espaço aproximado $A = (U, \tilde{C})$, considerando-se C = {presença de anticorpos, doenças autoimune}.

Tem-se: $U/\tilde{C} = C' = \{C'_1, C'_2, C'_3\}$ e $U/\tilde{D} = D' = \{D'_1, D'_2, D'_3\}$.

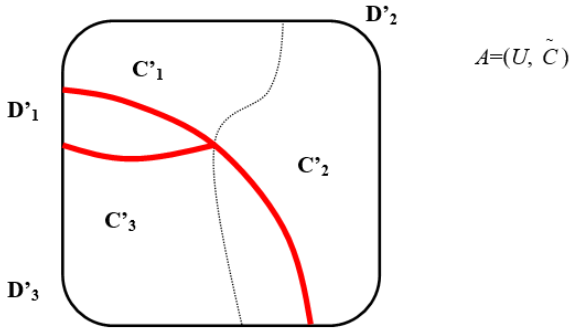

Figura 2.19 – Espaço aproximado $A = (U, \tilde{C})$, considerando-se C = {presença de anticorpos, doenças autoimune} e D = {imunidade}.

Podendo ser identificados os conjuntos:

- $U/\tilde{C} = C' = \{C'_1, C'_2, ..., C'_3\} = \{\{e_1, e_2, e_7, e_{11}\}, \{e_3, e_5, e_6, e_9\}, \{e_4, e_8, e_{10}\}\}$
- $U/\tilde{D} = D' = \{D'_1, D'_2, D'_3\} = \{\{e_1\}, \{e_2, e_3, e_5, e_6, e_7, e_{11}\}, \{e_4, e_8, e_9, e_{10}\}\}$

A região positiva de D em C será:

$$\text{pos}(C, D) = \bigcup_{1<i<3} \{A_{C-\text{inf}}(D'_i) \mid D'_i \in D'\} = A_{C-\text{inf}}(D'_1) \cup A_{C-\text{inf}}(D'_2) \cup A_{C-\text{inf}}(D'_3) =$$

$$= \emptyset \cup \emptyset \cup (C'_3) = \{e_4, e_8, e_{10}\}$$

A Figura 2.20 ilustra as regiões positiva e negativa no espaço aproximado $A = (U, \tilde{C})$ formado a partir do SRC, sendo C = {presença de anticorpos, doenças autoimune} e D = {imunidade}. Podendo a presença de anticorpos ($V1$) variar dentro do domínio {b – baixa, m – média, a – alta} e doenças autoimune ($V2$) no domínio {s – sim, n – não}.

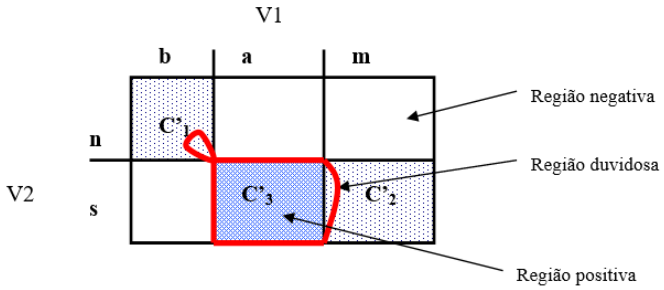

Figura 2.20 – Região positiva, negativa e duvidosa de $D \subseteq U$.

Portanto,

$$k(C, D) = \frac{|\text{pos}(C, D)|}{|U|} = \frac{3}{11} = 0.27$$

isso significa que presença de anticorpos e doenças autoimune não são suficientes para, sozinhos, definir a imunidade.

Exemplo 2.8: Considere $C = \{\text{Obesidade, presença de anticorpos}\}$ e $D = \{\text{imunidade}\}$ do SRC do Exemplo 2.6, no espaço aproximado $A = (U, \tilde{C})$, representado pela Tabela 12

Podendo ser identificados os conjuntos:

- $U/\tilde{C} = C' = \{C'_1, C'_2, \ldots, C'_6, C'_7\} =$
  $\{\{e_1\}, \{e_2, e_7\}, \{e_3, e_5, e_6\}, \{e_4, e_8\}, \{e_9\}, \{e_{10}\}, \{e_{11}\}\}$
- $U/\tilde{D} = D' = \{D'_1, D'_2, D'_3\} =$
  $\{\{e_1\}, \{e_2, e_3, e_5, e_6, e_7, e_{11}\}, \{e_4, e_8, e_9, e_{10}\}\}$

Como mostra a Figura 2.21 e 2.22, observa-se então que:

$$\text{pos}(C, D) = \bigcup_{1<i<3} \{A_{C-\text{inf}}(D'_i) \mid D'_i \in D'\} = A_{C-\text{inf}}(D'_1) \cup A_{C-\text{inf}}(D'_2) \cup A_{C-\text{inf}}(D'_3) =$$

$$= (C_1) \cup (C_2 \cup C_3 \cup C_7) \cup (C_4 \cup C_5 \cup C_6) = \{e_1, e_2, e_7, e_3, e_5, e_6, e_1 1, e_4, e_8, e_9, e_{10}\}$$

Tabela 12 – SRC que descreve 11 objetos (exames) usando 3 atributos (2 condições e 1 decisão)

U	**Obesidade**	**Presença de Anticorpos**	**Imunidade**
e_1	baixa	baixa	baixo
e_2	média	baixa	médio
e_3	média	média	médio
e_4	média	alta	alto
e_5	média	média	médio
e_6	média	média	médio
e_7	média	baixa	médio
e_8	média	alta	alto
e_9	alta	média	alto
e_{10}	alta	alta	alto
e_{11}	alta	baixa	médio

A Figura 2.23 ilustra as regiões positiva e negativa no espaço aproximado $A = (U, \tilde{C})$ formado a partir do SRC fornecido pela Tabela 12, sendo $C = \{\text{obesidade, presença de anticorpos}\}$ e $D = \{\text{imunidade}\}$. Podendo a Obesidade $(V1)$ e a presença de anticorpos $(V2)$ variar dentro do domínio $\{b - \text{baixa}, m - \text{média}, a - \text{alta}\}$.

Portanto,

$$k(C, D) = \frac{|\text{pos}(C, D)|}{|U|} = \frac{11}{11} = 1$$

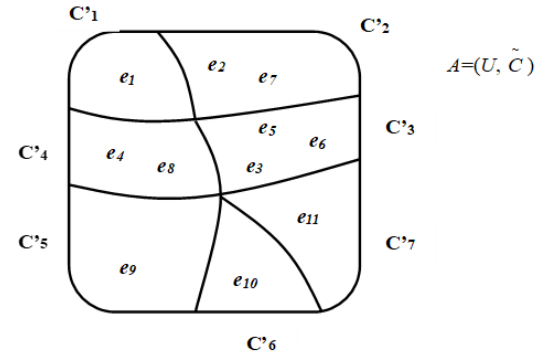

Figura 2.21 – Espaço aproximado considerando-se C= {obesidade, presença de anticorpos}.

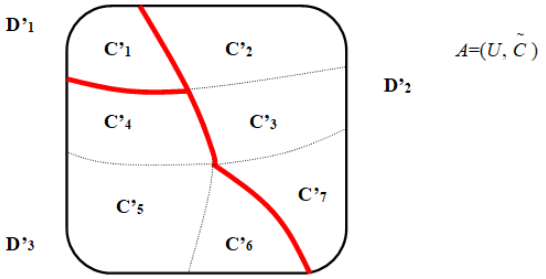

Figura 2.22 – Espaço aproximado considerando-se C= {obesidade, presença de anticorpos} e D= {imunidade}.

isso significa que obesidade e presença de anticorpos são suficientes para, sozinhos, definir a imunidade. Esse exemplo leva à noção de conjunto mínimo de atributos, que pode ser chamado também de reduto. Um reduto tem o mesmo poder de determinação que o conjunto original.

Como podem ser encontrados mais de um reduto para o mesmo SRC, é possível escolher um reduto de acordo com os atributos disponíveis, viabilizando a tomada de decisão. Se um SRC como $S = (U, Q, V, \rho)$ e o reduto encontrado e escolhido de Q em S, o subsistema $S^* = (U, P, V, \rho)$ é chamado de sistema reduzido. S e S^* são equivalentes.

Existem outras formas de encontrar redutos e suas aproximações; um exemplo são os algoritmos chamados redutores relativos ao objeto, utilizados para geração de regras de decisão.

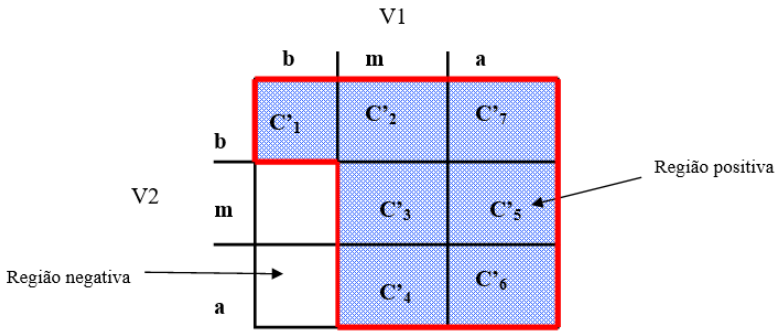

Figura 2.23 – Região positiva e negativa de $D \subseteq U$.

Esses redutores podem ser calculados usando heurísticas baseadas, por exemplo, na abordagem de discernibilidade e raciocínio booleano [61].

2.3.5 Dicernibilidade e Raciocínio Booleano

As relações de *discernibilidade/indiscernibilidade* estão intimamente relacionadas e pertencem às relações mais importantes da Teoria dos Conjuntos Aproximados. Na abordagem de conjuntos aproximados, para $S = (U, Q, V, \rho)$ e $P \subseteq Q$ a relação de discernibilidade $DIS(P) \subseteq U \times U$ é definida por $xDIS(P)y$ se e somente se não $(xI(P)y)$.

Uma *indiscernibilidade* pode ser definida por $x \in I(y)$ e a *discernibilidade* por $I(x) \cap I(y) = \emptyset$ para quaisquer objetos x, y onde $I(x) = P(x), I(y) = P(y)$ e geralmente $I(x), I(y)$ são vizinhanças de objetos, não necessariamente definidas pela relação de equivalência [61][62].

Um estudo para o cálculo de redutos foi desenvolvido por Skowron e Rauszer [63], baseado na noção de matriz de discernibilidade e álgebra booleana. Na matriz de discernibilidade, tanto as linhas como as colunas da matriz correspondem aos objetos contidos no universo do SRC. Um elemento da matriz representa o conjunto de todos os atributos nos quais os dois objetos correspondentes possuem valores distintos. Pode-se construir uma função de discernibilidade booleana a partir de uma relação de discernibilidade, com atributos como variáveis booleanas [63][64]. A matriz de discernibilidade é simétrica e com a diagonal vazia.

Seja um SRC como $S = (U, Q, V, \rho)$, sua matriz de discernibilidade $M = (M(x, y))$ é a $|U| \times |U|$ matriz, na qual o elemento $M(x, y)$ para um par de elementos (x, y) é definido pela equação (2.3):

$$M(x, y) = \{a \in Q \mid I_a(x) \neq I_a(y)\} \tag{2.3}$$

A função de *Discernibilidade* é dada pela equação (2.4):

$$F(Q) = \bigwedge_{(i,j) \in U \times U} \left(\bigvee M(i, j) \right) \tag{2.4}$$

Onde $m(i, j) = \bigvee \{a \mid a \in M(i, j)\}$, sendo que $\wedge$ representa o operador AND (produto) e $\vee$ representa o operador OR (soma).

Exemplo 2.9: Seja um SRC representado pelas informações da Tabela 13, deseja se saber as condições, que definem os indivíduos mais habilitados, para o cargo de Analista de Sistemas em uma determinada empresa. Os candidatos são analisados quanto tenham (S) ou não (N) alguma experiência profissional (EXP), se têm ou não conhecimento em inglês (ING) e se possuem conhecimento baixo (B), médio (M) ou alto (A) em linguagens de programação (LING).

Para construir a matriz de discernibilidade apresentada na Tabela 15, é necessário gerar inicialmente a Tabela 14 contendo apenas os exemplos distintos referentes aos atributos de condição:

Tabela 13 – SRC que descreve 5 objetos usando 4 atributos (3 condições e 1 decisão)

U	**EXP**	**ING**	**LING**	**Analista**
e_1	N	S	A	Média
e_2	N	S	A	Média
e_3	S	S	A	Maior
e_4	N	N	M	Baixa
e_5	N	N	M	Baixa

Tabela 14 – SRC com objetos distintos

U	**EXP**	**ING**	**LING**
e_1	N	S	A
e_2	S	S	A
e_3	N	N	M

Pode-se verificar que a interseção entre os objetos e_1 e e_2 será o atributo experiência profissional (**EXP**), sendo o único atributo que contém valores diferentes entre os objetos. O conhecimento em inglês (**ING**) e em linguagem de programação (**LING**) possuem valores diferentes para os objetos e_1 e e_3. Para os objetos e_2 e e_3, todos os atributos possuem valores diferentes.

De acordo com a Tabela 15, a função de discernibilidade será:

$$F(Q) = \text{EXP} \wedge (\text{ING} \vee \text{LING}) \wedge (\text{EXP} \vee \text{ING} \vee \text{LING})$$

Pela álgebra booleana tem-se a expressão simplificada:

$$F(Q) = \text{EXP} \wedge (\text{ING} \vee \text{LING})$$

A expressão $F(Q)$ apresentada na forma da "Soma do Produto":

$$F(Q) = (\text{EXP} \wedge \text{ING}) \vee (\text{EXP} \wedge \text{LING})$$

Dessa expressão obtém-se os possíveis redutos $\{\{\text{EXP}, \text{ING}\}, \{\text{EXP}, \text{LING}\}\}$ para o SRC referente à Tabela 13. Isto indica que experiência profissional e conhecimento em inglês ou experiência profissional e conhecimento em linguagem de programação, são suficientes para, sozinhos, definir a contratação no cargo de analista de sistema. Fornecendo um ou mais conjuntos mínimos ou redutos com o mesmo poder de decisão que o conjunto original. Seja o reduto escolhido $P = \{\text{EXP}, \text{ING}\}$ ou $P = \{\text{EXP}, \text{LING}\}$, $P \subseteq Q$ e o sistema $S = (U, Q, V, \rho)$ é

Tabela 15 – Matriz de Discernibilidade

U	e_1	e_2	e_3
e_1	-	-	-
e_2	EXP	-	-
e_3	ING, LING	EXP, ING, LING	-

equivalente ao subsistema $S^* = (U, P, V, \rho)$.

2.3.6 Cálculo de Redutos e Complexidade Computacional

A obtenção dos redutos mínimos de um SRC de dimensão elevada geralmente consiste em um problema determinístico, de complexidade computacional igual a $O(kn^2)$, onde n é o número de objetos e k é o número de atributos da tabela de dados, sendo assim, crescente com o volume de dados do sistema.Comprovou-se que encontrar o conjunto de todos os redutos, ou encontrar um reduto ótimo (ou seja, um reduto com o número mínimo de atributos), são problemas NP–completos ou NP–difíceis. No entanto, muitos métodos heurísticos para encontrar algoritmos redutores eficientes, do ponto de vista do tempo necessário para o cálculo da solução, foram investigados retornando soluções promissoras. Muitos desses métodos são baseados em matrizes de discernibilidade [61][62][64][65].

Hoa [65], apresenta alguns algoritmos eficientes para o cálculo de redutos, onde não há a necessidade de armazenar a matriz de discernibilidade. Propõe heurísticas eficientes para computação dos conceitos básicos da Teoria de Conjuntos Aproximados como aproximações inferiores e superiores, regiões positivas, com complexidade.

$$O(PU \log U)$$

Essas heurísticas apresentam bom desempenho, sendo viáveis para grandes tabelas de dados.

2.4 Considerações Finais

Neste capítulo, foram expostos os princípios teóricos que fundamentaram o desenvolvimento do modelo proposto neste trabalho. A compreensão dos conceitos relacionados aos agrotóxicos e às formas de intoxicação, à ciência de dados e seus modelos, e à teoria de conjuntos aproximados foi essencial para a construção e compreensão do modelo. O próximo capítulo apresentará o modelo proposto.

3 Modelo proposto CRD–TCA

Nesse capítulo será apresentada uma proposta para a descoberta de conhecimento de dados, *Knowledge Discovery Data* (KDD), denominada Ciclo de Refinamento de Dados usando a Teoria de Conjuntos Aproximados (CRD–TCA), com o objetivo de fornecer uma metodologia estruturada, baseada no CVD–CI, mas com duas novas etapas, sendo elas o Refinamento e o Treinamento que nesse caso, será realizado utilizando a TCA.

O modelo proposto abrange todas as etapas do processo KDD. O quadro geral do modelo é apresentado na Figura 3.1, na qual delineamos a sequência de etapas que constituem esse processo. Nesta figura, o pipeline do modelo é observável, ilustrando as distintas fases e sua sequência. Em resumo, o modelo adota uma abordagem cíclica que incorpora a coleta de dados, exclusão de dados irrelevantes, armazenamento de dados e refinamento contínuo de dados, com a capacidade de eliminar informações desnecessárias e incorporar novos dados conforme necessário. Subsequentemente, durante o processo de treinamento, podem surgir insights indicando a necessidade de refinamentos adicionais. Ao final do ciclo, o modelo transmite conhecimento para facilitar a tomada de decisões. Na fase subsequente, a recuperação de dados é baseada nos resultados da etapa de refinamento, possibilitando a execução de novos processos de treinamento.

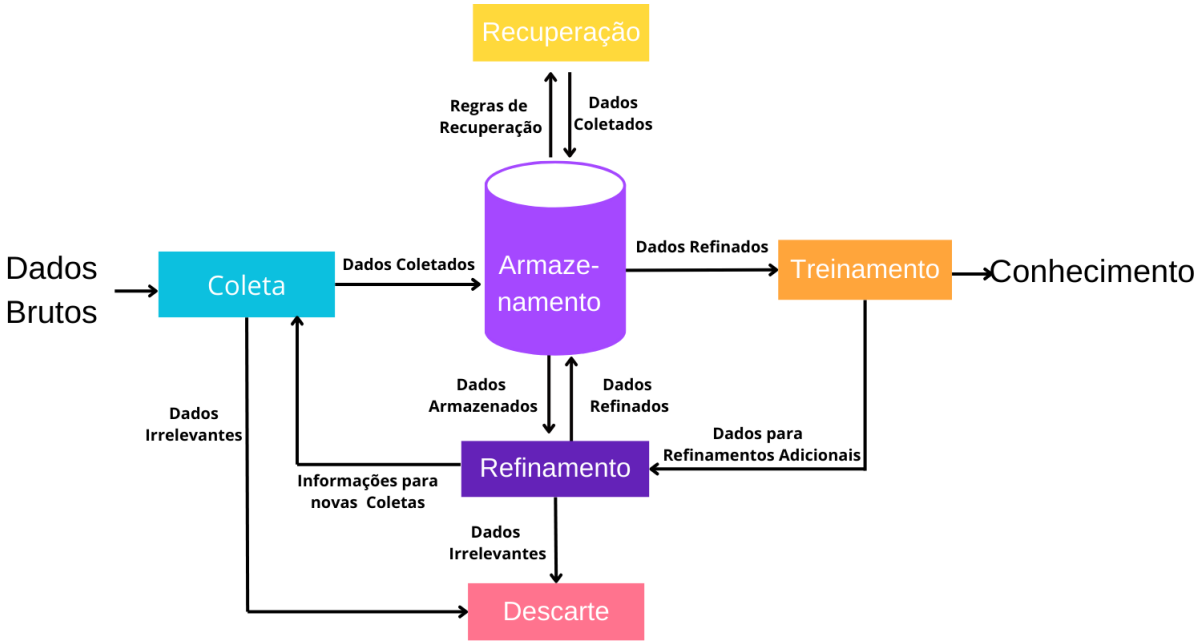

Figura 3.1 – Modelo CRD–TCA.

