UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

JOÃO PAULO PEREIRA SANTA CLARA
MURILLO IAGO MOREIRA

APLICAÇÃO DA METODOLOGIA CRISP-DM: UM ESTUDO SOBRE DIABETES UTILIZANDO BASES DE DADOS PÚBLICAS

PONTA GROSSA
2024

JOÃO PAULO PEREIRA SANTA CLARA
MURILLO IAGO MOREIRA

APLICAÇÃO DA METODOLOGIA CRISP-DM: UM ESTUDO SOBRE DIABETES UTILIZANDO BASES DE DADOS PÚBLICAS

Trabalho de conclusão de curso de graduação apresentada como requisito para obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador(a): Prof Esp Marcos Vinicius Fidelis

PONTA GROSSA
2024

JOÃO PAULO PEREIRA SANTA CLARA
MURILLO IAGO MOREIRA

CONSTRUÇÃO DE CLASSIFICADORES CRISP-DM: UM ESTUDO SOBRE DIABETES UTILIZANDO BASES DE DADOS PÚBLICAS

Trabalho de conclusão de curso de graduação apresentada como requisito para obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador: Prof. Esp. Marcos Vinicius Fidelis

Data de aprovação: 29/maio/2024

_____
Marcos Vinicius Fidelis
Especialização
Universidade Tecnológica Federal do Paraná

_____
Simone de Almeida
Doutorado
Universidade Tecnológica Federal do Paraná

_____
Geraldo Ranthum
Doutorado
Universidade Tecnológica Federal do Paraná

PONTA GROSSA
2024

Dedico este Trabalho de Conclusão de Curso à todas as pessoas da minha família, a meus amigos e colegas de graduação, ao nosso orientador Marcos Vinicius Fidelis pelo conhecimento que nos passou e a todos os companheiros e futuros companheiros de computação.

AGRADECIMENTOS

Primeiramente gostaríamos de agradecer a Deus pelo decorrer do nosso curso em Análise e Desenvolvimento de Sistemas, mesmo passando por diversas dificuldades somos extremamente gratos pelos ensinamentos de nossos professores e professoras durante nossa formação.

Gostaríamos de agradecer também a nossas famílias, por aqueles que estão conosco e por aqueles que já se foram, nós os agradecemos por todo o apoio que nos deram no decorrer de nossa jornada, buscando sempre nos aperfeiçoar cada dia mais.

Somos gratos a nosso excelente orientador e professor Marcos Vinicius Fidelis que nos ensinou primeiramente sobre o conceito de mineração de dados em primeiro momento e nos aceitou como orientandos para a realização deste trabalho de conclusão de curso. Somos extremamente gratos professor Fidelis por nos ajudar em nossa graduação.

A todos os professores do Departamento Acadêmico de Informática conhecido como DAINF nós os agradecemos por seus ensinamentos, curiosidades e inspirações para todos os alunos do curso de Análise e Desenvolvimento de Sistemas.

Nossos mais sinceros agradecimentos a todos vocês, aos colegas e amigos que fizemos no decorrer do curso e a futuras amizades que podem vir a surgir com o tempo. Muito Obrigado!

Seja você quem for, seja qual for a posição social que você tenha na vida, a mais alta ou a mais baixa, tenha sempre como meta muita força, muita determinação e sempre faça tudo com muito amor e com muita fé em Deus, que um dia você chega lá. De alguma maneira você chega lá.
(Senna; Ayrton, 1994).