Este modelo propõe uma nova abordagem sobre os pilares do modelo CVD-CI. O modelo CRD-TCA é sustentado pelos pilares fundamentais de privacidade, integração, qualidade, respeito aos direitos autorais e preservação de dados. Esses pilares são ilustrados na Figura 3.2 e orientam as etapas do modelo de acordo com a necessidade e relevância para cada etapa.

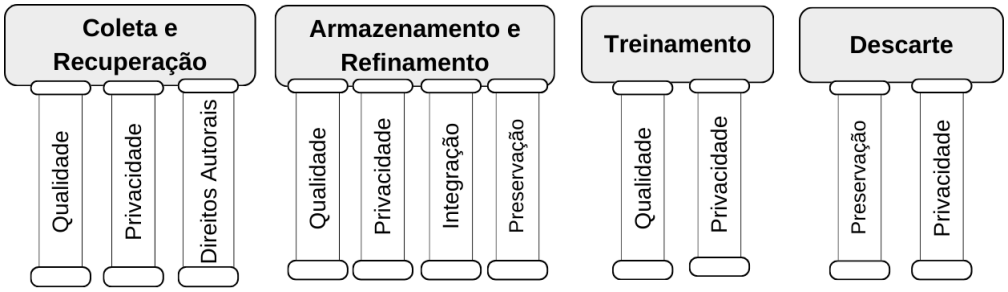

Figura 3.2 – Pilares do modelo CRD–TCA.

A seguir, detalhamos todas as etapas do modelo proposto, abordando os pilares a serem considerados em cada etapa, bem como a aplicação de cada etapa no contexto do projeto PCS, cujo objetivo é a assistência ao diagnóstico médico.

3.1 Coleta

A coleta de dados é a etapa mais importante do projeto de Ciência de Dados, pois os dados são a base de todas as etapas do modelo. A aquisição de dados pode ocorrer através de pesquisa de dados (coleta manual) ou de fontes de dados existentes (coleta digital). Independentemente do método escolhido, os seguintes pilares devem ser considerados:

- Qualidade: O planejamento é crucial na coleta manual, particularmente quando os dados ainda não existem e precisam ser coletados. Isso pode envolver a criação de perguntas e respostas padronizadas para garantir a qualidade da coleta de dados e evitar múltiplas respostas para a mesma pergunta. No caso da coleta digital, que se baseia em dados existentes, é essencial identificar os repositórios e determinar se os dados estão devidamente padronizados.

- Privacidade: Na coleta manual, o acesso aos dados deve ser limitado às pessoas responsáveis pela coleta, que devem possuir treinamento técnico apropriado. No contexto da coleta digital, é essencial definir níveis de acesso para prevenir violações de confidencialidade e garantir conformidade com as leis que regulam a proteção de dados digitais.

- Direitos Autorais: É essencial obter autorização documentada do proprietário dos dados para uso de acordo com as diretrizes éticas e a legislação vigente.

Para documentar a coleta manual e digital, foram elaborados os formulários apresentados nas Figuras 3.3 e 3.4, que podem ser aplicados em qualquer domínio.

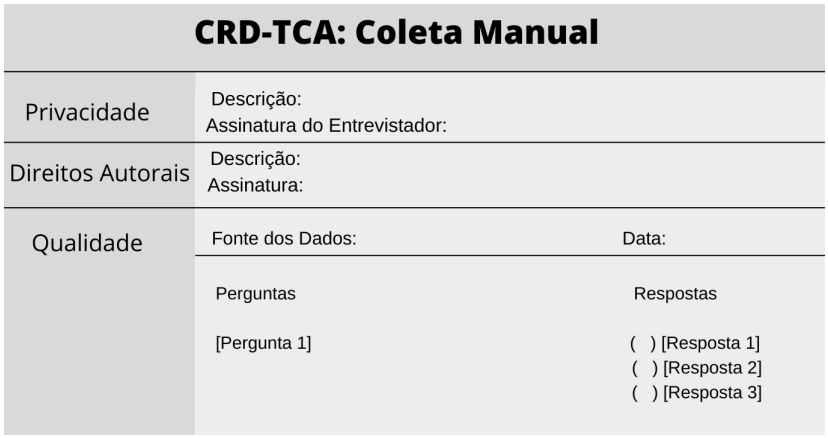

Figura 3.3 – Formulário para Coleta manual do modelo CRD–TCA.

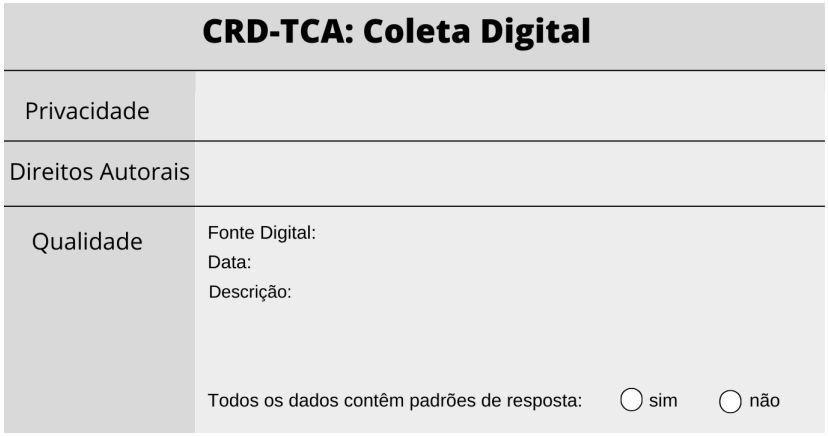

Figura 3.4 – Formulário para Coleta digital do modelo CRD–TCA.

3.1.1 Etapa de Coleta no projeto PCS

No projeto PCS, foram utilizados dados originados de um estudo de pesquisa científica relatado na seção 2.1.5. Este estudo foi aprovado pelo Comitê de Ética em Pesquisa da Universidade José do Rosário Vellano, em colaboração com a Universidade Federal de Alfenas (números de protocolo 149718 e 415856). É importante enfatizar que, após receberem as informações, todos os voluntários forneceram seu consentimento para participar do estudo assinando um Termo de Consentimento Informado [4].

Nesta pesquisa científica, foram coletadas 1.027 amostras de trabalhadores rurais no sul de Minas Gerais, Brasil. A coleta foi realizada manualmente, utilizando uma ficha de investigação de exposição a agrotóxicos, disponível no Anexo A. Essa ficha abrange informações socioeconômicas, histórico de intoxicação por agrotóxicos e hospitalização, além do uso de equipamentos de proteção individual. Também foram coletadas amostras de sangue para a avaliação de biomarcadores de exposição a agrotóxicos, bem como a detecção de possíveis sequelas renais e hepáticas.

É crucial destacar que uma parte significativa desses dados apresentava níveis de ruído, com

apenas alguns pontos de dados caracterizados pela presença de padrões de resposta consistentes. Essa coleta manual foi digitalizada em uma planilha, que deu origem à coleta digital do projeto PCS, documentada conforme o modelo CRD–TCA, na Figura 3.5.

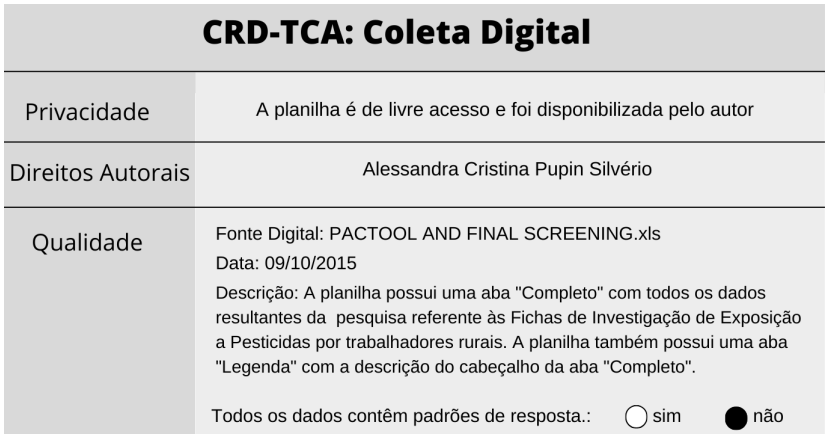

Figura 3.5 – Formulário de Coleta digital do modelo CRD–TCA aplicado ao projeto PCS.

3.2 Descarte

Informações irrelevantes podem potencialmente desviar e diminuir a eficiência computacional. Portanto, quando certos elementos são identificados como irrelevantes no conjunto de dados utilizado para apoiar a tomada de decisões, sua exclusão torna-se essencial. Essas informações devem ser arquivadas em um repositório separado para garantir a privacidade e preservar os dados para possível uso futuro.

3.3 Armazenamento

Para persistir os dados, é necessário a utilização de um modelo físico e lógico estruturado. Atualmente os mais usados são os modelos de banco de dados relacionais, banco de dados orientado a objetos e planilhas eletrônicas.

3.3.1 Bancos de Dados Relacionais

Nos bancos de dados relacionais os dados são organizados de acordo com o contexto em subconjuntos denominados tabelas e essas tabelas através de seus próprios dados criam relações entre si. Cada dado da tabela é representado por uma coluna e cada coluna armazena somente um formato de dado, conforme mostra a Figura 3.6. As linhas representam uma única entidade da tabela, conforme ilustrado na Figura 3.7.

Geralmente os bancos de dados relacionais estão acoplados a um Sistema gerenciador de Banco de Dados (SGBD), que já possui suporte para atender os quatro pilares descritos no modelo CRD–TCA para a etapa de armazenamento.

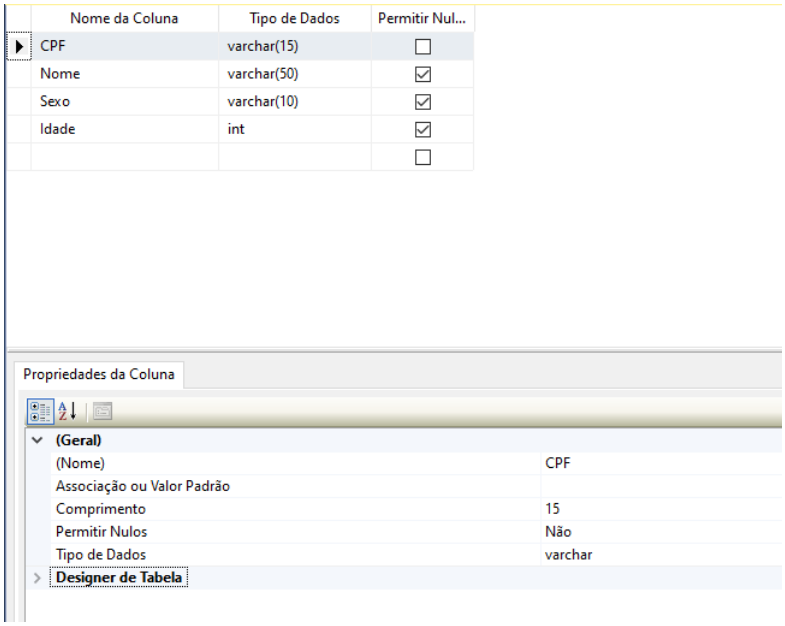

Figura 3.6 – Exemplo de definição de dados em bancos de dados relacionais.

	CPF	Nome	Sexo	Idade
	333.333.333-33	Adolfo	Masculino	20
	444.444.444-44	Luísa	Feminino	22
	555.555.555-55	Miguel	Masculino	18

Figura 3.7 – Exemplo de tabela em bancos de dados relacionais.

- Privacidade: Permite que múltiplos usuários acessem os dados de forma segura, através da definição de senhas e restrições de acesso aos dados.
- Integração: As relações entre as tabelas permitem a integração dos dados de forma consistente. Facilidade de importação e exportação através de arquivo texto, viabilizando a integração dos dados com outros modelos de armazenamento.
- Qualidade: Sendo o modelo projetado devidamente de acordo com as normas exigidas para o banco de dados relacional, os dados não apresentarão inconsistências e serão de fácil acesso. Os SGBDs possuem uma linguagem estruturada de consulta e manutenção dos dados chamada “Structured Query Language” (SQL) que garantem informações de forma rápida, seguras e de qualidade.
- Preservação: Possui arquivo de metadados ou dicionário de dados, que são informações que descrevem os dados, como nome do dado, formato e descrição textual com informações sobre o dado, permitindo a manutenção eficaz ao longo do tempo. As facilidades de Integração com outras plataformas também garantem a preservação dos dados com o avanço tecnológico.

3.3.2 Bancos de Dados Orientados a Objetos

Nos bancos de dados orientados a objetos, os dados são agrupados de acordo com seu contexto em estruturas denominadas classes. As classes mantêm a definição lógica dos dados como o nome e o formato dos dados. Para cada classe é possível criar várias instâncias de valores denominadas objetos da classe, que serão persistidas pelo banco de dados, conforme ilustra a Figura 3.8.

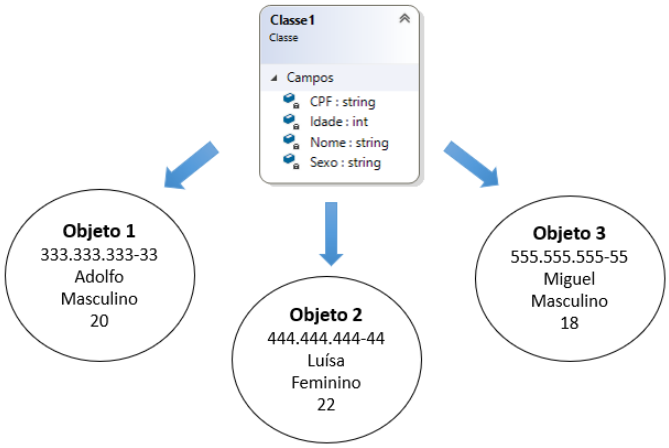

Figura 3.8 – Modelo de dados orientado a objetos.

Existem dois diferenciais para o uso desse modelo, a primeira é que atualmente na área de desenvolvimento de software, as linguagens orientadas a objetos são as mais utilizadas, sendo assim a integração com um banco de dados orientado a objetos se faz de forma natural. A segunda é que nesse modelo é fácil trabalhar com dados complexos, como textos longos e imagens. Porém, devido ao acesso rápido dos dados e a tradição no uso, o modelo de bancos de dados relacionais se destaca em relação ao uso de bancos de dados orientado a objetos. Os sistemas gerenciadores de banco de dados orientado a objetos (SGBDOO), tal como os SGBD relacionais possuem mecanismos que dão suporte a todos os pilares que sustentam a etapa de armazenamento do modelo CRD–TCA [66].

3.3.3 Planilhas Eletrônicas

As planilhas eletrônicas são modeladas na forma de uma matriz onde cada posição de linha e coluna é denominada célula. A primeira linha da matriz pode ser usada para a descrição de cada coluna. As planilhas são sistemas abertos de fácil modelagem e não exige que existam regras entre os dados. As planilhas dispõem de muitos recursos de cálculos e gerações de gráficos para análise. Com um conhecimento de lógica de programação é possível desenvolver lógicas que envolvam os dados disponíveis na planilha. A Figura 3.9, mostra o exemplo de uma planilha. Quanto aos pilares que sustentam a etapa de armazenamento do modelo CRD–TCA, as planilhas oferecem alguns recursos:

- Privacidade: Permitem a definição de acesso de usuários através de senhas, porém não oferecem recursos de restrições de privacidade por dados.
- Integração: A integração com outras tecnologias pode ser realizada através de arquivo texto, não apresenta integração entre os dados.
- Preservação: A preservação através de metadados pode ser planejada e anexada à planilha, através de uma nova planilha com esse objetivo. A preservação também se dá pela facilidade de integração por arquivo texto com tecnologias novas.
- Qualidade: Esse quesito deixa a desejar com relação aos bancos de dados, pois os dados não precisam estar organizados por contexto ou formatos e não possuem interação entre eles, dificultando a criação de visualizações que permitam analisar os dados de diferentes formas.

	A	B	C	D	E
1	CPF	Nome	Sexo	Idade	
2	333.333.333-33	Adolfo	Masculino	20 anos	
3	444.444.444.44	Luísa	Feminino	22	
4	555-555-555-55	Miguel	Masc.	18,5	
5					

Planilha1

Figura 3.9 – Exemplo de planilha eletrônica.

3.3.4 Etapa de Armazenamento no projeto PCS

A abordagem CRD–TCA utiliza o modelo de banco de dados relacional em conjunto com um SGBD para garantir a segurança, integração, qualidade e preservação dos dados armazenados, possibilitando análises das informações contidas no repositório. Esta análise permite a identificação de dados irrelevantes que podem ser subsequentemente descartados e, quando necessário, a criação de dados para informações ausentes, mas essenciais para a pesquisa em questão. No projeto PCS, os dados obtidos durante a fase de coleta foram armazenados em uma planilha. No entanto, nesse formato de armazenamento, os dados carecem de interação entre si, dificultando a criação de visualizações que permitam a análise dos dados de diferentes maneiras. Para cumprir os objetivos da etapa de armazenamento, o conteúdo da planilha foi transferido para um banco de dados relacional. Posteriormente, foi possível excluir informações consideradas irrelevantes para o processo de diagnóstico, como detalhes de endereço e avaliações de serviços médicos. Além disso, elementos essenciais foram incorporados, a saber, um "ID" usado para a identificação única de cada amostra e "Diagnóstico" empregado para a categorização e rotulagem de cada amostra. A Figura 3.10 mostra o formulário recomendado pelo modelo CRD–TCA para documentar a etapa de armazenamento e exibir dados relacionados ao projeto PCS. Este formulário fornece os nomes dos campos designados para o armazenamento de cada informação dentro do banco de dados escolhido, com o objetivo de permitir ao cientista

de dados inserir descrições textuais para cada campo, marcar campos para exclusão e adicionar novos campos conforme necessário.

CRD-TCA: Formulário de Armazenamento

Banco de dados: Pesquisa
Tabela: Amostra_Original

Nome original do dado	Nome do campo na tabela	Formato
Identificador	Id	int
Número do Formulário	Sujeito	varchar (50)
Município de residência	Residencia	varchar (50)
Sexo	Sexo	varchar (50)
Gestante	Gestante	varchar (50)
⋮	⋮	⋮
Diagnóstico	Diagnostico	varchar (20)

Figura 3.10 – Formulário de Armazenamento do modelo CRD–TCA aplicado ao projeto PCS.

O formulário com todos os dados do projeto PCS encontra-se no Anexo B.

3.4 Refinamento

Refinar, apurar ou aprimorar os dados, é uma tarefa importante para a extração do conhecimento. Normalmente os algoritmos de extração do conhecimento são projetados para trabalhar com dados normalizados ou com categorias de dados. Dados numéricos podem assumir valores contínuos que dependendo da necessidade precisam ser transformados em dados discretos. Essa transformação é conhecida como discretização. Após a discretização os dados podem ser tratados como dados nominais. Seja o exemplo do dado altura de uma Pessoa, que pode assumir qualquer valor numérico, esse pode ser mapeado para valores discretos nominais como “baixo” e “Alto”. Essas transformações reduz o conjunto dos dados e torna a informação mais concisa e interpretável. Existem vários trabalhos que propõem métodos estatísticos como o uso de histograma e algoritmos de Inteligência Artificial para realizar a discretização dos valores de dados contínuos em faixas de valores discretos [67]. Porém essa discretização geralmente é realizada pelo especialista do domínio dos dados. Um modelo de previsão preciso é baseado em um eficiente conceito de transformação de valor de domínio [68].

Em Ciência de Dados, o processo de refinamento de dados é uma etapa crítica que envolve procedimentos essenciais, como lidar com valores ausentes, padronização, eliminação de informações irrelevantes e transformação quando necessário. Para manter a preservação e qualidade dos dados ao longo do tempo, o modelo CRD–TCA propõe o uso de um formulário de Refinamento, cuidadosamente elaborado para documentar o processo de refinamento de cada dado armazenado, enquanto a privacidade e integração dos dados são mantidas pelo SGBD relacional. Este formulário abrange informações cruciais, como a natureza dos dados, da seguinte forma:

- Contínuo: Se é um dado que representa valores contínuos, como valores numéricos.

- Discretizado: Valores que já se encontram discretizados obedecendo regras e padrões.
- Multivalorado: Valores que se encontram em forma de textos.
- Nulo: Valores nulos, inexistentes, que não constam no armazenamento.

A natureza dos dados influencia significativamente na estratégia de refinamento. Para dados contínuos, é necessário estabelecer regras de discretização para criar intervalos definidos, tornando-os adequados para análises subsequentes. Para dados discretizados, é essencial documentar as regras previamente aplicadas, possibilitando a compreensão e replicação do processo. Por fim, ao lidar com dados nulos ou multivalorados, é aconselhável avaliar a viabilidade de retornar à fase de coleta.

O formulário de Refinamento deve sinalizar também o status dos dados durante o processo de refinamento, indicando se está concluído ou deve ser descartado.

Nesta fase, a colaboração do especialista do domínio desempenha um papel crucial, indicando quais dados podem ser descartados e orientando o refinamento de acordo com seu conhecimento, garantindo uma representação adequada dos dados.

3.4.1 Etapa de Refinamento no projeto PCS

A Figura 3.11 apresenta o formulário de Refinamento do modelo proposto aplicado a alguns dados do projeto PCS. Neste contexto, “Creatinina” é destacada como um exemplo de dado contínuo que passou por discretização. Os dados “Produto que teve contato pela última vez” e “Diagnóstico”, ainda não estão concluídos, porque dados do tipo multivalorado e nulo são mais complexos para discretizar, e isso será explicado na próxima seção. O dado “Relação de Trabalho” já foi discretizado na etapa de coleta, mas foi identificado como supérfluo para o domínio do problema e considerado como descarte.

O formulário completo de refinamento dos dados do projeto PCS, encontra-se no Anexo C

3.4.2 Refinamento de Dados Multivalorado ou Nulo

Quando os dados se encontram multivalorado ou nulo, é necessário retornar a etapa de coleta, conforme mostra a Figura 3.12, para analisar a possibilidade de uma nova coleta dos dados de forma padronizada. Caso não seja possível realizar a coleta na mesma fonte, deve-se verificar a possibilidade de fontes alternativas.

Para dados do tipo multivalorado que podem ser agrupados em classes discretizadas, pode-se verificar a possibilidade de usar as informações da resposta do dado textual, como padrões de perguntas. Nesse caso, o modelo CRD–TCA propõe um formulário para refinamento de dados multivalorados, onde se sugere desagregar a informação em unidades individuais e atribuir uma nova classificação a cada unidade da informação.

No caso de dados do tipo nulo, que não apresentam nenhuma informação, é necessário voltar à etapa de coleta para realizar a coleta do dado de forma padronizada.

CRD-TCA: Formulário de Refinamento

Nome Original	Tipo	Discretização	Concluído
CREATININA	Contínuo	Homens: [Val < 0,9] Baixo, [Val ≥ 0,9 e Val ≤ 1,3] Normal Mulheres: [Val < 0,6] Baixo, [Val ≥ 0,6 e Val ≤ 1,1] Normal	Sim
Produto teve contato pela última vez	Multivalorado		Não
Relação de Trabalho	Discretizado	Proprietário, Assalariado, Meeiro, Volante	Descarte
Diagnóstico	Nulo		Não

Figura 3.11 – Formulário de Refinamento do modelo CRD–TCA aplicado ao projeto PCS.

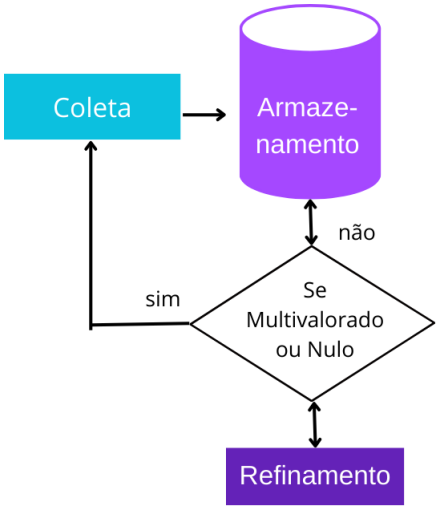

Figura 3.12 – Ciclo do Refinamento de dados do tipo multivalorado ou nulo.

3.4.3 Exemplo de Refinamento de Dados Multivalorado ou Nulo no projeto PCS

No projeto PCS, foram coletadas informações relacionadas ao último produto com o qual o trabalhador rural teve contato. Os entrevistados forneceram uma lista de produtos, como "Polyram, Supera, Bitrin", em formato de texto separado por vírgulas, indicando os nomes das marcas desses produtos. No entanto, para fins de análise, busca-se a categorização química desses produtos em vez de seus nomes de marca. Uma nova coleta de dados foi necessária, introduzindo categorias específicas para representar a classificação química. Cada resposta foi avaliada individualmente e associada à respectiva classificação química, conforme apresentado no formulário de refinamento de dado multivalorado do modelo CRD–TCA na Figura 3.13. O resultado completo do refinamento referente à classificação química dos produtos, encontra-se no Anexo D.

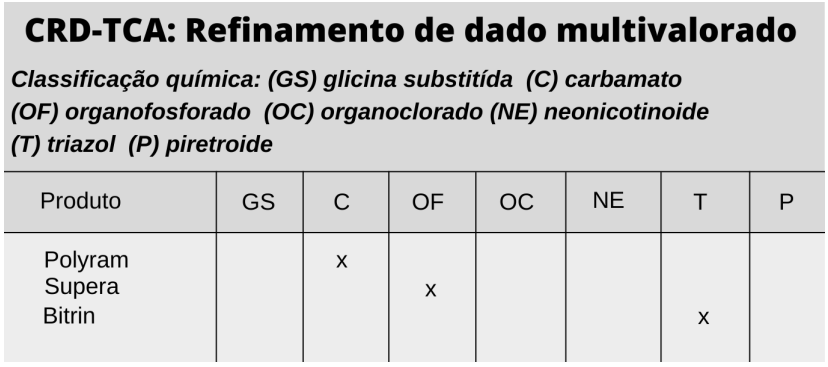

Figura 3.13 – Formulário de Refinamento de dado multivalorado.

No domínio do problema desse trabalho, foi identificado e armazenado um dado do tipo nulo, que é o dado "Diagnóstico". Justamente o dado que define o objetivo do domínio desse trabalho, sendo de suma importância no aprendizado de máquina da etapa de treinamento. A sua coleta foi realizada com base em dados existentes no banco de dados. A fonte da coleta foi o profissional do domínio. A Tabela 16 apresenta as classes de discretização possíveis para o diagnóstico, definidas pelo profissional do domínio.

Tabela 16 – Tipos de diagnósticos, especificações e acrônimos correspondentes.

Tipo	Especificação	Diagnóstico
Intoxicação Aguda	CH_P: Com Atividade Reduzida	IA–CHPD
	CH_P: Com Atividade Reduzida Altas Concentrações de Pesticidas	IA–AP–CHPD
	CH_P: Com Atividade Reduzida Baixas Concentrações de Pesticidas	IA–BP–CHPD
	CH_P e CH_E: Com Atividade Reduzida Creatinina Alterada (Problema Renal) Altas Concentrações de Pesticidas	IA–AP–CHPD–CHED–PRCR
	CH_P: Com Atividade Reduzida AST e ALT Alterados (Lesão Hepática)	IA–CHPD–ASTALT–LH
Intoxicação Crônica	CH_E: Com Atividade Reduzida	IC–CHED
	CH_E: Com Atividade Reduzida AST Alterado (Lesão Hepática)	IC–CHED–AST–LH
	Não Apresenta Intoxicação	NAI
	Não Tem Diagnóstico (Ausência de informações)	NTD

Após a definição dos tipos de diagnósticos, um formulário específico para a coleta do dado "Diagnóstico" foi desenvolvido, permitindo a seleção da sigla referente ao diagnóstico para cada amostra. A Figura 3.14 mostra esse formulário, que foi criado com a utilização de uma ferramenta de desenvolvimento de software da Microsoft, o Visual Studio 2022, que possui bibliotecas de acesso ao banco de dados usado nesse trabalho.

3.5 Treinamento

A etapa de Treinamento está relacionada ao aprendizado de máquina, uma parte importante da área de Inteligência Artificial. A Teoria de Conjuntos Aproximados (TCA) é um modelo de aprendizado de máquina que oferece um método de seleção, ajudando a identificar e classificar

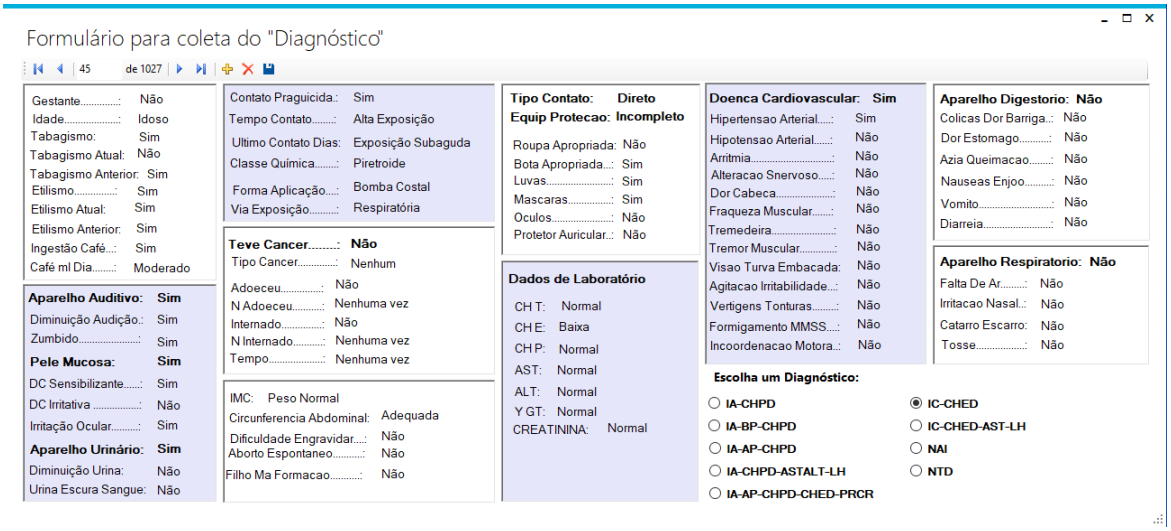

Figura 3.14 – Formulário para Coleta do dado “Diagnóstico”.

os recursos mais importantes em um conjunto de dados supervisionado, que possui um grande número de informações. A Teoria de Conjuntos Aproximados busca descrever aproximadamente o conhecimento impreciso a partir de um conhecimento preciso, lidando com incerteza nos dados. Embora a implementação clássica da teoria geralmente utilize algoritmos determinísticos, nem todos os métodos e implementações seguem essa abordagem, podendo variar em termos de determinismo conforme as escolhas específicas de modelagem e implementação. [14] [61].

Para iniciar o treinamento na TCA, os dados precisam estar devidamente preparados, e o modelo CRD–TCA se preocupa com essa preparação desde a primeira etapa do ciclo. Os pilares que sustentam a etapa de Treinamento do modelo CRD–TCA são a privacidade e a qualidade dos dados. Quanto à privacidade, esta é garantida em todo o modelo pelo SGBD. O SGBD também garante a qualidade dos dados em todo o modelo, evitando redundâncias e garantindo a integridade dos dados. A qualidade referente à discretização dos dados é alcançada ao final da etapa de Refinamento. Assim, a etapa de Treinamento recebe dados que já se encontram qualificados para esse propósito. Nessa etapa, portanto, a qualidade deverá ser medida pelo grau de satisfação dos resultados obtidos pelo treinamento, em conjunto com o profissional do domínio.

Como a TCA é um modelo de aprendizado de máquina supervisionado, os dados de condição e o dado de decisão precisam estar bem definidos. O conjunto de dados de condição pode conter todos os dados do conjunto de dados, ou podem conter conjuntos parciais, resultante de seleções de alguns dados do conjunto total dos dados. Contudo, cada conjunto, preparado para o treinamento, deve conter um dado de decisão na última posição do conjunto de dados. O treinamento com o algoritmo da TCA, possui duas fases, a geração de redutos e a extração de regras.

3.5.1 Geração de Redutos

Nessa primeira fase do treinamento, a geração de redutos, será realizada uma busca por subconjuntos que contenham a mesma relação de equivalência e o mesmo poder de decisão que o conjunto original. Essa busca pode ser realizada por algoritmos clássicos com o cálculo do grau de dependência ou através da matriz de discernibilidade. Os subconjuntos encontrados são chamados de redutos pela TCA.

Para essa fase, além da definição dos dados que farão parte do conjunto de treinamento, é necessário definir também quais exemplos, referentes a esses dados, irão fazer parte do treinamento. O objetivo do aprendizado de máquina supervisionado é construir um classificador baseado em exemplos, de forma que, ao final do treinamento, a máquina consiga classificar exemplos que não possuam valores para o dado de decisão.

É importante documentar o treinamento conduzido na geração de redutos, fornecendo as seguintes informações:

- Fonte dos dados e o número de exemplos ou amostras utilizados no treinamento.
- Conjunto de atributos selecionados como conjunto de condição e atributo de decisão. O conjunto de atributos de condição pode consistir em todos os dados de condição ou ser parcial, consistindo em subconjuntos dos atributos de condição.

O modelo CRD–TCA adota a técnica de Validação Cruzada (CV), empregando especificamente a validação cruzada k–fold [69], permitindo a definição do número k de partições e o volume de amostras em cada partição. O conjunto de dados é dividido em k subconjuntos. O modelo é então treinado k vezes, com cada iteração usando um subconjunto diferente como o conjunto de validação, enquanto os subconjuntos restantes formam coletivamente o conjunto de treinamento. Esse processo iterativo fornece uma avaliação mais robusta do desempenho do modelo, garantindo que cada ponto de dado seja usado tanto para treinamento quanto para validação ao longo das k iterações.

3.5.2 Etapa de Treinamento com Geração de Redutos no projeto PCS

No projeto PCS, foi realizada a CV seguindo estes passos:

- Foram selecionadas amostras de acordo com o tipo de diagnóstico, resultando em nove seleções distintas.
- Em seguida, essas amostras foram distribuídas para criar quatro partições de tamanhos proporcionais, cada uma contendo aproximadamente a mesma porcentagem para cada tipo de diagnóstico.

Essas partições foram usadas para conduzir quatro treinamentos distintos, alternando a partição de teste denominada Testes A, B, C e D, enquanto as partições restantes passaram por treinamento, conforme ilustrado na Figura 3.15.

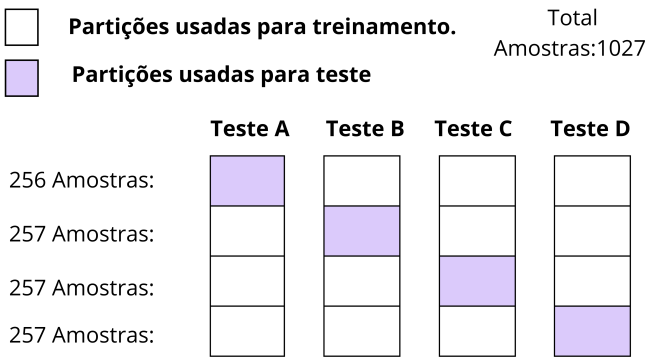

Figura 3.15 – Testes para validação cruzada.

A etapa de refinamento forneceu um conjunto de dados com 79 atributos de condição e um atributo de decisão para a etapa de treinamento. Foram organizados 12 agrupamentos diferentes com atributos de condição em colaboração com um especialista do domínio, conforme apresentado na Tabela 17. Foram criados 11 agrupamentos parciais, cada um com o propósito de analisar informações de testes laboratoriais juntamente com diferentes dados. Estes incluíram dados de testes laboratoriais combinados com informações sobre sistemas de órgãos específicos, como o sistema nervoso, cardiovascular, entre outros. Além disso, foram obtidos grupos combinando dados de testes laboratoriais com informações de exposição a pesticidas, dados pessoais e classes químicas. Em todos esses agrupamentos, o diagnóstico foi utilizado como atributo de decisão. Adicionalmente, foi criado um agrupamento abrangendo todos os dados de condição (Grupo 12). Para a realização dos agrupamentos utilizou-se a linguagem de consulta própria do SGBD.

Tabela 17 – Grupos de Treinamento.