RESUMO

Este estudo tem como objetivo principal explorar a eficácia e relevância do modelo CRISP-DM na análise de dados relacionados à diabetes, com o intuito de desenvolver modelos preditivos que contribuam para o entendimento da doença e suas complicações. Inicialmente, realizou-se uma revisão bibliográfica sobre o uso do CRISP-DM na literatura, destacando sua flexibilidade e capacidade de adaptação a diferentes projetos e conjuntos de dados. Essa fase permitiu uma compreensão detalhada da estrutura e das etapas do modelo, visando a aplicação prática no contexto da diabetes. Em seguida, foram analisados conjuntos de dados públicos sobre diabetes, provenientes de diversas fontes, como cuidados clínicos, registros de qualidade e dados administrativos. Cada fonte foi avaliada quanto às suas principais vantagens e desvantagens, levando em consideração aspectos como qualidade, representatividade e acessibilidade dos dados. Essa análise detalhada proporcionou uma visão abrangente das informações disponíveis e orientou a seleção dos conjuntos de dados mais adequados para o estudo. Durante o processo de análise, foi elaborada uma documentação das fases do modelo CRISP-DM, adaptada às necessidades e requisitos específicos do estudo de caso. Essa documentação não apenas descreveu cada etapa do processo, desde a compreensão do negócio até a implantação dos modelos, mas também forneceu orientações práticas para sua execução. Além disso, serviu como uma ferramenta educacional, facilitando a compreensão dos detalhes técnicos do estudo e promovendo a disseminação do conhecimento sobre mineração de dados. Os resultados obtidos do uso do CRISP-DM foram analisados e interpretados em relação aos objetivos estabelecidos no início do estudo. Os modelos preditivos desenvolvidos representam ferramentas que podem auxiliar na identificação de riscos da doença, na implementação de estratégias de prevenção e controle da diabetes e na aplicação de um modelo de processos norteando o processo de construção de classificadores. Em conclusão, este estudo demonstrou que o modelo CRISP-DM é uma abordagem eficaz e abrangente para a análise de dados relacionados à diabetes. Sua aplicação permitiu uma compreensão mais profunda dos padrões e fatores de risco associados à doença, contribuindo significativamente para o avanço do conhecimento na área da saúde e dos processos de mineração de dados. O sucesso deste projeto valida a importância do CRISP-DM como uma ferramenta poderosa para impulsionar futuras pesquisas e inovações no campo da mineração de dados e da saúde pública.

Palavras-chave: mineração de dados; diabetes; CRISP-DM; classificadores.

ABSTRACT

The main objective of this study is to explore the effectiveness and relevance of the CRISP-DM model in the analysis of diabetes-related data, with the aim of developing predictive models that contribute to the understanding of the disease and its complications. Initially, a literature review on the use of CRISP-DM in the literature was carried out, highlighting its flexibility and ability to adapt to different projects and datasets. This phase allowed a detailed understanding of the structure and stages of the model, aiming at its practical application in the context of diabetes. Next, public datasets on diabetes were analyzed, coming from various sources, such as clinical care, quality registries and administrative data. Each source was evaluated for its main advantages and disadvantages, taking into account aspects such as data quality, representativeness and accessibility. This detailed analysis provided a comprehensive view of the available information and guided the selection of the most appropriate datasets for the study. During the analysis process, a documentation of the phases of the CRISP-DM model was prepared, adapted to the specific needs and requirements of the case study. This documentation not only described each step of the process, from understanding the business to implementing the models, but also provided practical guidance for its execution. In addition, it served as an educational tool, facilitating the understanding of the technical details of the study and promoting the dissemination of knowledge about data mining. The results obtained from the use of CRISP-DM were analyzed and interpreted in relation to the objectives established at the beginning of the study. The predictive models developed represent tools that can assist in the identification of disease risks, in the implementation of diabetes prevention and control strategies, and in the application of a process model to guide the process of building classifiers. In conclusion, this study demonstrated that the CRISP-DM model is an effective and comprehensive approach for the analysis of diabetes-related data. Its application allowed a deeper understanding of the patterns and risk factors associated with the disease, contributing significantly to the advancement of knowledge in the health area and data mining processes. The success of this project validates the importance of CRISP-DM as a powerful tool to drive future research and innovation in the field of data mining and public health.

Keywords: data mining; diabetes; CRISP-DM; classifiers.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

CRISP-DM	*Cross Industry Standard Process for Data Mining*
DM	*Data Mining*
Diabetes	*Diabetes Mellitus*
KDD	*Knowledge Discovery in Databases*
MLP	*MultiLayer Perceptron*
SPSS	*Statistical Package for the Social Science*
TDSP	*Team Data Science Process*

SUMÁRIO

1 INTRODUÇÃO

Dados são importantes e continuam crescendo no mundo contemporâneo, são descritos como um conjunto de caracteres agrupados e com um propósito como informações registradas e analisadas. São a espinha dorsal da tecnologia além de serem o principal foco para a tomada de decisão como descrito por Chaudhary (2023), que em grande quantidade, proporciona uma oportunidade única para a aplicação de técnicas de mineração de dados, conhecidas como *Data Mining* – DM, que não visa apenas descobrir padrões diversos, mas, construir modelos preditivos. (Avelar; Rocha; Cruz; 2017 p.33 *apud* Zaki; Meira Junior, 2014).