Grupo	Descrição dos atributos do conjunto de condição	Total de atributos
1	Dados de contato com pesticidas, testes laboratoriais e classe química	22
2	Dados pessoais e dados de testes laboratoriais	23
3	Classe química e dados de testes laboratoriais	9
4	Dados clínicos sobre sintomas cardiovasculares e testes laboratoriais	12
5	Dados clínicos sobre alterações no sistema nervoso e testes laboratoriais	18
6	Dados clínicos sobre alterações no sistema digestivo e testes laboratoriais	15
7	Dados clínicos sobre alterações no sistema respiratório e testes laboratoriais	13
8	Dados clínicos sobre alterações no sistema auditivo e testes laboratoriais	11
9	Dados clínicos sobre alterações na pele e testes laboratoriais	12
10	Dados clínicos sobre alterações no trato urinário e testes laboratoriais	11
11	Dados sobre o diagnóstico de câncer e testes laboratoriais	10
12	Conjunto completo com todos os dados de condição	80

Esses 12 grupos foram empregados como critérios de seleção nas partições de treinamento resultantes da CV para condução dos testes A, B, C e D.

Para a execução dos treinamentos, utilizou-se o software RStudio ®, versão 1.4.1103, e a biblioteca RoughSets, versão 1.3-7 [70] [71]. A implementação foi realizada em dois scripts diferentes. O primeiro, denominado script 1, tem como objetivo encontrar todos os redutos

possíveis, usados no treinamento dos grupos 1 a 11. O segundo, denominado script 2, faz a busca por um único reduto e foi utilizado somente no grupo 12. A descrição das instruções utilizadas nesses dois scripts e a forma como foram implementados encontra-se no Anexo E.

A Figura 3.16 apresenta o formulário do modelo CRD–TCA usado para documentar os resultados do treinamento na geração de redutos. Por exemplo, Por exemplo, 18 atributos, dos quais 17 são de condição e um é de decisão, relacionados ao Grupo 5, foram selecionados e aplicados às partições de treinamento para o teste A. Durante este treinamento, identificamos sete subconjuntos com o mesmo poder de decisão do conjunto original.

CRD-TCA: Formulário de Treinamento - Geração de Redutos

Número de identificação: Grupo 5A **Data: 05/01/2023**

Fonte: Banco de Dados de Pesticidas - Tabela: Amostra
Conjunto de Condições: Alteração_Snervoso, Dor_Cabeça, Fraqueza_Muscular, Tremores, Tremor_Muscular, Visão_Embaçada, Agitação_Irritabilidade, Vertigem_Tontura, Formigamento_MMSS, Incoordenação_Motora, CH_T, CH_E, CH_P, AST, ALT, Y_GT, CREATININA.
Dados de Decisão: Diagnóstico **Total de Dados de Conjunto de Condições e Decisão: 18**

Redutos	Número de dados
CH_T, CH_E, CH_P, AST, CREATININA	5
CH_T, CH_E, CH_P, AST, Alteração_Snervoso	5
CH_T, CH_E, CH_P, AST, Dor_Cabeça	5
CH_T, CH_E, CH_P, AST, Incoordenação_Motora	5
CH_T, CH_E, CH_P, AST, Fraqueza_Muscular, Formigamento_MMSS	6
CH_T, CH_E, CH_P, AST, Tremores, Formigamento_MMSS	6
CH_T, CH_E, CH_P, AST, Tremor_Muscular, Formigamento_MMSS	6

Figura 3.16 – Formulário de Treinamento com Geração de Redutos do modelo CRD–TCA aplicado ao projeto PCS.

Entre os 12 agrupamentos treinados, o Grupo 5 confirmou a importância de alterações no Sistema Nervoso Central em indivíduos expostos a pesticidas, de acordo com os resultados apresentados na pesquisa científica conduzida por Silverio et al. [4].

3.5.3 Extração de Regras

Na segunda fase da etapa de Treinamento, inicia-se a extração de regras dos redutos considerados relevantes pelo profissional do domínio. As regras são formadas através da leitura condicional dos dados de condição em relação ao dado de decisão das instâncias da tabela de decisão gerada pelo reduto escolhido. Supondo que para o conjunto de condição $C = \{a, b, c, d, e\}$ e o dado de decisão $D = \{f\}$, o reduto escolhido seja $R = \{\{b, c\}\}$. E as instâncias de $\{b, c\}$ sejam representadas pela tabela de decisão ilustrada na Tabela 18.

Tabela 18 – Tabela de Decisão

b	c	f
1	0	0
1	1	1

A extração das regras é feita através da leitura condicional da seguinte forma:

- Se $b = 1$ e $c = 0$, então $f = 0$.
- Se $b = 1$ e $c = 1$, então $f = 1$.

Para estabelecer a origem das regras geradas, é crucial especificar o número de identificação do formulário que descreve a geração dos redutos, bem como o reduto selecionado nesse formulário para a extração das regras.

3.5.4 Etapa de Treinamento com Extração de Regras no projeto PCS

Nessa etapa foi realizada uma análise conduzida pelo especialista do domínio em Toxicologia, que selecionou os redutos considerados relevantes para extração de regras.

A Figura 3.17 apresenta o formulário de extração do modelo proposto, documentando, como exemplo, a seleção do reduto "CH_T, CH_E, CH_P, AST, CREATININA" a partir dos resultados do treinamento do Grupo 5 no Teste A, que resultou na geração de 27 regras de decisão.

CRD-TCA: Formulário de Treinamento - Extração de Regras

Grupo de Referência: Grupo 5A

Fonte: Banco de Dados de Pesticidas - Tabela: Amostra

Reduto Selecionado: CH_T, CH_E, CH_P, AST, CREATININA.

Número de Regras: 27

Armazenamento: Regras-G5A.csv

Figura 3.17 – Formulário de Treinamento com Extração de Regras do modelo CRD–TCA aplicado ao projeto PCS.

As regras associadas à aplicação deste reduto, que foram geradas durante o treinamento do Grupo 5 no Teste A, são apresentadas na Tabela 20 na Seção 4.2.

Após a geração das regras para todos os grupos e testes, foi desenvolvido um formulário específico para a avaliação das regras geradas para o diagnóstico, permitindo a seleção do arquivo referente às regras e o nome do grupo e teste no qual as regras serão avaliadas. A Figura 3.18 mostra esse formulário, desenvolvido no software da Microsoft, Visual Studio 2022. O código implementado encontra-se no Anexo G.

3.6 Recuperação

Nessa etapa, o formulário final de refinamento de dados é recuperado e disponibilizado para a coleta de novas informações, a fim de evitar a aquisição redundante de dados previamente descartados e garantir a conformidade com as categorias e regras de discretização estabelecidas na etapa de refinamento. As regras resultantes da etapa de treinamento também são disponibilizadas para auxiliar na tomada de decisão.

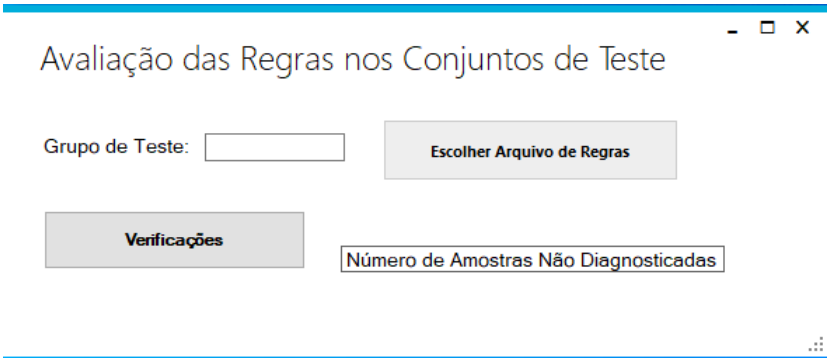

Figura 3.18 – Formulário para Avaliação das Regras no conjunto de teste do projeto PCS.

Novas amostras são incluídas nesta etapa para realizar treinamentos adicionais e aprimorar as regras geradas para a tomada de decisão. Esse procedimento é realizado mantendo o compromisso com os pilares de privacidade, direitos autorais e qualidade, conforme enfatizado durante a fase inicial de coleta de dados.

3.6.1 Etapa de Recuperação no projeto PCS

O modelo DRC-TCA disponibilizou o formulário resultante da etapa de refinamento, juntamente com as regras obtidas ao final da etapa de treinamento, para um Trabalho de Conclusão do Curso de Bacharelado em Ciência da Computação. Nesse trabalho, está sendo desenvolvida a interface, como mostram as figuras 3.19 e 3.20, para o projeto PCS, onde novas amostras serão coletadas e poderão contribuir para melhorar o treinamento.

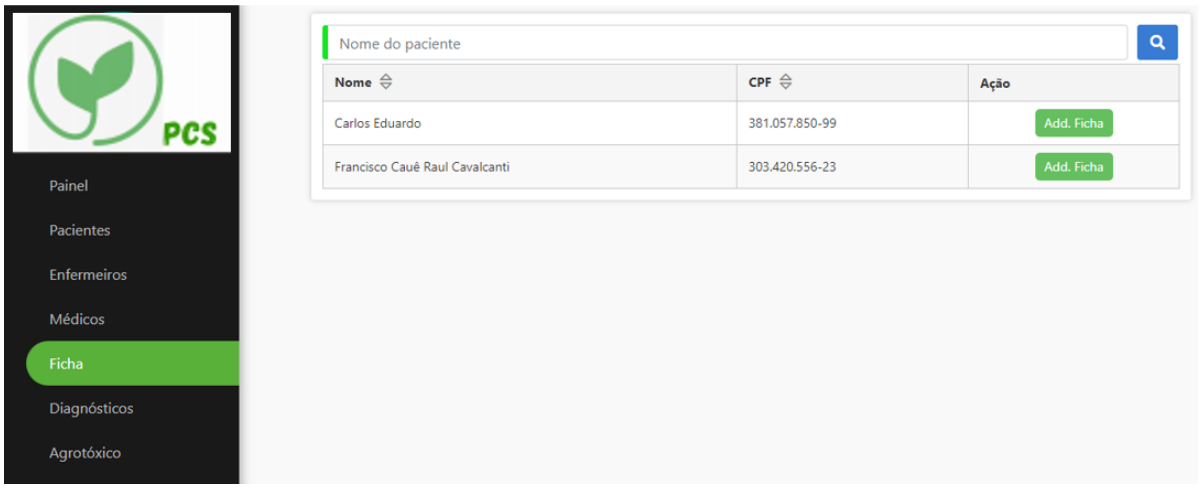

Figura 3.19 – Formulário Principal do projeto PCS.

3.7 Considerações finais

Neste capítulo, foi apresentado o modelo de Ciclo de Refinamento de Dados usando a Teoria de Conjuntos Aproximados e a importância da padronização da documentação nas etapas de

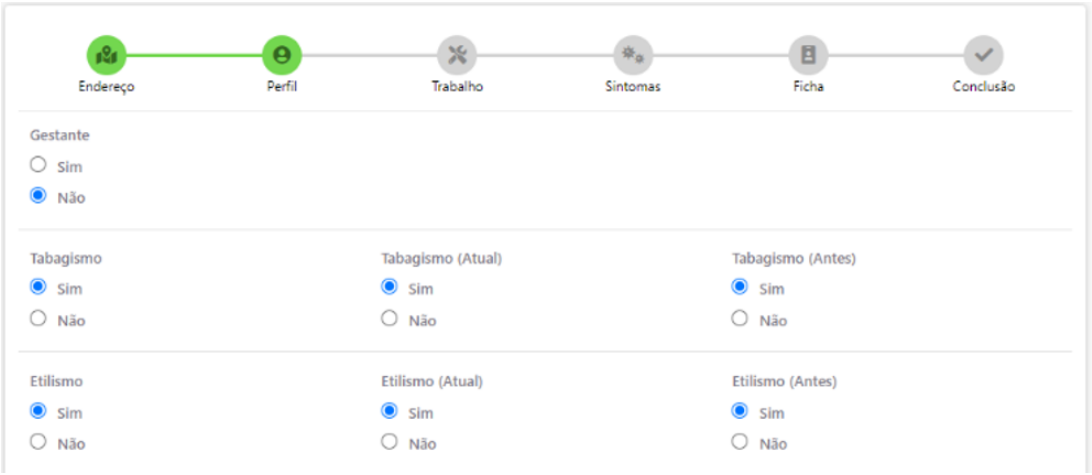

Figura 3.20 – Formulário do projeto PCS.

coleta, armazenamento, refinamento, treinamento, recuperação e descarte dos dados. A criação e implementação do CRD–TCA mostrou-se essencial para a organização e desenvolvimento eficiente de projetos de ciência de dados. A documentação adequada evita resultados insatisfatórios, ineficientes e redundantes, promovendo uma comunicação clara entre os diversos profissionais envolvidos no projeto.

O próximo capítulo abordará os testes e os resultados do modelo no contexto do projeto Plantando e Colhendo Saúde, destacando a aplicabilidade prática do CRD–TCA. Serão analisados os resultados obtidos e as melhorias propostas para otimizar o desempenho do modelo. Esta análise fornecerá uma visão detalhada das capacidades e limitações do modelo, oferecendo insights valiosos para futuras pesquisas e aplicações em diferentes áreas da ciência de dados.

4 Resultados e Discussões

Nesta seção, destacamos primeiramente o modelo CRD–TCA como uma metodologia de KDD, abordando suas contribuições e vantagens no processo de descoberta de conhecimento em bases de dados. Em seguida, exploramos os resultados empíricos da aplicação do modelo aos dados do projeto PCS, demonstrando a eficácia e a aplicabilidade prática do CRD–TCA na análise de dados reais.

4.1 Modelo CRD–TCA como uma Proposta de KDD

O modelo CRD–TCA demonstrou sua eficácia em abordar todas as etapas do KDD de acordo com a hierarquia de Ciência de Dados, conforme ilustrado na Figura 4.1. O modelo proposto fornece uma metodologia documentada que se alinha com cada etapa da pirâmide de KDD.

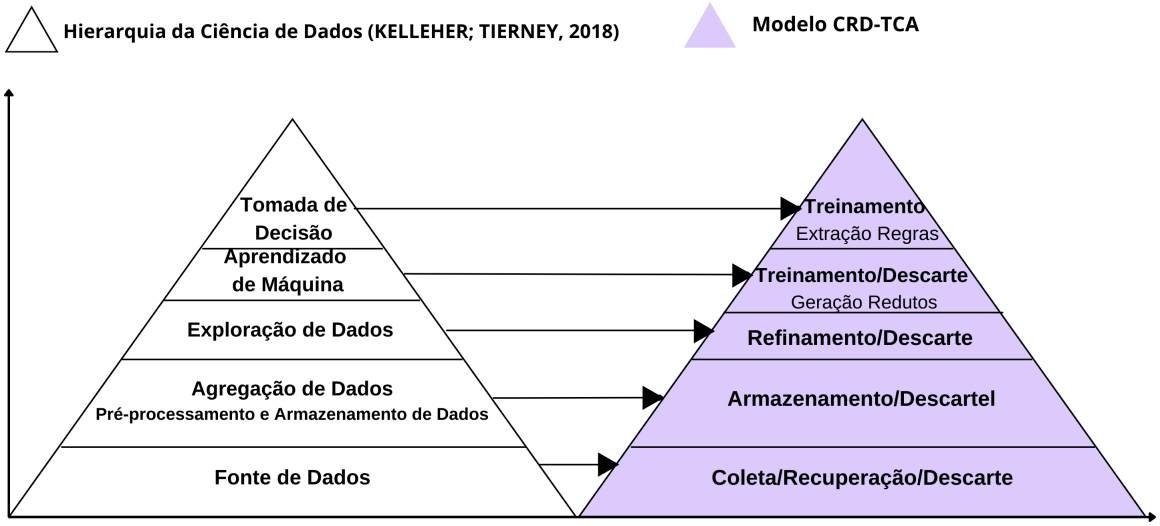

Figura 4.1 – Modelo CRD–TCA aplicado à hierarquia KDD em Ciência de Dados.

Como enfatizado por Kelleher e Tierney [10], é de suma importância que os resultados de um projeto de Ciência de Dados sejam comunicados de maneira acessível, para que até mesmo os membros da equipe sem formação técnica possam entendê-los. Nesse contexto, o modelo CRD–TCA destaca-se pela sua simplicidade, evitando documentação complexa e desnecessária, focando no essencial. Adicionalmente, o modelo oferece implicações práticas, tais como:

- Refinamento Eficaz de Dados: O modelo analisa, discretiza, identifica e elimina redundâncias, inconsistências e ruídos, resultando em dados mais limpos e precisos.
- Redução de Dimensionalidade: O modelo pode contribuir para a redução de dimensionalidade ao identificar atributos menos relevantes, o que é útil em conjuntos de dados extensos.

- Transparência na Tomada de Decisão: A transparência do modelo facilita a aceitação e confiança dos usuários, pois as regras geradas são compreensíveis.

O uso da documentação proposta pelo modelo pode ser aplicado a conjuntos de dados em diversos domínios, como saúde, marketing, finanças, entre outros, tornando o modelo valioso para uma ampla gama de profissionais que contribuem para projetos de Ciência de Dados.

A Figura 4.2 ilustra como o modelo recebeu dados ruidosos de pesquisas científicas anteriores e produziu os seguintes resultados:

- Um conjunto de regras de refinamento, que estabeleceu padrões para a recuperação de dados, expandindo assim o conjunto de amostras disponíveis para futuros treinamentos.
- Um conjunto de regras de decisão, que desempenhou um papel crucial na orientação da tomada de decisões dentro do escopo do projeto. O modelo contribuiu significativamente para o desenvolvimento do projeto PCS, incorporando essas regras como recursos de aprendizado de máquina.

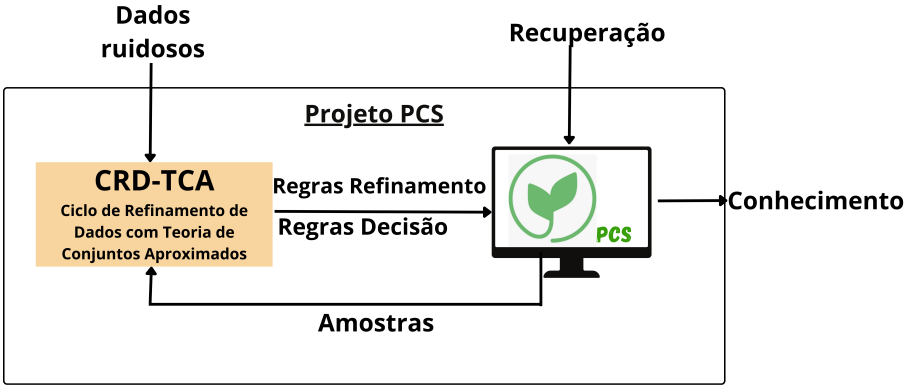

Figura 4.2 – Interação entre os dados da pesquisa científica e o modelo CRD–TCA no projeto PCS.

O modelo CRD–TCA, com sua metodologia cíclica, documentação específica para cada etapa e o uso de aprendizado de máquina, é capaz de lidar com desafios enfrentados pela Ciência de Dados, como grande volume de dados, dados ruidosos e, especialmente, a falta de direção que leva à conclusão do projeto, garantindo sua eficácia ao longo do tempo. No projeto PCS, onde os padrões de intoxicação por agrotóxicos podem evoluir devido a mudanças nas práticas agrícolas, tipos de agrotóxicos utilizados, entre outros fatores, o CRD–TCA, com sua capacidade de recuperação e adaptação a mudanças nos dados, pode lidar com essa dinâmica, garantindo que o modelo permaneça relevante e eficaz na assistência aos profissionais de saúde.

4.2 Resultados da aplicação do modelo CRD–TCA aos dados do projeto PCS.

Através da análise dos resultados obtidos na aplicação do modelo aos dados do projeto PCS, observou-se uma redução progressiva na quantidade de informações da etapa de coleta de dados

até a conclusão da etapa de treinamento. Inicialmente, houve um conjunto de 121 atributos, dos quais 41 foram excluídos ao longo do processo.

Em todas as 48 sessões de treinamento, com base nos 12 agrupamentos aplicados nos Testes A, B, C e D, foram identificados vários redutos. Uma síntese desses resultados encontra-se no Anexo F. A Tabela 19 apresenta os redutos mínimos, com apenas 5 atributos, juntamente com a acurácia de validação das regras geradas.

Tabela 19 – Melhores resultados do treinamento em grupo.

Menores Redutos	Acurácia dos testes				Média
	A (%)	B (%)	C (%)	D (%)	± Desvio Padrão
CH_T, CH_E, CH_P, AST, Ultimo_ContatoDias	95,75	96,89	98,05	98,05	97,19 ± 1,10
CH_T, CH_E, CH_P, AST, Modo_Aplicacao	96,91	98,44	98,05	98,44	97,96 ± 0,72
CH_T, CH_E, CH_P, AST, Tipo_Contato	98,07	99,22	98,05	98,83	98,54 ± 0,58
CH_T, CH_E, CH_P, AST, CREATININA	99,61	99,61	98,44	97,67	98,83 ± 0,95
CH_T, CH_E, CH_P, AST, Circunferencia_Abdominal	98,46	99,22	98,83	98,83	98,84 ± 0,31
CH_T, CH_E, CH_P, AST, Aborto_Expontaneo	99,61	99,22	98,44	98,83	99,03 ± 0,50
CH_T, CH_E, CH_P, AST, Alteracao_SNervoso	99,61	99,61	98,44	98,83	99,13 ± 0,58
CH_T, CH_E, CH_P, AST, Dor_Cabeca	98,84	99,22	99,22	98,83	99,03 ± 0,22
CH_T, CH_E, CH_P, AST, Incordenacao_Motora	99,23	99,22	99,22	98,83	99,13 ± 0,20
CH_T, CH_E, CH_P, AST, Vomito	98,46	99,61	99,22	98,44	98,93 ± 0,58

No contexto da análise apresentada na Tabela 19, é evidente que todos os redutos alcançaram uma acurácia consideravelmente alta. Qualquer um desses redutos pode ser utilizado para classificar as novas amostras. No entanto, neste cenário específico, o profissional do domínio optou pelo reduto "CH_T, CH_E, CH_P, AST, CREATININA" devido à sua acurácia notável, bem como à sua inclusão de informações de análises laboratoriais.

A Tabela 20 apresenta as regras associadas à aplicação deste reduto, que foram geradas durante o treinamento do Grupo 5 no Teste A. Por exemplo, a primeira regra nesta tabela pode ser interpretada da seguinte forma:

Se CH_T = "Baixo", CH_E = "Baixo", CH_P = "Baixo", AST = "Normal" e CREATININA = "Normal", então o diagnóstico é "IA–AP–CHPD".

Recall e acurácia (conforme Equações (4.1) e (4.2)) foram determinados com base nos valores obtidos nas classificações de verdadeiros positivos (VP), verdadeiros negativos (VN) e falsos negativos (FN).

$$Recall = \frac{VP}{VP + FN} \times 100 \tag{4.1}$$

$$Acuracia = \frac{VP + VN}{N} \times 100 \tag{4.2}$$

Tabela 20 – Tabela de decisão resultante do reduto “CH_T, CH_E, CH_P, AST, CREATININA” no Teste A.

CH_T	CH_E	CH_P	AST	CREATININA	DIAGNÓSTICO
Baixo	Baixo	Baixo	Normal	Normal	IA–AP–CHPD
Baixo	Baixo	Baixo	Normal	Baixo	IA–AP–CHPD–CHED–PRCR
Normal	Normal	Baixo	Normal	Normal	IA–BP–CHPD
NAI	Normal	Baixo	NAI	NAI	IA–CHPD
Baixo	Normal	Baixo	Normal	Normal	IA–CHPD
Normal	Normal	Baixo	NAI	NAI	IA–CHPD
Normal	Normal	Baixo	Baixo	Normal	IA–CHPD–ASTALT–LH
Baixo	Baixo	Normal	Normal	Normal	IC–CHED
Normal	Baixo	Normal	Normal	Normal	IC–CHED
Baixo	Normal	Normal	Normal	Normal	IC–CHED
Normal	Baixo	Normal	NAI	NAI	IC–CHED
Baixo	Baixo	Normal	Normal	Baixo	IC–CHED
Normal	Baixo	Normal	Normal	Baixo	IC–CHED
Baixo	Baixo	Normal	Baixo	Normal	IC–CHED–AST–LH
Baixo	Normal	Normal	Baixo	Normal	IC–CHED–AST–LH
Normal	Baixo	Normal	Baixo	Normal	IC–CHED–AST–LH
Normal	Normal	Normal	Normal	Normal	NAI
Normal	Normal	Normal	Baixo	Normal	NAI
Normal	Normal	Normal	Baixo	Baixo	NAI
Normal	Normal	Normal	Normal	NAI	NAI
Normal	Normal	Normal	Normal	Baixo	NAI
Normal	Normal	Normal	NAI	NAI	NAI
Baixo	Normal	Normal	NAI	NAI	NTD
NAI	NAI	NAI	Baixo	Normal	NTD
NAI	NAI	NAI	NAI	NAI	NTD
NAI	NAI	NAI	Normal	Normal	NTD
Baixo	Baixo	Normal	NAI	NAI	IC–CHED

Na análise realizada, as amostras corretamente diagnosticadas foram consideradas verdadeiros positivos, enquanto as amostras não classificadas foram consideradas falsos negativos devido à ausência de regras para sua classificação. No caso do diagnóstico em questão, nenhuma regra com mais de um diagnóstico possível ocorreu, portanto, nenhuma amostra foi considerada falso positivo ou verdadeiro negativo. Devido à ausência de amostras falsos positivos, os cálculos de Precisão resultaram em valores indefinidos com erros de divisão por zero, impedindo outros cálculos como o F1-Score.

A Tabela 21 fornece uma visão detalhada da acurácia para o reduto “CH_T, CH_E, CH_P, AST, CREATININA” em cada teste, considerando todos os tipos de diagnósticos.

Tabela 21 – Acurácia do reduto “CH_T, CH_E, CH_P, AST, CREATININA” em cada teste, considerando todos os tipos de diagnósticos.

Testes	N	VP	VN	FP	FN	Acurácia(%)
A	256	255	0	0	1	99,61
B	257	256	0	0	1	99,61
C	257	253	0	0	4	98,44
D	257	251	0	0	6	97,67

No reduto escolhido, foram realizados cálculos para determinar o recall e a acurácia de cada

Tabela 22 – Recall de cada diagnóstico referente às regras geradas por “CH_T, CH_E, CH_P, AST, CREATININA”.

Diagnóstico	Recall dos testes (%)			
	A	B	C	D
IA–AP–CHPD	100,00	100,00	100,00	100,00
IA–AP–CHPD–CHED–PRCR	100,00	100,00	100,00	100,00
IA–BP–CHPD	100,00	100,00	100,00	100,00
IA–CHPD	100,00	0,00	100,00	0,00
IA–CHPD–ASTALT–LH	100,00	100,00	100,00	100,00
IC–CHED	98,04	100,00	94,12	100,00
IC–CHED–AST–LH	100,00	100,00	100,00	100,00
NAI	100,00	100,00	100,00	98,43
NTD	100,00	100,00	66,67	33,33

Tabela 23 – Acurácia de cada diagnóstico referente às regras geradas por “CH_T, CH_E, CH_P, AST, CREATININA”.

Diagnóstico	Acurácia dos testes (%)			
	A	B	C	D
IA–AP–CHPD	100,00	100,00	100,00	100,00
IA–AP–CHPD–CHED–PRCR	100,00	100,00	100,00	100,00
IA–BP–CHPD	100,00	100,00	100,00	100,00
IA–CHPD	100,00	99,61	100,00	99,61
IA–CHPD–ASTALT–LH	100,00	100,00	100,00	100,00
IC–CHED	99,61	100,00	98,83	100,00
IC–CHED–AST–LH	100,00	100,00	100,00	100,00
NAI	100,00	100,00	100,00	98,83
NTD	100,00	100,00	99,61	99,22

diagnóstico, e os resultados são apresentados nas Tabelas 22 e 23. Nesse caso, os diagnósticos que não estavam sendo verificados foram considerados como verdadeiros negativos.

No conjunto de 1027 amostras, houve sobreamostragem para alguns diagnósticos e subamostragem para outros. Durante a validação cruzada (CV), tentou-se alcançar uma distribuição uniforme dessas amostras entre diferentes diagnósticos. No entanto, nos Testes B e D, foi observado que os diagnósticos em “IA–CHPD” registraram um recall de 0% devido à incapacidade de gerar regras eficazes para a classificação deste diagnóstico no conjunto de testes. Considerando que novas coleções de amostras estão em andamento como parte do Projeto PCS e que os resultados iniciais já estão beneficiando os profissionais de saúde, não foi considerado necessário recorrer a metodologias de simulação para balanceamento de classes. Ajustes nas amostras por meio do processo de recuperação serão necessários no futuro para abordar esse desequilíbrio.

Conforme estabelecido pelo modelo CRD–TCA, na etapa de recuperação, novas amostras são coletadas com base em todas as informações obtidas durante a etapa de refinamento. A Tabela 24 apresenta uma parte dessas informações, especificamente as regras de refinamento dos dados presentes nos redutos listados na Tabela 19.

Após a conclusão da etapa de recuperação, será possível realizar novas sessões de treinamento com o objetivo de incorporar um número crescente de casos diversos de intoxicação por agrotóxicos, melhorando assim a capacidade do modelo de lidar com vários diagnósticos com

maior precisão.

Tabela 24 – CRD–TCA: Regras de Refinamento para Recuperação.

Nome original	Regras de Refinamento
Último contato (em dias) com um pesticida	Desinformado, [0, 7 dias] - Exposição Aguda [8, 30 dias] - Exposição Subaguda, [31, 90 dias] - Exposição Subcrônica, [91, *] - Exposição Crônica
Método de aplicação do produto	Bomba dorsal, Mangueira, Trator sem cabine, Trator com cabine fechada, Desligado
Tipo de contato	Direto, Indireto, Sem contato
Alteração do sistema nervoso, Dor de cabeça, Incoordenação motora, Vômito, Aborto espontâneo	Sim / Não
CH_T	[Val < 15.5] Baixo,[Val ≥ 15.5] Normal
CH_E	[Val < 32] Baixo, [Val ≥ 32] Normal
CH_P	[Val < 1.3] Baixo, [Val ≥ 1.3] Normal
AST	[Val < 4] Baixo, [Val ≥ 4 e Val ≤ 36] Normal
CREATININA	Homens: [Val < 0.9] Baixo, [Val ≥ 0.9 e Val ≤ 1.3] Normal Mulheres: [Val < 0.6] Baixo, [Val ≥ 0.6 e Val ≤ 1.1] Normal
Circunferência abdominal	Homens: [*, 101] Adequada, [102, *] Inadequada Mulheres: [*, 87] Adequada, [88, *] Inadequada

4.3 Considerações finais

Neste capítulo, os resultados obtidos da aplicação do modelo CRD–TCA no projeto PCS, proporcionou uma avaliação dos melhores conjuntos de condição para a tomada de decisão no diagnóstico. As análises realizadas destacam a relevância dos dados estudados e suas inter–relações.

O próximo capítulo abordará as conclusões finais e as perspectivas futuras para o modelo CRD–TCA. Destacará suas contribuições para a área de saúde, ressaltando sua eficácia na assistência aos profissionais de saúde e na seleção eficiente de atributos. Além disso, discutirá os desafios remanescentes que precisam ser enfrentados para sua implementação e aprimoramento em contextos de pesquisa.

5 Conclusão

O modelo CRD–TCA demonstrou sua eficácia em abordar todas as etapas do processo KDD de acordo com a hierarquia da Ciência de Dados, conforme ilustrado na Figura 4.1. Sua simplicidade e praticidade o tornam adequado para profissionais de diversos campos envolvidos em projetos de Ciência de Dados. Além disso, ele lidou com sucesso com dados ruidosos e removeu informações desnecessárias, resultando em redução da incerteza no conjunto de dados. Isso levou a um conjunto de dados preparado para treinamento e, consequentemente, à derivação de regras de decisão bem validadas. Esses resultados contribuíram significativamente para o projeto PCS, melhorando sua eficiência e utilidade, pois foram obtidas 27 regras de decisão com 99,61% de precisão diagnóstica. Essas regras servem de apoio para a tomada de decisão dos profissionais de saúde e contribuem para a saúde dos trabalhadores agrícolas, o que é crucial para garantir a produtividade agrícola e a qualidade do produto.

Identificamos modelos de Ciência de Dados com descrições puramente teóricas e, separadamente, encontramos aplicações de aprendizado de máquina assumindo que os dados já estão preparados para uso, o que dificultou a apresentação de uma análise comparativa com o modelo proposto.

Em trabalhos futuros, é recomendado:

- Explorar a conexão do modelo CRD–TCA com bancos de dados orientados a objetos e bancos de dados orientados a grafos, permitindo maior flexibilidade no armazenamento e recuperação de dados.
- Expandir o escopo incorporando a documentação necessária à modelagem CRD–TCA para utilizar métodos adicionais de aprendizado de máquina. Essa abordagem permitirá que cientistas de dados avaliem e escolham o método mais adequado com base nas características específicas de seus problemas, permitindo análises comparativas no mesmo conjunto de dados.
- Para melhorar a acessibilidade e utilidade prática do modelo, uma possível extensão seria criar uma API para implementar um aplicativo móvel de coleta de dados.

Esses esforços têm o potencial de ampliar o impacto do modelo e ampliar sua aplicabilidade em projetos de Ciência de Dados.

Publicação relacionada

Uma parte significativa dos resultados desta pesquisa foi publicada no artigo:

J. C. S. Carvalho, T. C. Pimenta, A. C. P. Silverio, M. A. Carvalho and J. P. C. S. Carvalho, "A New Data Science Model With Supervised Learning and its Application on Pesticide Poisoning Diagnosis in Rural Workers,"in IEEE Access, vol. 12, pp. 40871-40882, 2024, doi: 10.1109/ACCESS.2024.3375764.

Referências

[1] Pecuária e Abastecimento Ministério da Agricultura. Ministério da agricultura, pecuária e abastecimento, 2024. URL https://www.gov.br/agricultura/pt-br/assuntos/insumos-agropecuarios/insumos-agricolas/agrotoxicos. Online.

[2] Daniel Azevedo Duarte. Quem são os maiores exportadores agrícolas do mundo? https://news.agrofy.com.br/noticia/202377/quem-sao-os-maiores-exportadores-agricolas-do-mundo. [Online; Acesso em 09 de Junho de 2024].

[3] J.-N. Aubertot, F. Robin, J. Lamichhane, K. Wick, and R. A. Deguine. Integrated pest management: good intentions, hard realities. a review. *Agronomy for Sustainable Development*, 41, June 2021. doi: 10.1007/s13593-021-00642-x.

[4] A. C. P. Silvério. *Aplicação de bioindicadores e avaliação clínica em trabalhadores rurais expostos aos praguicidas organofosforados visando subsidiar a implantação de uma rede de atenção primária à saúde*. Dissertação de doutorado, Universidade Federal de Alfenas - UNIFAL, 2016.

[5] E. R. M. Cabral. Exposição aos agrotóxicos: Implicações na saúde de trabalhadores agrícolas de uma região de campinas, sp. Master's thesis, Universidade Estadual de Campinas - Faculdade de Ciências Médicas - UNICAMP, Campinas, 2012.

[6] Sistema Nacional de Informações Tóxico-Farmacológicas - SINITOX. Casos registrados de intoxicação humana e envenenamento. http://www.fiocruz.br/sinitox/. [Online; Acesso em 05 Maio 2023].

[7] OIT - Organização Internacional do Trabalho. Agriculture: a hazardous work. https://www.ilo.org/safework/info/WCMS_110188/lang--es/index.htm. [Online; Acesso em 05 Maio 2023].

[8] Antonio Peña-Fernández, María Peña, M.C. Lobo, and Mark Evans. Interventions to enhance the teaching of toxicology at a uk university. pages 7126–7130, 07 2018. doi: 10.21125/edulearn.2018.1683.

[9] Iqbal H. Sarker. Data science and analytics: An overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science*, 2(5):377, 2021. ISSN 2661-8907. doi: 10.1007/s42979-021-00765-8.

[10] J. D. Kelleher and B. Tierney. *Data Science*. The MIT Press, Cambridge, MA, 2018.

[11] Jeffrey Saltz and Iva Krasteva. Current approaches for executing big data science projects—a systematic literature review. *PeerJ Computer Science*, 8:e862, 02 2022. doi: 10.7717/peerj-cs.862.

[12] Tijl De Bie, Luc De Raedt, José Hernández-Orallo, Holger H. Hoos, Padhraic Smyth, and Christopher K I Williams. Automating data science: Prospects and challenges. *Communications of the ACM*, 65(3):76–87, feb 2022. ISSN 0001-0782. doi: 10.1145/3495256.

[13] Z. Pawlak. *Rough sets: theoretical aspects of reasoning about data*. Kluwer, London, 1991.

[14] D.P. Acharjya and Ajith Abraham. Rough computing — a review of abstraction, hybridization and extent of applications. *Engineering Applications of Artificial Intelligence*, 96:103924, 2020. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2020.103924. URL https://www.sciencedirect.com/science/article/pii/S0952197620302529.

[15] T. H. A. Silva, D. G. M. Nobre, and T. P. R. Bachur. Caminhos para a consolidação da disciplina de toxicologia médica em um curso de medicina. *REBES - Revista Brasileira de Educação e Saúde*, 9(4):145–149, 2019.

[16] Diretores da IBM e da Gap. Por que 87% dos projetos de data science não saem do papel? ilumeo.com.br, 10 07 2019. [Online], 2019. URL https://ilumeo.com.br/categorias/2019-07-29-por-que-87-dos-projetos-de-data-science-nao-saem-do-papel/. [Acesso em 24 de maio de 2023].

[17] Inigo Martinez, Elisabeth Viles, and Igor G Olaizola. A survey study of success factors in data science projects. In *IEEE International Conference on Big Data (Big Data)*, pages 2313–2318, 2021. doi: 10.1109/bigdata52589.2021.9671588.