O *Data Mining* é uma das fases do processo de extração de conhecimento em bases de dados conhecido como KDD – *Knowledge Discovery in Databases,* apresentada como:

> Uma etapa do processo KDD que consiste na aplicação de algoritmos de análise e descoberta de dados que, sob limitações aceitáveis de eficiência computacional, produzem uma enumeração específica de padrões sobre os dados (Fayyad et al, 1996, p. 83).

A ideia de buscar conhecimento precisa de organização e uma apresentação para quem for utilizar, podendo ser um simples relatório ou algo com maior grau de complexidade, descrição da fase de implantação do modelo CRISP-DM (C*ross Industry Standard Process for Data Mining*), um modelo de processo hierárquico proposto por Chapman et al (2000) que, em conjunto a métodos e técnicas de mineração de dados e ajuda de profissionais mais experientes, pode ser uma ferramenta valiosa para ajudar os analistas de mineração de dados menos experientes a compreender o valor e as etapas de todo o processo de mineração de dados. (Shearer; 2000 p. 19).

O diabetes - Diabetes Mellitus, atualmente um dos principais problemas de saúde, segundo Brunner e Suddarth (2006) já é a quarta maior causa de morte no Brasil, conforme referenciado por Pace e Nunes (2006). A escolha de abordar a diabetes surge em resposta ao alarmante aumento no número de casos. Em 2021, a Fiocruz registrou 17 milhões de diagnósticos de diabetes no Brasil, colocando-o como o quinto país com o maior número de casos. Globalmente, a OMS relatou cerca de 422 milhões de casos de diabetes, resultando em 1,5 milhão de mortes anuais relacionadas à doença (OMS, 2023).

Diante disso, é possível obter vastas bases de dados sobre diabetes de várias fontes, como cuidados clínicos, registros de qualidade e dados administrativos. Cada fonte tem suas vantagens e desvantagens (Wild et al, 2016). Por exemplo, os dados de ensaios clínicos podem oferecer alta qualidade, mas podem ser limitados em representatividade, enquanto os dados de fontes não destinadas à pesquisa podem ser mais acessíveis, embora possam conter lacunas. É crucial interpretar os resultados com cautela, considerando as limitações de cada conjunto de dados.

No entanto, o grande volume de dados e a complexidade das informações (Chaudhary, 2023) requerem uma abordagem metodológica robusta para extrair *insights* significativos. Como podemos aplicar de forma eficaz processos de mineração a grandes volumes de dados de forma flexível, adaptável e com enfoque interativo?

Sendo assim, o intuito deste trabalho é o estudo da metodologia CRISP-DM utilizando bases de dados relacionados à diabetes, a fim de contribuir para o estudo de Mineração de Dados e ajudar no entendimento da doença e suas complicações. A análise comparativa desta metodologia com outras técnicas utilizadas para o mesmo fim, visa contribuir para uma compreensão mais aprofundada dos padrões e fatores de risco associados à diabetes. Além disso, uma parte crucial deste estudo, é a elaboração de uma documentação abrangente do processo de mineração de dados, destinada tanto a especialistas quanto a pessoas interessadas que não estejam familiarizadas com os detalhes técnicos.

Esta documentação visa fornecer uma explicação clara e acessível de todas as etapas do processo, desde a coleta e preparação dos dados até a interpretação dos resultados. Desta forma, pretende-se tornar a informação sobre a análise de dados mais acessível e compreensível para um público mais amplo, facilitando a disseminação do conhecimento e o entendimento dos *insights* derivados da mineração de dados em relação à diabetes.

1.1 Justificativa

Segundo Sridharan (2018), a metodologia CRISP-DM – *Cross Industry Standard Process for Data Mining* é flexível em relação à tecnologia e aos problemas enfrentados, onde pode ser adaptado para atender às necessidades específicas de diferentes projetos e conjuntos de dados. A alta flexibilidade é importante ao lidar com grandes volumes de dados, onde as estratégias de análise podem precisar ser ajustadas conforme novas informações são descobertas ou requisitos de negócios

evoluem. O CRISP-DM ainda enfatiza a natureza interativa do processo de mineração de dados, permitindo que as equipes revisem e refinem etapas anteriores conforme avançam no projeto. Isso é importante ao lidar com grandes volumes de dados, onde a análise inicial pode revelar *insights* que exigem uma abordagem diferente ou mais detalhada (Chapman, 2000).