[18] Jeffrey Saltz, Nicholas Hotz, David Wild, and Kyle Stirling. Exploring project management methodologies used within data science teams. In *Americas Conference on Information Systems 2018*, Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018. Association for Information Systems, 2018. ISBN 9780996683166. Publisher Copyright: © 2018 Association for Information Systems. All rights reserved.; 24th Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018 ; Conference date: 16-08-2018 Through 18-08-2018.

[19] D. Araújo, M. A. A. Llarema, S. d. A. Siebra, and G. A. Dias. Contribuições para a gestão de dados científicos: análise entre modelos de ciclo de vida dos dados. *LIINC*, 2:15, 2019.

[20] R. Bochner. Óbito ocupacional por exposição a agrotóxicos utilizado como evento sentinela: quando o pouco significa muito. *Visa em Debate - Sociedade, ciência e tecnologia*, 3(4):39–49, 2015.

[21] R. Tooge. Quem criou o termo 'agrotóxico' e por que não 'pesticida' ou 'defensivo agrícola'. Globo.com, 07 10 2019. [Online]. URL https://g1.globo.com/economia/agronegocios/noticia/2019/10/07/quem-criou-o-termo-agrotoxico-e-por-que-nao-pesticida-ou-defensivo-agricola.ghtml. [Acesso em 10 de maio de 2023].

[22] C. R. A. Mendes, C. E. P. Mendes, S. E. Santos, K. S. R. Luz, and L. P. Santana. AgrotÓxicos: principais classificações utilizadas na agricultura brasileira - uma revisão de literatura. *Revista Maestria*, 17:95–107, 2019.

[23] P. P. C. Mineau. The impact of the nation's most widely used insecticides on birds. *Am.Bird Conserv.*, March 2013.

[24] L. M. Bombardi. *Geografia do Uso de Agrotóxicos no Brasil e Conexões com a União Europeia*. FFLCH - USP, São Paulo, 2017.

[25] C. R. M. Araújo, V. L. d. A. Santos, and G. A. A. Acetilcolinesterase - ache: Uma enzima de interesse. *Revista Virtual de Química*, 8(6):1818–1834, 2016.

[26] COLUNISTA PORTAL - EDUCAÇÃO. Portal educação. Portal Educação, 2020. [Online]. URL https://siteantigo.portaleducacao.com.br/conteudo/artigos/farmacia/historico-dos-anticolinesterasicos-agonistas-colenergicos-de-acao-indireta/46773. [Acesso em 25 de maio de 2023].

[27] J. P. Q. W. Z. S. P. E. K. Gang Shao. Exploring potential roles of academic libraries in undergraduate data science education curriculum development. *The Journal of Academic Librarianship*, 47(2), 2021.

[28] Brasil. Lei n. 13.709, de 14 de agosto de 2018. lei geral de proteção de dados pessoais (lgpd). Brasília, 2018. [Online]. URL http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. [Acesso em 02 de junho de 2023].

[29] H. S. M. a. S. M. Silva. Data science in public mental health: a new analytic framework. In *IEEE Symposium on Computers and Communications (ISCC)*, pages 1123–1128, 01 06 2019.

[30] P. D. McNicholas. Data science. *FACETS*, 4:131–135, 08 2019. doi: 10.1139/facets:2019-0005.

[31] H. Harmouch. Evaluating four of the most popular open source and free data mining tools. *IJASR International Journal of Academic Scientific Research*, 3:13–23, 2015.

[32] Jeff Russell. *Agile Data Science 2.0: Building Full-Stack Data Analytics Applications with Spark*. O'Reilly Media, Sebastopol, CA, USA, 2017. ISBN 978-1491960110.

[33] R. Wirth and J. Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, vol. 1*. Springer-Verlag London, UK, 2000.

[34] Farhad Foroughi and Peter Luksch. Data science methodology for cybersecurity projects. *CoRR*, abs/1803.04219, 2018. URL http://arxiv.org/abs/1803.04219.

[35] R. C. G. Santana. Ciclo de vida dos dados: Uma perspectiva a partir da ciência da informação. *Inf.Inf., Londrina*, 21(2):116–142, 2016.

[36] A. Beal. Agile data science: How is it different? [Online], 2022. URL https://www.modernanalyst.com/Resources/Articles/tabid/115/ID/6081/categoryId/20/Agile-Data-Science-How-is-it-Different.aspx. [Acesso em 07 de maio de 2023].

[37] Iñigo Martinez, Elisabeth Viles, and Igor G. Olaizola. Data science methodologies: Current challenges and future approaches. *Big Data Research*, 24:100183, 2021. ISSN 2214-5796. doi: https://doi.org/10.1016/j.bdr.2020.100183. URL https://www.sciencedirect.com/science/article/pii/S2214579620300514.

[38] P. C. T. Gomes. Machine learning para todos, de forma simples e com exemplos! DataGeeks, 26 06 2019. URL https://www.datageeks.com.br/machine-learning/. [Acesso em 15 de junho de 2023].

[39] H. Honda, M. Facure, and P. Yaohao. The three types of machine learning. Lamfo UnB, 27 07 2017. URL https://lamfo.unb.br/blog-en/. [Acesso em 15 de junho de 2023].

[40] T. Ludemir. Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados*, pages 85–94, 01 04 2021. doi: doi.org/10.1590/s0103-4014.2021.35101.007.

[41] P. A. Harlianto, N. A. Setiawan, and T. B. Adji. Comparison of machine learning algorithms for soil. In *2017 3rd International Conference on Science and Technology - Computer (ICST)*, pages 7–10, 2017. doi: 10.1109/ICSTC.2017.8011843.

[42] Z. Pawlak. Rough real functions and rough controllers. Research Report 1/95, The Institute of Computer Science (ICS) ¬ Warsaw University of Technology (WUT), Warsaw,ICS¬WUT, Janeiro 1995.

[43] B. R. Slowinski. Rough set approach to decision analysis. *AI Expert*, pages 19–25, March 1995.

[44] W. P. Ziarko. *Rough Sets, fuzzy sets and knowledge discovery*. Springer-Verlag, London, 1994.

[45] L. F. Medeiros. *Redes Neurais em Delphi*. Visual Books, Florianópolis, 2006.

[46] M. Grubler. Entendendo o funcionamento de uma rede neural artificial. aibrasil, 11 Junho 2018. URL https://medium.com/brasil-ai/entendendo-o-funcionamento-de-uma-rede-neural-artificial-4463fcf44dd0. [Acesso em 29 Junho 2023].

[47] Z. Pawlak. Rough sets. *International Journal of Computer and Information Sciences*, 11 (5):341–356, 1982.

[48] Z. Pawlak. Rough classification. *International Journal of Man-Machine Studies*, (20): 469–483, 1984.

[49] Z. Pawlak. Rough sets and fuzzy sets. Technical Report 540, ICS, Warsaw, March 1984.

[50] Z. Pawlak. On learning - a rough set approach. In *Lecture Notes in Computer Science*, volume 28, pages 197–227. 1985.

[51] Z. Pawlak. Rough sets and fuzzy sets. *Fuzzy Sets and Systems*, (17):99–102, 1985.

[52] Z. Z. C. S. B. M. M. G. L. R. B. d. F. O. & B. Chelly Dagdia. Rough set theory as a data mining technique: A case study in epidemiology and cancer incidence prediction. *Machine Learning and Knowledge Discovery in Databases*, 11053:440–455, 2019.

[53] M. S. &. Z. J. &. J. D. &. N. A. &. D. S. Pathan. Identifying stroke indicators using rough sets. *IEEE Access*, 8, 2020. doi: 10.1109/ACCESS.2020.3039439.

[54] B. R. Slowinski. *Intelligent Decision Support*. Kluwer Academic Publisher, 1992.

[55] Z. Pawlak. Information systems: theoretical foundations. *Information Systems*, 6(3): 205–218, 1981.

[56] Z. Pawlak. Rough relations. ftp at ftp.ii.pw.edu.pl/pub/reports/, ICS PAS Report 435, Warsaw, June 1981.

[57] Z. Pawlak. Rough functions. ftp at ftp.ii.pw.edu.pl/pub/reports/, ICS PAS Report 467, Warsaw, ICS, December 1981.

[58] J. Q. Uchôa. *Representação e indução de conhecimento usando teoria de conjuntos aproximados*. PhD thesis, UFSCar, 1998.

[59] A. &. Z. W. Szladow. Rough sets: working with imperfect data. *AI Expert*, (July):36–41, 1993.

[60] W. P. Ziarko. The discovery, analysis, and representation of data dependencies in databases. In *Knowledge Discovery in Databases*, pages 195–209. MIT, Boston, 1991.

[61] A. S. Z. Pawlak. Rudiments of rough sets. *Information Sciences*, (177):3–27, 2007.

[62] A. S. Z. Pawlak. Rough sets and boolean reasoning. *Information Sciences*, (177):41–73, 2007.

[63] C. R. A. Skowron. *The discernibility matrices and functions in information systems*. Kluwer, Dordrecht, 1992.

[64] Y. Z. Yiyu Yao. Discernibility matrix simplification for constructing attribute reducts. *Information Sciences*, 179(7):867–882, 2009.

[65] N. S. Hoa. Some efficient algorithms for rough set methods. In *Proceedings IPMU'96 Granada (Spain)*, pages 1541–1457, 1996.

[66] A. C. GALANTE, E. L. R. MOREIRA, and F. C. BRANDÃO. Banco de dados orientados a objetos: Uma realidade. fsma.edu. URL http://www.fsma.edu.br/si/edicao3/banco_de_dados_orientado_a_objetos.pdf. [Acesso em 05 de junho de 2023].

[67] W. &. W. C. &. L. J. &. Y. B. &. L. Y. &. W. J. Chen. Benchmarking discretisation level of continuous attributes: Theoretical and experimental approaches. In *2019 IEEE International Conference on Big Data*, pages 3623–3631, 2019. doi: 10.1109/BigData47090.2019.9006513.

[68] G. &. R. D. &. P.-R. O. &. S. P. Dimić. Descriptive statistical analysis in the process of educational data mining. pages 388–391, 2019. doi: 10.1109/TELSIKS46999.2019.9002177.

[69] Daniel Berrar. Cross-validation. In Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach, editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 542–545. Academic Press, Oxford, 2019. ISBN 978-0-12-811432-2. doi: https://doi.org/10.1016/B978-0-12-809633-8.20349-X. URL https://www.sciencedirect.com/science/article/pii/B978012809633820349X.

[70] C. Bergmeir. cran.r-project.org. [Online], 12 2019. URL https://cran.r-project.org/web/packages/RoughSets. [Acesso em 07 de março de 2023].

[71] L. S. Riza, A. Janusz, C. Bergmeir, C. Cornelis, F. Herrera, D. Slezak, and J. M. Benitez. Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "RoughSets". *Information Sciences*, 278:68–89, 2014.

A Anexo: Ficha de Investigação de Exposição aos Praguicidas

FICHA DE INVESTIGAÇÃO DE EXPOSIÇÃO AOS PRAGUICIDAS

(Quando a questão não se aplicar anotar o número 99)

I. DADOS DE IDENTIFICAÇÃO

1. Data:		**2. Nº:**	
3. Nome do município de residência:	**3.1 ( )**	(1)Zona Rural	(2)Zona urbana
4. Endereço: (Rua, Av. etc.):			**Nº:**
5. Bairro:	**6. Ponto de Referência:**	**7. Telefone:**	
8. Nome do município onde trabalha:	**9. Local de Trabalho:**		

II. DADOS DO PACIENTE

10. Nome do Paciente:			**11. Sexo: ( )**	(1)M	(2)F
12.Gestante:( )				(1)Sim	(2)Não
13. Data de Nascimento:	**14. Idade:**		**15. Anos de estudo*:**		
16.Tabagismo: ( )	**16.1** ( )Atual	**16.2**( )Anterior		(1)Sim	(2)Não
17.Etilismo: ( )	**17.1** ( )Atual	**17.2**( )Anterior		(1)Sim	(2)Não
18. Ingestão de café: ( )	**18.1** Quantidade:	ml/dia		(1)Sim	(2)Não

*A correspondência é feita de tal modo que cada série concluída com aprovação corresponde a 1 ano de estudo.

III. DADOS OCUPACIONAIS

19. Relação de Trabalho: ( )	(1)Proprietário	(2)Assalariado	(3)Meeiro/Arrendatário	(4)Volante
	(5)Outro:			
20. Função: ( )	(1)Administrativa	(2)Téc.Agrícola/Agrônomo	(3) Aplicador na Pecuária	
	(4)Puxa Mangueira	(5)Aplicador/Preparador de Calda	(6)Outros: (Agricultura Familiar)	
21. Contato com Praguicidas: ( )			(1)Sim	(2)Não
22. Há quanto tempo tem contato com praguicidas(venenos)?anos				
23. Frequência do contato com praguicidas: (Anotar valor da multiplicação ano/meses/dias): ( )				
Quantos meses por ano?	Quantos dias por mês?	Quantas horas por dia?		
24. Quando foi a última vez que teve contato (em dias) com um praguicida?				
25. Com qual produto teve contato pela última vez?				
26. Como aplica os produtos? ( )	(1)Bomba costal(mochila)	(2)Mangueira	(3)Trator sem cabide	
(4)Trator com cabine fechada	(5)Outros (especificar):			
27. Praguicidas de maior utilização (até três)	**27.1**Nome comercial:			
	27.2 Princípio Ativo ou Classe Toxicológica:			
	27.3 Cultura/Lavoura:			

28.Principal Via de Exposição:( )	(1)Cutânea	(2)Digestiva	(3)Respiratória	(4)Outra:
29. Já ficou doente por causa do veneno?( )			(1)Sim	(2)Não
30. Quantas vezes você ficou doente por causa do veneno?( )		(1)Uma única vez	(2)Mais de uma vez	
31. Alguma vez teve que ser internado?( )			(1)Sim	(2)Não
32. Quantas vezes?( )		(1)Uma única vez	(2)Mais de uma vez	
33. Há quanto tempo isto aconteceu?( )	(1)Há menos de 10 anos	(2)Há mais de 10 anos		
34.Tipo de Contato:( )	(1)Direto	(2)Indireto	(3)Sem contato	
35. Utiliza Equipamentos de Proteção Individual:( )	**35.1** ( )Roupa impermeável apropriada			
35.2 ( ) Bota apropriada	**35.3** ( ) Luvas	**35.4** ( ) Máscaras		
35.5 ( ) Óculos de proteção	**35.6** ( ) Protetor auricular	(1) Completo	(2) Incompleto	(3) Não

IV. DADOS CLÍNICOS

36.Apresenta Doença Cardiovascular:()	**36.1**( )Hipertensão arterial (pressão alta)		
36.2( )Hipotensão arterial (Pressão baixa)	**36.3**( )Arritmia (batedeira)	(1)Sim	(2)Não
37. Apresenta algum sinal/sintoma referente ao Sistema Nervoso Central Periférico?()			
37.1( ) Dor de cabeça	**37.2**( ) Fraqueza muscular	**37.3** ( )Tremedeira	
37.4( ) Tremormuscular?....Palpebral?	**37.5**( )Visão Turva/Vista embaçada		
37.6 ( )Agitação/Irritabilidade	**37.7** ( )Vertigens/Tonturas	**37.8**( )Formigamento em MMSS	
37.9( )Incoordenação Motora (Se não compreende pergunta, faça o teste Índex-nariz.)		(1)Sim	(2)Não
38. Do Aparelho Digestório? ()	**38.1**( ) Cólicas/Dor de barriga	**38.2**( ) Dor de estômago	
38.3( )Azia/Queimação	**38.4**( )Náuseas/Enjoo	**38.5**( )Vômito	**38.6**( ) Diarreia
39. Do Aparelho Respiratório? ( )	**39.1**( ) Falta de ar	**39.2**( )Irritação Nasal (coceira/ardência)	
39.3( ) Catarro ou escarro	**39.4**( )Tosse	(1)Sim	(2)Não
40. Do Aparelho Auditivo: ( )	**40.1**( ) Diminuição da audição	**40.2**( )Zumbido	(1)Sim (2)Não
41. De Pele e Mucosa? O Sr (a) tem alguma coceira relacionada ao uso do agrotóxico? ( )			
41.1( )A coceira veio depois de algum tempo que o sr(a) começou a trabalhar com o produto?(DC Sensibilizante)			
41.2 ( ) Ou ela aparece logo que usa/prepara o produto?(DC Irritativa)			
41.3 ( ) O Sr (a) tem irritação ocular (coceira, vermelhidão...), por causa do produto?		(1)Sim	(2)Não
42. Do Aparelho Urinário:()	**42.1** ( )Diminuição da urina (pouco)	**42.2**( )Urina escura/com sangue	
42.3 ( ) Outro:		(1)Sim	(2)Não
43.Exposição Raio X ( )	**43.1** Data da última exposição:	(1)Sim	(2)Não

V. NEOPLASIA

44. Tem/Teve Câncer? ( )	**44.1 Qual Tipo?**		
45. Alguém da Família tem/teve Câncer? ( )	**45.1 Qual Tipo?**		
45.2 É da Região? ( )		(1)Sim	(2)Não

VI. DADOS LABORATORIAIS

Resultado do Exame de Colinesterase (Método de Ellman)		**VR**	**IBMP**
46.1Ch-T		15,5 – 31	25% de inibição da atividade da Ch-T
46.2Ch- E		32 – 58	30% de inibição da atividade Ch-E
46.3Ch-P		1,3 – 7,8	50% de inibição da atividade Ch-P

47. AST:	**48.ALT:**
49. γ- GT:	**50. CREATININA:**

VII. CONDUTA

51. Encaminhado ao Ambulatório de Doença Ocupacional? ( )	(1)Sim	(2)Não

VIII. AVALIAÇÃO NUTRICIONAL

52. Nos últimos 7 dias, em quantos dias você comeu os seguintes alimentos ou bebidas?								
ALIMENTO/BEBIDA	**Não comi**	**1 dia**	**2 dias**	**3 dias**	**4 dias**	**5 dias**	**6 dias**	**Todos os 7**
1. Salada crua (alface, tomate, cenoura, pepino, repolho, etc.).								
2. Legumes e verdura cozidos (couve, abóbora, chuchu, brócolis, espinafre, etc.) (não considerar batata e mandioca								
3. Frutas frescas ou salada de frutas								
4. Feijão								
5. Leite ou iogurte								
6. Batata frita, batata de pacote e salgados fritos (coxinha, quibe, pastel etc.)								
7. Hambúrguer e embutidos (salsicha, mortadela, salame, presunto, linguiça etc.)								
8. Bolachas/biscoitos salgados ou salgadinhos de pacote								
9. Bolachas/biscoitos doces ou recheados, doces, balas e chocolates (em barra ou bombom)								
10. Refrigerante (não considerar *diet* ou *light*)								