Foi realizada uma busca no portal ACM *Digital Library* sobre artigos de análise de dados e classificadores. No portal foram encontrados mais de 6 artigos relevantes sobre o uso de *Machine Learning* aplicado à doença Diabetes, onde 6 deles foram selecionados para estudo e nenhum utilizou o CRISP-DM.

Também foi realizada uma busca em bases de dados públicas disponíveis, um dos locais escolhidos foi o *site* da *UCI Machine Learning Repository* no qual encontram-se 622 bases de dados e, 3 delas são de diabetes com dados mais recentes de 2020 e mais antigos da década de 90, permitindo analisar dados antigos e mais recentes para esta proposta. As bases contam com os seguintes números de Hits:

Tabela 1 - Ocorrências Diabetes Data Set (Abril de 2023)

Data sets	Ocorrências
General Diabetes Data Set	747700 *Hits*
Diabetes 130-US hospitals for years 1999-2008 Data Set	439732 *Hits*
Early-stage diabetes risk prediction dataset.	114833 *Hits*

Fonte: Autoria própria (2024)

A escolha do tema diabetes também se deu pelo seu número de pessoas que convivem com a doença que de acordo com a *International Diabetes Federation* fez um levantamento que no ano de 2021 existem cerca de 537 milhões de diabéticos, além de estimativas como a de em 2030 cerca de 537 milhões de pessoas na casa dos 20 aos 79 anos vivem com a doença, e uma projeção futura no qual no ano de 2045 cerca de 783 milhões de pessoas devem ter esta doença.

De acordo com o artigo publicado Análise de dados epidemiológicos e clínicos em pacientes com diabetes mellitus tipo 2, Freitas (2019) comenta sobre a área da saúde, que na análise de estática os dados permitem avaliar melhor a incidência, prevalências, causas e impacto de uma doença numa população, região ou grupo

étnico. Além disso, permite avaliar os fatores de risco e quantificar a eficácia de medidas de prevenção e tratamentos. Nos últimos anos, a análise de dados se tornou uma ferramenta importante para o estabelecimento de políticas públicas e como suporte para o desenvolvimento da medicina de precisão.

1.2 Objetivos

O Objetivo geral é aplicar a metodologia CRISP-DM, a fim de documentar de forma abrangente o processo previsto de Análise de Dados, tornando a informação acessível e compreensível para um público amplo.

Para conseguir atingir esse objetivo foi traçado os seguintes objetivos específicos:

1. Realizar uma revisão bibliográfica sobre artigos que utilizam o modelo de processos CRISP-DM.
2. Documentar detalhadamente o processo de mineração de dados seguindo as etapas realizadas em cada fase do CRISP-DM, adaptado às necessidades do projeto.
3. Implementar algoritmos de classificação utilizando ferramentas do WEKA.
4. Avaliar e comparar o desempenho obtido através dos modelos desenvolvidos, identificando os melhores resultados.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem como objetivo estabelecer as bases para a compreensão deste trabalho. Na seção 2.1 é apresentado o processo de extrair conhecimento a partir de bases de dados, conhecido como *Knowledge Discovery in Databases* - KDD, auxiliando na compreensão da mineração de dados, posteriormente abordando o modelo dos processos *Cross Industry Standard Process for Data Mining* - CRISP-DM.

Na seção 2.2.1 são apresentados os artefatos das fases do CRISP-DM, que é abordado na construção da solução para a proposta do trabalho.

No item 2.3 referente a outras metodologias existentes, é demonstrado o uso do *Statistical Package for the Social Science* (SPSS), utilizado para análise estatísticas avançadas, baseadas em algoritmos, utilizado principalmente nas áreas das Ciências Sociais e Ciências Exatas.

A metodologia, presente na seção 3, utiliza como base a utilização do *software Waikato Environment for Knowledge Analysi*s (WEKA), para a mineração de dados a partir de padrões encontrados para gerar soluções e explorar os classificadores utilizados que são embarcados ao WEKA para o pré-processamento de dados.

2.1 *Knowledge Discovery in Databases* – KDD

Chaudhary (2023) ressalta que os dados, sejam eles qualitativos ou quantitativos, representam um tesouro de informações extraídas de diversas fontes, como pesquisas científicas, registros governamentais e interações nas mídias sociais. Esses dados são a espinha dorsal da sociedade moderna, impulsionando a estratégia e a inovação. Eles emergem de transações comerciais e interações digitais, refletindo nossa existência coletiva.

No entanto, o verdadeiro potencial dos dados é revelado quando são analisados, transformando-se em insights valiosos que orientam decisões estratégicas em diversos setores. Com o avanço da tecnologia, como algoritmos avançados e inteligência artificial, somos capazes de processar grandes volumes de dados e descobrir padrões complexos, impulsionando inovações em áreas como veículos autônomos e saúde personalizada.

Em uma ampla variedade de campos de dados coletados e acumulados em um ritmo dramático, existe uma necessidade urgente de uma nova geração de tecnologia computacional, técnicas e ferramentas para ajudar os humanos a extraírem

informações completas do rápido crescimento volumes de dados. Essas técnicas e ferramentas são objetos de descoberta de conhecimento em bancos de dados (KDD) como observa (Fayyad, 1996).

O processo KDD é um método para extrair informações úteis dos dados, com a mineração de dados como uma de suas etapas cruciais, onde algoritmos específicos são aplicados para identificar padrões. Conforme descrito por Fayyad (1996), o KDD abrange uma série de etapas, desde a preparação e limpeza dos dados até a interpretação dos resultados, visando garantir a extração de conhecimento significativo e evitar descobertas de padrões sem sentido. Esta abordagem evoluiu a partir da interseção de diversos campos, incluindo aprendizado de máquina, reconhecimento de padrões, estatísticas, buscando extrair insights de alto nível a partir de grandes conjuntos de dados.

Enquanto compartilha afinidades com a estatística em métodos exploratórios de análise, o KDD se distingue por sua abordagem voltada para modelos de extração e operação em conjuntos de dados maiores e mais complexos. Além disso, o KDD está intimamente ligado ao armazenamento de dados e à análise de dados transacionais para suportar a tomada de decisões. Assim, o KDD se destaca como uma abordagem abrangente e multifacetada para a extração de conhecimento valioso a partir de dados. (Fayyad, 1996).

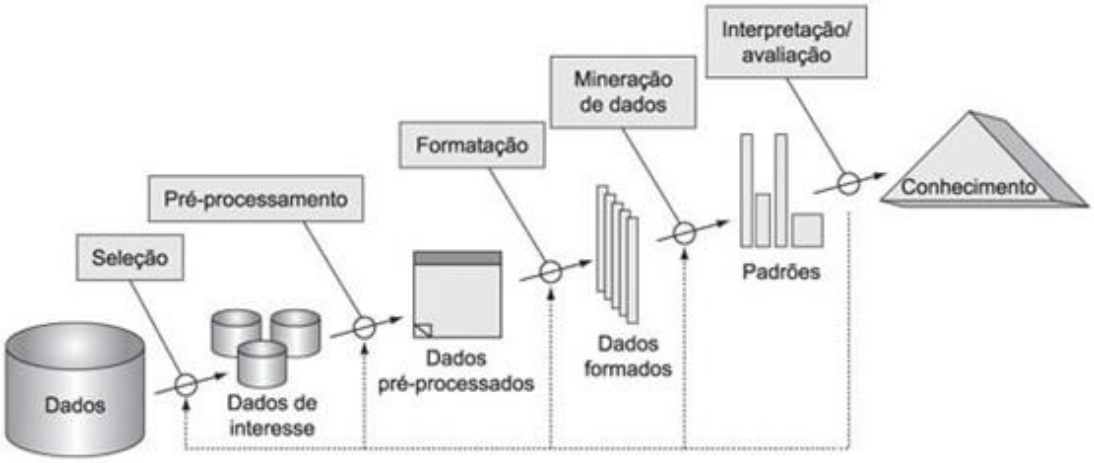
Brachman e Anand (1996 *apud* Fayyad et al., 1996, p 84) oferecem uma visão prática do processo de Descoberta de Conhecimento em Bases de Dados (KDD), destacando a natureza interativa do processo. Eles definem os seguintes passos básicos:

1. Identificar os objetivos do processo KDD a partir da perspectiva do cliente para compreender as áreas de aplicação e o conhecimento existente relacionado.
2. Crie um conjunto de dados de destino: Selecione um conjunto de dados ou concentrar em um subconjunto de variáveis ou amostra de dados na qual a pesquisa será realizada.
3. Limpeza e pré-processamento de dados: etapas básicas, como remoção de ruído, modelagem de ruído ou coleta de informações necessárias para cálculos, estratégias para lidar com campos de dados ausentes, dadas informações de séries temporais e alterações conhecidas.