IX. DADOS ANTROPOMÉTRICOS

Peso	
Altura	
53. ÍMC	
54. CA	

X. INFORMAÇÕES ADICIONAIS

55. Você (ou sua esposa) teve dificuldade engravidar	
56. Você (ou sua esposa) teve abortamento espontâneo	
57. Tem filhos	
58. algum filho com mal formação (congênita)	
59. Você toma algum medicamento de uso contínuo	
59.1. Nome do medicamento	
60. Usou remédio para micose nos últimos dois meses	
60.1. Nome do remédio	

B Anexo: Formulário Completo de Armazenamento no projeto PCS

CRD-TCA: Formulário para Armazenamento		
Banco de dados: Pesquisa Tabela: Amostra_Original		
Nome original do dado	**Nome do campo na tabela**	**Formato**
Identificador	Id	int
2. Número do formulário	Sujeito	varchar(50)
3. Município de residência	Residencia	varchar(50)
3.1 Zona de residência	Zona_Res	varchar(50)
8. Nome do município onde trabalha	Municipio_Trabalha	varchar(50)
11. Sexo	Sexo	varchar(50)
12. Gestante	Gestante	varchar(50)
14. Idade	Idade	int
15. Anos de estudo	Anos_Estudo	int
16. Tabagismo	Tabagismo	varchar(50)
16.1 Atual	Atual1	varchar(50)
16.2 Anterior	Anterior1	varchar(50)
17. Etilismo	Etilismo	varchar(50)
17.1 Atual	Atual2	varchar(50)
17.2 Anterior	Anterior2	varchar(50)
18. Ingestão de café	Ingestao_Cafe	varchar(10)
18.1 Quantidade em ml/dia	Cafe_mlDia	int
19. Relação de Trabalho	Relacao_Trabalho	varchar(50)
20. Função	Funcao	varchar(50)
21. Contato com Praguicidas	Contato_Praguicida	varchar(50)
22. Tempo tem contato com praguicidas em anos	Tempo_Contato	varchar(50)
23. Frequência do contato com praguicidas - multiplicação ano/mês/dias	Frequencia_Contato	varchar(50)
24. Última vez que teve contato (em dias) com um praguicida	Ultimo_Contato	varchar(50)
25. Produto teve contato pela última vez	Produto	varchar(50)
26. Como aplica os produtos	Forma_Aplicacao	varchar(50)
27.1 Nome comercial do Praguicidas de maior utilização	Nome_Comercial	varchar(50)
27.3 Cultura/Lavoura	Cultura	varchar(50)

Nome original do dado	Nome do campo na tabela	Formato
28. Principal Via de Exposição	Via_Exposicao	varchar(50)
29. Já ficou doente por causa do veneno	Adoeceu	varchar(50)
30. Quantas vezes você ficou doente por causa do veneno	N_Adoeceu	varchar(50)
31. Alguma vez teve que ser internado	Internado	varchar(50)
32. Quantas vezes	N_Internado	varchar(50)
33. Há quanto tempo isto aconteceu	Tempo	varchar(50)
34. Tipo de Contato	Tipo_Contato	varchar(50)
35. Utiliza Equipamentos de Proteção Individual	Equip_Protecao	varchar(50)
35.1 Roupa impermeável apropriada	Roupa_Aproriada	varchar(50)
35.2 Bota apropriada	Bota_Apropriada	varchar(50)
35.3 Luvas	Luvas	varchar(50)
35.4 Máscaras	Mascaras	varchar(50)
35.5 Óculos de proteção	Oculos	varchar(50)
35.6 Protetor auricular	Protetor_Auricular	varchar(50)
36. Apresenta Doença Cardiovascular	Doenca_Cardiovascular	varchar(50)
36.1 Hipertensão arterial	Hipertensao_Arterial	varchar(50)
36.2 Hipotensão arterial	Hipotensao_Arterial	varchar(50)
36.3 Arritmia	Arritmia	varchar(50)
37. Apresenta algum alteração Sistema Nervoso	Alteracao_Snervoso	varchar(50)
37.1 Dor de cabeça	Dor_Cabeca	varchar(50)
37.2 Fraqueza muscular	Fraqueza_Muscular	varchar(50)
37.3 Tremedeira	Tremedeira	varchar(50)
37.4 Tremor muscular	Tremor_Muscular	varchar(50)
37.5 Visão Turva/Vista embaçada	VisaoTurva_Embacada	varchar(50)
37.6 Agitação/Irritabilidade	Agitacao_Irritabilidade	varchar(50)
37.7 Vertigens/Tonturas	Vertigens_Tonturas	varchar(50)
37.8 Formigamento em MMSS	Formigamento	varchar(50)
37.9 Incoordenação Motora	Incoordenacao_Motora	varchar(50)
38. Do Aparelho Digestório	Aparelho_Digestorio	varchar(50)
38.1 Cólicas/Dor de barriga	Colicas_DorBarriga	varchar(50)
38.2 Dor de estômago	Dor_Estomago	varchar(50)
38.3 Azia/Queimação	Azia_Queimacao	varchar(50)
38.4 Náuseas/Enjoo	Nauseas_Enjoo	varchar(50)
38.5 Vômito	Vomito	varchar(50)
38.6 Diarreia	Diarreia	varchar(50)
39. Do Aparelho Respiratório	Aparelho_Respiratorio	varchar(50)
39.1 Falta de ar	Falta_DeAr	varchar(50)
39.2 Irritação Nasal (coceira/ardência)	Irritacao_Nasal	varchar(50)
39.3 Catarro ou escarro	Catarro_Escarro	varchar(50)
39.4 Tosse	Tosse	varchar(50)

Nome original do dado	Nome do campo na tabela	Formato
40. Do Aparelho Auditivo	Aparelho_Auditivo	varchar(50)
40.1 Diminuição da audição	Diminuicao_Audicao	varchar(50)
40.2 Zumbido	Zumbido	varchar(50)
41. Pele e mucosa	Pele_Mucosa	varchar(50)
41.1 DC Sensibilizante	DC_Sensibilizante	varchar(50)
41.2 DC Irritativa	DC_Irritativa	varchar(50)
41.3 Irritação ocular	Irritacao_Ocular	varchar(50)
42. Do Aparelho Urinário	Aparelho_Urinario	varchar(50)
42.1 Diminuição da urina	Diminuicao_Urina	varchar(50)
42.2 Urina escura/com sangue	UrinaEscura_Sangue	varchar(50)
42.3 Outro	Outro	varchar(50)
43. Exposição Raio X	Exposicao_RaioX	varchar(50)
43.1 Data da última exposição	Data_Exposicao	varchar(50)
44. Tem/Teve Câncer	Teve_Cancer	varchar(50)
44.1 Qual tipo (SNC)	SNC_Cancer	varchar(50)
44.1 Qual tipo (Digestório)	Digestorio_Cancer	varchar(50)
44.1 Qual tipo (Respiratório)	Respiratorio_Cancer	varchar(50)
44.1 Qual tipo (Glandular)	Glandular_Cancer	varchar(50)
44.1 Qual tipo (Pele/Osso/Sangue)	Pele_Osso_Sangue _Cancer	varchar(50)
45. Alguém da Família tem/teve Câncer	Familia_Cancer	varchar(50)
45.1 Qual tipo (SNC)	SNC _CancerFamilia	varchar(50)
45.1 Qual tipo (Digestório)	Digestorio _CancerFamilia	varchar(50)
45.1 Qual tipo (Respiratório)	Respiratorio _CancerFamilia	varchar(50)
45.1 Qual tipo (Glandular)	Glandular _CancerFamilia	varchar(50)
45.1 Qual tipo (Pele/Osso/Sangue)	Pele_Osso_Sangue _CancerFamilia	varchar(50)
45.2 É da região	Eda_Regiao	varchar(50)
46.1 Ch-T1	CH_T1	varchar(50)
46.1 Ch- T2	CH_T	varchar(50)
46.2 Ch- E1	CH_E1	varchar(50)
46.2 Ch- E2	CH_E	varchar(50)
46.3 Ch-P1	CH_P1	varchar(50)
46.3 Ch-P2	CH_P	varchar(50)
47. AST1	AST1	varchar(50)
47. AST2	AST	varchar(50)
48. ALT1	ALT1	varchar(50)
48. ALT2	ALT2	varchar(50)
49. γ- GT1	Y_GT1	varchar(50)
49. γ- GT2	Y_GT	varchar(50)
50. CREATININA1	CREATININA1	varchar(50)
50. CREATININA2	CREATININA	varchar(50)

Nome original do dado	Nome do campo na tabela	Formato
51. Encaminhado ao Ambulatório de Doença Ocupacional	Ambulatorio	varchar(50)
52. Hábito alimentar	Habito_Alimentar	varchar(50)
53. IMC	IMC	varchar(50)
54. Circunferência abdominal	Circunferencia _Abdominal	varchar(50)
55. Você (ou sua esposa) teve dificuldade engravidar	Dificuldade_Engravidar	varchar(50)
56. Você (ou sua esposa) teve abortamento espontâneo	Aborto_Espontaneo	varchar(50)
57. Tem filhos	Tem_Filhos	varchar(50)
58. Algum filho com malformação (congênita)	Filho_MaFormacao	varchar(50)
59. Você toma algum medicamento de uso contínuo	Medicamento_Continuo	varchar(50)
59.1 Nome do medicamento	Medicamento	varchar(50)
60. Usou remédio para micose nos últimos dois meses	Remedio_Micose	varchar(50)
60.1 Nome do remédio	Nome_Remedio	varchar(50)
Diagnóstico	Diagnostico	varchar(50)

C Anexo: Formulário Completo do processo de Refinamento no projeto PCS

CRD-TCA: Formulário para Refinamento			
Nome original do dado	**Tipo**	**Discretização**	**Concluído**
11. Sexo	Discretizado	Masculino / Feminino	Descarte
12. Gestante	Discretizado	Sim / Não	Sim
14. Idade	Contínuo	[0, 12] – Criança [13, 17] – Jovem [18, 59] – Adulto [60, *] – Idoso	Sim
16. Tabagismo	Discretizado	Sim / Não	Sim
16.1 Atual	Discretizado	Sim / Não	Sim
16.2 Anterior	Discretizado	Sim / Não	Sim
17. Etilismo	Discretizado	Sim / Não	Sim
17.1 Atual	Discretizado	Sim / Não	Sim
17.2 Anterior	Discretizado	Sim / Não	Sim
18. Ingestão de café	Discretizado	Sim / Não	Sim
18.1 Quantidade em ml/dia	Contínuo	[0 ml] – Sem consumo [1, 100mL] – Consumo Baixo [101, 300mL] – Consumo Moderado [301, *] – Consumo Alto	Sim
19. Relação de Trabalho	Discretizado	(1) Proprietário (2) Assalariado (3) Meeiro (4) Volante (5) Outro	Descarte
20. Função	Discretizado	Administrativa, Agrônomo, Aplic.Pecuária, Puxa Mangueira, Aplicador /Preparador de Calda, Outros(Agric.Familiar)	Descarte

Nome original do dado	Tipo	Discretização	Concluído
21. Contato com Praguicidas	Discretizado	Sim / Não	Sim
22. Tempo tem contato com praguicidas em anos (Exposição)	Contínuo	Não Informado - NI, [1, 3 anos] – Baixa, [4, 6] – Média, [7, *] – Alta	Sim
23. Frequência do contato com praguicidas - multiplicação ano/mese/dias	Contínuo	Sem Padronização	Descarte
24. Última vez que teve contato (em dias) com um praguicida (Exposição)	Contínuo	Não informado - NI, [0, 7 dia] – Aguda, [8, 30 dias] – Subaguda, [30, 90 dias] – Subcrônica, [91, *] – Crônica	Sim
25. Produto químico teve contato pela última vez	Multivalorado		Não
26. Como aplica os produtos	Discretizado	BombaCostal, Mangueira, Trator sem cabine, Trator Cabine fechada, NI (Não Informado)	Sim
27.1 Nome comercial do Praguicida de maior utilização	Multivalorado		Descarte
27.3 Cultura/Lavoura	Multivalorado		Descarte
28. Principal Via de Exposição	Discretizado	Cutânea, Digestiva, Respiratória	Sim
29. Já ficou doente por causa do veneno	Discretizado	Sim / Não	Sim
30. Quantas vezes você ficou doente por causa do veneno	Discretizado	Uma única vez, Mais de uma vez, Nenhuma vez	Sim
31. Alguma vez teve que ser internado	Discretizado	Sim / Não	Sim
32. Quantas vezes	Discretizado	Uma única vez, Mais de uma vez, Nenhuma vez	Sim
33. Há quanto tempo isto aconteceu	Discretizado	Menos de 10 anos, Mais de 10 anos, Nenhuma vez	Sim
34. Tipo de Contato	Discretizado	Direto, Indireto, Sem contato	Sim

Nome original do dado	Tipo	Discretização	Concluído
35. Utiliza Equipamentos de Proteção Individual	Discretizado	Completo, Incompleto, Não	Sim
35.1 Roupa impermeável apropriada	Discretizado	Sim / Não	Sim
35.2 Bota apropriada	Discretizado	Sim / Não	Sim
35.3 Luvas	Discretizado	Sim / Não	Sim
35.4 Máscaras	Discretizado	Sim / Não	Sim
35.5 Óculos de proteção	Discretizado	Sim / Não	Sim
35.6 Protetor auricular	Discretizado	Sim / Não	Sim
36. Apresenta Doença Cardiovascular	Discretizado	Sim / Não	Sim
36.1 Hipertensão arterial	Discretizado	Sim / Não	Sim
36.2 Hipotensão arterial	Discretizado	Sim / Não	Sim
36.3 Arritmia	Discretizado	Sim / Não	Sim
37. Apresenta alguma alteração Sistema Nervoso	Discretizado	Sim / Não	Sim
37.1 Dor de cabeça	Discretizado	Sim / Não	Sim
37.2 Fraqueza muscular	Discretizado	Sim / Não	Sim
37.3 Tremedeira	Discretizado	Sim / Não	Sim
37.4 Tremor muscular	Discretizado	Sim / Não	Sim
37.5 Visão Turva/Vista embaçada	Discretizado	Sim / Não	Sim
37.6 Agitação/Irritabilidade	Discretizado	Sim / Não	Sim
37.7 Vertigens/Tonturas	Discretizado	Sim / Não	Sim
37.8 Formigamento em MMSS	Discretizado	Sim / Não	Sim
37.9 Incoordenação Motora	Discretizado	Sim / Não	Sim
38. Do Aparelho Digestório	Discretizado	Sim / Não	Sim
38.1 Cólicas/Dor de barriga	Discretizado	Sim / Não	Sim
38.2 Dor de estômago	Discretizado	Sim / Não	Sim
38.3 Azia/Queimação	Discretizado	Sim / Não	Sim
38.4 Náuseas/Enjoo	Discretizado	Sim / Não	Sim
38.5 Vômito	Discretizado	Sim / Não	Sim
38.6 Diarreia	Discretizado	Sim / Não	Sim
39. Do Aparelho Respiratório	Discretizado	Sim / Não	Sim
39.1 Falta de ar	Discretizado	Sim / Não	Sim
39.2 Irritação Nasal (coceira/ardência)	Discretizado	Sim / Não	Sim
39.3 Catarro ou escarro	Discretizado	Sim / Não	Sim

Nome original do dado	Tipo	Discretização	Concluído
39.4 Tosse	Discretizado	Sim / Não	Sim
40. Do Aparelho Auditivo	Discretizado	Sim / Não	Sim
40.1 Diminuição da audição	Discretizado	Sim / Não	Sim
40.2 Zumbido	Discretizado	Sim / Não	Sim
41. Pele e mucosa	Discretizado	Sim / Não	Sim
41.1 DC Sensibilizante	Discretizado	Sim / Não	Sim
41.2 DC Irritativa	Discretizado	Sim / Não	Sim
41.3 Irritação ocular	Discretizado	Sim / Não	Sim
42. Do Aparelho Urinário	Discretizado	Sim / Não	Sim
42.1 Diminuição da urina	Discretizado	Sim / Não	Sim
42.2 Urina escura/com sangue	Discretizado	Sim / Não	Sim
42.3 Outro	Discretizado	Sim / Não	Descarte
43. Exposição Raio X	Discretizado	Sim / Não	Descarte
43.1 Data da última exposição	Contínuo		Descarte
44. Tem/Teve Câncer	Discretizado	Sim / Não	Sim
44.1 Qual tipo	Discretizado	SNC, Digestório, Respiratório, Reprodutor, Glandular, Pele _Osso _Sangue, Nenhum	Sim
45. Alguém da Família tem/teve Câncer	Discretizado	Sim / Não	Descarte
45.1 SNC	Discretizado	Sim / Não	Descarte
45.1 Digestório	Discretizado	Sim / Não	Descarte
45.1 Respiratório	Discretizado	Sim / Não	Descarte
45.1 Reprodutor	Discretizado	Sim / Não	Descarte
45.1 Glandular	Discretizado	Sim / Não	Descarte
45.1 Pele _Osso _Sangue	Discretizado	Sim / Não	Descarte
46.1 Ch- T1	Contínuo	[VR < 15,5] - Baixo, [VR >= 15,5] - Normal	Sim
46.2 Ch- E1	Contínuo	[VR < 32] - Baixo, [VR >= 32] - Normal	Sim

Nome original do dado	Tipo	Discretização	Concluído
46.3 Ch-P1	Contínuo	[VR < 1,3] - Baixo, [VR >= 1,3] - Normal	Sim
47. AST1	Contínuo	[VR < 4] - Baixo, [VR >= 4 e VR <= 36] - Normal	Sim
48. ALT1	Contínuo	[VR < 4] - Baixo, [VR >= 4 e VR <= 32] - Normal	Sim
49. γ- GT1	Contínuo	[VR < 7] - Baixo, [VR >= 7 e VR <= 50] - Normal	Sim
50. CREATININA1	Contínuo	HOMENS [VR < 0,9] - Baixo, [VR >= 0,9 e VR <= 1,3] - Normal MULHERES [VR < 0,6] - Baixo, [VR >= 0,6 e VR <= 1,1] - Normal	Sim
51. Encaminhado ao Ambulatório de Doença Ocupacional	Discretizado	Sim / Não	Descarte
52. Hábito alimentar	Contínuo	(0)Totalmente inadequado, (1) Inadequado, (2) Razoável, (3) Saudável, (4)Totalmente saudável, (9) Não Informado	Descarte
53. IMC	Contínuo	<18,5 - Baixo Peso, 18,5-24,99 - Peso Normal, 25-29,99 - Sobrepeso, 30-34,99 - Obeso I, 35-39,99 - Obeso II, >40 - Obeso III	Sim
54. Circunferência abdominal	Contínuo	Homem [*, 101] – Adequada, [102, *] - Inadequada Mulher [*, 87] – Adequada, [88, *] - Inadequada	Sim