4. Redução e projeção de dados: Encontre recursos úteis para exibir dados de acordo com a finalidade do seu trabalho. Use técnicas de redução ou transformação de dimensionalidade para reduzir o número de variáveis válidas a serem consideradas ou para encontrar uma representação invariante dos dados.
5. Adapte os objetivos do processo KDD (etapa 1) ao seu método específico de mineração de dados (por exemplo, agregação, classificação, regressão, *clustering*, etc.).
6. Selecione um algoritmo de mineração de dados: Escolha o método que deseja usar para encontrar padrões em seus dados. Ele define algoritmos e parâmetros apropriados para adequar um método específico de mineração de dados aos critérios gerais do processo KDD.
7. Mineração de Dados: Encontrar padrões de interesse em uma determinada forma de representação ou conjunto de representações (regras ou árvores de classificação, regressão, agrupamento, etc.). Os usuários podem beneficiar muito seus métodos de mineração de dados seguindo corretamente as etapas anteriores.
8. Interpretação de padrões de resultados. Você pode retornar a qualquer uma das etapas 1 a 7 para iterações adicionais. Esta atividade também pode incluir a visualização do modelo/modelo resultante ou a visualização de dados do modelo resultante.
9. Integração do conhecimento descoberto: Integre este conhecimento em outros sistemas para ações futuras ou simplesmente documente e reporte às partes interessadas. Envolve também examinar e resolver potenciais conflitos com conhecimentos pré-concebidos (ou adquiridos).

A explicação das fases pode ser compreendida a partir da Figura abaixo (1):

Figura 1 – Etapas do processo de KDD



Fonte: Fayyad et al. (1996).

De acordo com Weiss e Kulikowski (1991) e Hand (1981), a classificação, como método de mineração de dados, envolve a aprendizagem de uma função que classifica um item de dados em uma das várias classes predefinidas. Exemplos de sua aplicação incluem a classificação de tendências nos mercados financeiros (Apte e Hong, 1996) e a identificação automatizada de objetos de interesse em grandes bancos de dados de imagens (Fayyad, Djorgovski e Weir, 1996).

Para a regressão, uma função que mapeia um item de dados para uma variável de previsão com valor real, aplicações diversas são mencionadas, como prever a quantidade de biomassa em uma floresta com base em medições de micro-ondas de sensoriamento remoto, entre outros exemplos citados (apud Fayyad, Piatetsky-Shapiro e Smyth, 1996).

Jain e Dubes (1988) e Titterington, Smith e Makov (1985) explicam que o *clustering*, uma tarefa descritiva, busca identificar um conjunto finito de categorias ou clusters para descrever os dados. Exemplos de sua aplicação incluem a descoberta de subpopulações homogêneas para consumidores em bancos de dados de marketing e a identificação de subcategorias de espectros de medições infravermelhas do céu (Cheeseman e Stutz, 1996) (apud Fayyad, Piatetsky-Shapiro e Smyth, 1996).

Em relação à modelagem de dependências, Glymour et al. (1987) e Heckerman (1996) sugerem que este método consiste em encontrar um modelo que descreva dependências significativas entre variáveis, existindo em dois níveis: o estrutural e o quantitativo. Redes de dependência probabilísticas, conforme descrito por esses

autores, estão encontrando aplicações em diversos campos, como o desenvolvimento de sistemas especialistas médicos probabilísticos a partir de bancos de dados e a modelagem do genoma humano (apud Fayyad, Piatetsky-Shapiro e Smyth, 1996) (apud Fayyad, Piatetsky-Shapiro e Smyth, 1996).

Para a detecção de alterações e desvios, Berndt e Clifford (1996), Guyon, Matic e Vapnik (1996), Kloesgen (1996), Matheus, Piatetsky-Shapiro e McNeill (1996) e Basseville e Nikiforov (1993) afirmam que essa área se concentra na descoberta das alterações mais significativas nos dados a partir de valores previamente medidos ou normativos (apud Fayyad, Piatetsky-Shapiro e Smyth, 1996).