Nome original do dado	Tipo	Discretização	Concluído
55. Você (ou sua esposa) teve dificuldade engravidar	Discretizado	Sim / Não	Sim
56. Você (ou sua esposa) teve abortamento espontâneo	Discretizado	Sim / Não	Sim
57. Tem filhos	Discretizado	Sim / Não	Descarte
58. Algum filho com mal formação (congênita)	Discretizado	Sim / Não	Sim
59. Você toma algum medicamento de uso contínuo	Discretizado	Sim / Não	Descarte
59.1 Nome do medicamento	Multivalorado	Lista de Medicamento	Descarte
60. Usou remédio para micose nos últimos dois meses	Discretizado	Sim / Não	Descarte
60.1 Nome do remédio	Multivalorado		Descarte
Diagnóstico	Nulo	Lista de Diagnósticos	Sim

D Anexo: Resultado do Refinamento do dado Produto (Dado Multivalorado) no projeto PCS

Produto	**Classe Química**
Abamectina	Avermectina
Acefato FERSOL	Organofosforado
Acefato NORTOX	Organofosforado
Acetamiprid NORTOX	Neonicotinoide
Actara 250 WG	Neonicotinoide
Afalon	Glicina Substituída
Agritoato	Organofosforado
Akito	Piretroide
Aldicarbe	Carbamato
Ally	Metsulfurom Metílico
Alpha-Cipermetrina	Piretroide
Alto	Triazol
Alto 100	Triazol
Amistar	Estrobilurina
Astro	Organofosforado
Atrazina	Triazina
Attamix	Piretroide
Avant 750 SP	Organofosforado
Bamako 700 WG	Neonicotinoide
Baysiston	Organofosforado
Baytan	Triazol
Bifentrina 100 ec Nortox	Piretroide
Bitrin 100 EC	Piretroide
Bolfo	Carbamato
Bravonil	Isoftalonitrila
Butox	Piretroide
Cabrio Top	Ditiocarbamato/Triazol
Cantus	Anilida
Captus	Ciclodienoclorado
Carbaril	Carbamato
Carbofuran	Carbamato
Carbofurano	Carbamato
Carbosulfan	Carbamato
Cefanol	Organofosforado
Cercobin	Benzimidazol

Produto	**Classe Química**
Cerconil	Isoftalonitrila
Clordano	Organoclorado
Clorpirifós	Organofosforado
Clotianidina	Neonicotinoide
Colosso	Organofosforado
Connect	Piretroide
Cuprozeb	Ditiocarbamato
Curyom	Organofosforado
Cyptrin	Piretroide
Danimen 300 EC	Piretroide
DDT	Piretroide
Decis 25 EC	Piretroide
Deltamethrin	Piretroide
Deltametrina	Piretroide
Diazinon	Organofosforado
Difenoconazol	Carbamato
Dimetoato	Organofosforado
Dimexion	Organofosforado
Dinno	Neonicotinoide
Disulfoton	Organofosforado
Dithane NT	Ditiocarbamato
Dual Gold	Cloroacetanilida
Endossulfam	Organoclorado
Engeo Pleno	Piretroide
Esfenvalerato	Piretroide
Etofenprox	Piretroide
Evidence	Neonicotinoide
Fastac	Piretroide
Fenitrothion	Organofosforado
Fenpropathrin	Piretroide
Fibronil	Pirazol
Finale	Homoalanina Substituída
Flex	Éter Difenílico
Flumyzin	Ciclohexenodicarboximida
Flupiradifurona	Butenolidas
Folicur	Triazol
Folidol	Organofosforado
Furazin 310 FS	Carbamato
Galop M	Picloram/2,4-D
Gesaprim 500	Triazina
Glifosato	Glicina Substituída

Produto	**Classe Química**
Glifozap	Glicina Substituída
Gramocil	Ureia + Bipiridilio
Granutox	Organofosforado
Imidacloprido	Neonicotinoide
Impact	Triazol
Inside FS	Neonicotinoide
Karate - Lambda-Cialotrina	Piretroide
Karate Zeon	Piretroide
Klorpan	Organofosforado
K-Othrine	Piretroide
Kraft	Avermectina
Lambda-Cialotrina	Piretroide
Lannate	Carbamato
Lembra	Glicina Substituída
Lepecid	Organofosforado
Losban	Organofosforado
Malationa	Organofosforado
Manzate	Ditiocarbamato
Metamidofós	Organofosforado
Metiocarbe	Carbamato
Metiram	Carbamato
Metomil	Carbamato
Metoxocloro	Piretroide
Monocrotophos	Organofosforado
Mospilam	Neonicotinoide
Nativo	Triazol
Nitenpiram	Neonicotinoide
Opera	Triazol
Orthene	Organofosforado
Oxamil	Carbamato
Paration Metílico	Organofosforado
Phorate	Organofosforado
Pirate	Pirazol
Piretrinas	Piretroide
Pirimicarb	Carbamato
Pirimifós Metil	Organofosforado
Polyram	Carbamato
Polytrin	Organofosforado
Premier	Neonicotinoide
Premier Plus	Neonicotinoide
Premio	Antranilamida

Produto	**Classe Química**
Primeplus Br - Flumetralina	Neonicotinoide
Priori Xtra	Triazol
Prisma	Carbamato
Profenofós	Organofosforado
Proof	Triazina
Pyridaphenthion	Organofosforado
Regente	Pirazol
Ridomil	Ditiocarbamato
Riza	Triazol
Roundup	Glicina Substituída
Rubric - Epoxiconazol	Triazol
Rumo - Indoxacarbe	Oxadiazina
Sabre	Organofosforado
Sanson	Sulfonilureia
Score	Triazol
Servin 480SC	Carbamato
Shake	Triazol
Simboll	Triazol
Sulfoxaflor	Sulfoxaminas
Sumiban 150 SC	Piretroide
Supera	Inorgânico
Tamaron	Organofosforado
Tebuconazole	Triazol
Teflutrina	Piretroide
Terbufós	Organofosforado
Thiamethoxam	Neonicotinoide
Thiodan	Ciclodienoclorado
Thiodicarb	Carbamato
Tiacloprido	Neonicotinoide
Trebon 100 SC	Éter Difenílico
Triade	Triazol
Triazofós	Organofosforado
Triclorfon	Organofosforado
Triflumezoprim	Neonicotinoide
Trinca	Piretroide
Verdadero	Triazol
Verdict	Ácido Ariloxifenoxipropiônico
Zap	Glicina Substituída
Zartan	Glicina Substituída

E Anexo: Scripts em Linguagem R usados no Treinamento no projeto PCS

```r
#SCRIPT 1
#Leitura do arquivo csv.
TabelaFonte <- read.table("[caminho\\arquivo]", header=TRUE, sep=";")
#Converção de uma tabela do tipo data.frames em um objeto do tipo DecisionTable.
# O número 15 indica quantos atributos possui a tabela e o atributo de decisão.
Relacao <- SF.asDecisionTable(dataset=TabelaFonte, decision.att = 15, indx.nominal = 15)
#Construção da matriz de discernibilidade para a geração dos redutos.
res <- BC.discernibility.mat.RST(Relacao, range.object = NULL)
reduto <- FS.all.reducts.computation(res)
#Salva arquivo de redutos.
sink(file = "[caminho\\arquivo]")
print(reduto)
sink()
## Geração da da tabela de decisão do reduto escolhido.
# O indx de cada reduto se encontra no arquivo referente ao reduto.
new.Relacao <- SF.applyDecTable(Relacao, reduto, control = list(indx.reduct = 1))
#Geração de regras e eliminação de regras duplicadas.
Regras <- structure(new.Relacao, class = "data.frame")
dl <- Regras[!duplicated(Regras),]
#Salva arquivo de regras.
write.table(dl, file = "[caminho\\arquivo]", row.names = FALSE, sep = ",")
```

```r
#SCRIPT 2
#Leitura do arquivo csv.
TabelaFonte <- read.table("[caminho\\arquivo]", header=TRUE, sep=";")
#Converção de uma tabela do tipo data.frames em um objeto do tipo DecisionTable.
# O número 15 indica quantos atributos possui a tabela e o atributo de decisão.
Relacao <- SF.asDecisionTable(dataset=TabelaFonte, decision.att = 80, indx.nominal = 80)
#Calculo de reduções de decisão aproximadas com base na aproximação de subconjuntos de atributos.
reduto <-FS.greedy.heuristic.reduct.RST(Relacao,qualityF = X.nOfConflicts)
#Salva arquivo de redutos.
sink(file = "[caminho\\arquivo]")
print(reduto)
sink()
## Geração da da tabela de decisão do reduto escolhido.
# O indx de cada reduto se encontra no arquivo referente ao reduto.
new.Relacao <- SF.applyDecTable(Relacao, reduto, control = list(indx.reduct = 1))
#Geração de regras e eliminação de regras duplicadas.
Regras <- structure(new.Relacao, class = "data.frame")
dl <- Regras[!duplicated(Regras),]
#Salva arquivo de regras.
write.table(dl, file = "[caminho\\arquivo]", row.names = FALSE, sep = ",")
```

F Anexo: Síntese dos Formulários de Treinamento para Geração de Redutos no projeto PCS

Grupo: 1 - Testes: A,B,C,D		
Redutos	**Atributos**	**Testes**
Ultimo_ContatoDias2, CH_T, CH_E, CH_P, AST	5	A,B,C,D
Forma_Aplicacao, CH_T, CH_E, CH_P, AST	5	A,B,C,D
Tipo_Contato, CH_T, CH_E, CH_P, AST	5	A,B,C,D
CH_T, CH_E, CH_P, AST, CREATININA	5	A,B,C,D
ClasseQuimica, CH_T, CH_E, CH_P, AST	5	C
Tempo_Contato2, CH_T, CH_E, CH_P, AST	5	A
Tempo_Contato2, ClasseQuimica, CH_T, CH_E, CH_P, AST	6	B,D
ClasseQuimica, Equip_Protecao, CH_T, CH_E, CH_P, AST	6	A,B,D
ClasseQuimica, Roupa_Apropriada, CH_T, CH_E, CH_P, AST	6	A,B,D
ClasseQuimica, Bota_Apropriada, CH_T, CH_E, CH_P, AST2	6	A,B,D
ClasseQuimica, Luvas, CH_T, CH_E, CH_P, AST	6	A,B,D

Grupo: 2 - Testes: A,B,C,D		
Redutos	**Atributos**	**Testes**
CH_T, CH_E, CH_P, AST, CREATININA	5	A,B,C,D
CH_T, CH_E, CH_P, AST, IMC2	5	D
CH_T, CH_E, CH_P, AST, Circunferencia_Abdominal2	5	A,B,C,D
CH_T, CH_E, CH_P, AST, Aborto_Espontaneo	5	A,B,C,D
Idade2, CH_T, CH_E, CH_P, AST	5	A
Idade2, CH_T, CH_E, CH_P, AST, IMC2	6	B,C
Tabagismo, CH_T, CH_E, CH_P, AST, IMC2	6	A,B,C
Tabagismo_Atual, CH_T, CH_E, CH_P, AST, IMC2	6	A,B,C
Tabagismo_Anterior, CH_T, CH_E, CH_P, AST, IMC2	6	A,B,C
Etilismo, CH_T, CH_E, CH_P, AST, IMC2	6	A,B,C
Etilismo_Atual, CH_T, CH_E, CH_P, AST, IMC2	6	A,B,C
Etilismo_Anterior, CH_T, CH_E, CH_P, AST, IMC2	6	A,B,C
Cafe_mlDia2, CH_T, CH_E, CH_P, AST, IMC2	6	A,B,C
CH_T, CH_E, CH_P, AST, IMC2, Filho_MaFormacao	6	A,B,C

Grupo: 3 - Testes: A,B,C,D		
Redutos	**Atributos**	**Testes**
CH_T, CH_E, CH_P, AST, CREATININA	5	A,B,C,D
ClasseQuimica, CH_T, CH_E, CH_P, AST	5	C

Grupo: 4, 8, 9, 10, 11 - Testes: A,B,C,D		
Redutos	**Atributos**	**Testes**
CH_T, CH_E, CH_P, AST, CREATININA	5	A,B,C,D

Grupo: 5 - Testes: A,B,C,D		
Redutos	**Atributos**	**Testes**
Alteracao_Snervoso, CH_T, CH_E, CH_P, AST	5	A,B,C,D
Dor_Cabeca, CH_T, CH_E, CH_P, AST	5	A,B,C,D
Incoordenacao_Motora, CH_T, CH_E, CH_P, AST	5	A,B,C,D
CH_T, CH_E, CH_P, AST, CREATININA	5	A,B,C,D
Fraqueza_Muscular, CH_T, CH_E, CH_P, AST	5	B
Tremedeira, CH_T, CH_E, CH_P, AST	5	B
Fraqueza_Muscular, Formigamento_MMSS, CH_T, CH_E, CH_P, AST	6	A,C,D
Tremedeira, Formigamento_MMSS, CH_T, CH_E, CH_P, AST	6	A,C,D
Tremor_Muscular, Formigamento_MMSS, CH_T, CH_E, CH_P, AST	6	A,B,C,D
VisaoTurva_Embacada, Formigamento_MMSS, CH_T, CH_E, CH_P, AST	6	D
Vertigens_Tonturas, Formigamento_MMSS, CH_T, CH_E, CH_P, AST	6	D
Tremor_Muscular, VisaoTurva_Embacada, CH_T, CH_E, CH_P, AST	6	B
VisaoTurva_Embacada, Agitacao_Irritabilidade, CH_T, CH_E, CH_P, AST	6	B
Tremor_Muscular, Agitacao_Irritabilidade, CH_T, CH_E, CH_P, AST	6	B
Agitacao_Irritabilidade, Vertigens_Tonturas, CH_T, CH_E, CH_P, AST	6	B
VisaoTurva_Embacada, Agitacao_Irritabilidade, Formigamento_MMSS, CH_T, CH_E, CH_P, AST	7	C
Agitacao_Irritabilidade, Vertigens_Tonturas, Formigamento_MMSS, CH_T, CH_E, CH_P, AST	7	C

Grupo: 6 - Testes: A,B,C,D		
Redutos	**Atributos**	**Testes**
Vomito, CH_T, CH_E, CH_P, AST	5	A,B,C,D
CH_T, CH_E, CH_P, AST, CREATININA	5	A,B,C,D
Aparelho_Digestorio, CH_T, CH_E, CH_P, AST	5	B
Dor_Estomago, CH_T, CH_E, CH_P, AST	5	B
Azia_Queimacao, CH_T, CH_E, CH_P, AST	5	B
Aparelho_Digestorio, Diarreia, CH_T, CH_E, CH_P, AST	6	A,C,D
Colicas_DorBarriga, Diarreia, CH_T, CH_E, CH_P, AST	6	D
Dor_Estomago, Diarreia, CH_T, CH_E, CH_P, AST	6	A,C,D
Azia_Queimacao, Diarreia, CH_T, CH_E, CH_P, AST	6	A,C,D
Nauseas_Enjoo, Diarreia, CH_T, CH_E, CH_P, AST	6	A,B,C,D

Grupo: 7 - Testes: A,B,C,D		
Redutos	**Atributos**	**Testes**
CH_T, CH_E, CH_P, AST, CREATININA	5	A,B,C,D
Aparelho_Respiratorio, Irritacao_Nasal, CH_T, CH_E, CH_P, AST	6	D
Falta_DeAr, Irritacao_Nasal, CH_T, CH_E, CH_P, AST	6	D
Falta_DeAr, Catarro_Escarro, CH_T, CH_E, CH_P, AST	6	D
Falta_DeAr, Tosse, CH_T, CH_E, CH_P, AST	6	C
Irritacao_Nasal, Tosse, CH_T, CH_E, CH_P, AST	6	A
Aparelho_Respiratorio, Irritacao_Nasal, Tosse, CH_T, CH_E, CH_P, AST	7	B,C
Falta_DeAr, Irritacao_Nasal, Tosse, CH_T, CH_E, CH_P, AST	7	B
Aparelho_Respiratorio, Catarro_Escarro, Tosse, CH_T, CH_E, CH_P, AST	7	B
Falta_DeAr, Catarro_Escarro, Tosse, CH_T, CH_E, CH_P, AST	7	A,B

G Anexo: Código do Formulário de Avaliação de Regras no projeto PCS

```
class ClasseRegras
    {
        private SqlConnection cn = new SqlConnection();
        private SqlCommand cd = new SqlCommand();
        private SqlDataReader dr;
        public int c;

        public void Conectar()
        {
            string s = "";
            s = @"Server=.\SQLEXPRESS;Database=Doutorado;UID=sa;PWD=123";
            cn.ConnectionString = s;
        }

        public void AlterarDiagnostico(string sql)
        {
            Conectar();
            cn.Open();
            cd.Connection = cn;
            cd.CommandText = sql;
            cd.ExecuteNonQuery();
            cn.Close();
        }

        public void Verifica(string sql)
        {
            c = 0;
            Conectar();
            cn.Open();
            cd.Connection = cn;
            cd.CommandText = sql;
            dr = cd.ExecuteReader();
            while (dr.Read())
            {
                c++;
            }
            cn.Close();
        }

}
```

```
public partial class FrmAvalia : MetroFramework.Forms.MetroForm
    {
        public FrmAvalia()
        {
            InitializeComponent();
        }
        private int contador = 0;
        private string sql = "";
        private string[] campos;
        private string[] valores;
        private ClasseRegras regras = new ClasseRegras();
        private void btnArquivo_Click(object sender, EventArgs e)
        {
            sql += "Update " + txtTabelaTeste.Text + " set DiagnosticoRegras = ''";
            regras.AlterarDiagnostico(sql);
            sql = "";
            char[] charsToTrim = { '"', '\\' };
            if (openFileDialog1.ShowDialog() == DialogResult.OK)
            {
                StreamReader sr = new StreamReader(openFileDialog1.FileName,Encoding.UTF8);
                while (!sr.EndOfStream)
                {
                    string linha = sr.ReadLine();
                    if (contador == 0)
                    {
                        campos = linha.Split(',');
                        for (int i = 0; i < campos.Length; i++)
                        {
                            campos[i] = campos[i].Trim(charsToTrim);
                        }
                    }
                    else
                    {
                        valores = linha.Split(',');
                        for (int i = 0; i < valores.Length; i++)
                        {
                            valores[i] = valores[i].Trim(charsToTrim);
                        }
                    }
                    if (contador > 0)
                    {
                        sql += "Update " + txtTabelaTeste.Text;
                        sql += " set DiagnosticoRegras = '" + valores[5] + "' ";
                        sql += "where " + campos[0] + "= '" + valores[0] + "' ";
                        sql += "and " + campos[1] + "= '" + valores[1] + "' ";
                        sql += "and " + campos[2] + "= '" + valores[2] + "' ";
                        sql += "and " + campos[3] + "= '" + valores[3] + "' ";
                        sql += "and " + campos[4] + "= '" + valores[4] + "' ";

                        regras.AlterarDiagnostico(sql);
                        sql = "";
                    }
                    contador++;
                }
                MessageBox.Show("Fim");
            }
        }

        private void btnVerifica_Click(object sender, EventArgs e)
        {
            sql += "Select * from " + txtTabelaTeste.Text;
            sql += "Where Diagnostico <> DiagnosticoRegras";
            regras.Verifica(sql);
            label1.Text= regras.c.ToString();
            sql = "";
        }
    }
}
```