2.2 *Cross Industry Standard Process for Data Mining* - CRISP-DM

Segundo Azevedo e Santos (2008) por meio de um consorcio entre Daimler Chrysler (atual Mercedes-Benz Group), SPSS (*Statistical Package for Social Science for Windows*, conhecido como IBM SPSS, para softwares estatísticos) e NCR (focada em *data Warehousing*) nos anos 2000 foi produzido a metodologia para a mineração de dados conhecido como CRISP-DM. em conjunto a métodos e técnicas de mineração de dados e ajuda de profissionais mais experientes, pode ser uma ferramenta valiosa para ajudar os analistas de mineração de dados menos experientes a compreender o valor e as etapas de todo o processo de mineração de dados. (Shearer; 2000 p. 19).

Descrito por Sridharan (2018) está metodologia é flexível em relação à tecnologia e aos problemas enfrentados, onde pode ser adaptado para atender às necessidades específicas de diferentes projetos e conjuntos de dados. A alta flexibilidade é importante ao lidar com grandes volumes de dados, onde as estratégias de análise podem precisar ser ajustadas conforme novas informações são descobertas ou requisitos de negócios evoluem.

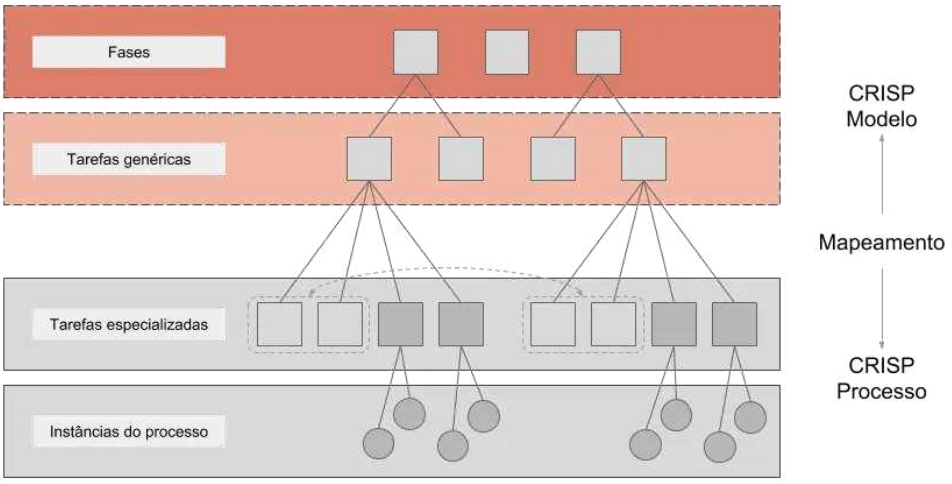
O CRISP-DM ainda enfatiza a natureza interativa do processo de mineração de dados, permitindo que as equipes revisem e refinem etapas anteriores conforme avançam no projeto. Isso é importante ao lidar com grandes volumes de dados, onde a análise inicial pode revelar insights que exigem uma abordagem diferente ou mais detalhada (Chapman et al. 2000).

Chapman et al. (2000) descreve a metodologia CRISP-DM com um conjunto de tarefas em quatro níveis de abstração (do geral ao específico): fases, tarefa genérica, tarefa especializada e instância de processo:

- No nível superior, o processo de mineração de dados é organizado em diversas fases; cada fase consiste em várias tarefas genéricas de segundo nível.
- Este segundo nível é denominado genérico, porque pretende ser geral o suficiente para cobrir todas as situações possíveis de mineração de dados.
- O terceiro nível de tarefa especializada, é o local para descrever como as ações no tarefas genéricas devem ser realizadas em determinadas situações específicas.
- O quarto nível, a instância do processo, é um registro das ações, decisões e resultados de um compromisso real de mineração de dados.

Abaixo é possível identificar uma visualização dessa divisão (Ver figura 2).

Figura 2 – Quatro níveis da metodologia CRISP-DM



Fonte: Chapman (2000), adaptado por Karina Moura (2019)

Shearer (2000) destaca que o modelo CRISP-DM fornece um plano completo para conduzir um projeto de mineração de dados, desde iniciantes até especialistas. Ele descreve a estrutura do processo que é demonstrado na Figura 2 logo abaixo, as fases desse processo. Nas palavras do autor:

> As flechas indicam as dependências mais importantes e frequentes entre as fases, enquanto o círculo externo simboliza a natureza cíclica da própria mineração de dados e ilustra que as lições aprendidas durante o processo de mineração de dados e da solução implantada podem desencadear novas questões de negócios, muitas vezes mais focadas" Shearer (2000, p 14).

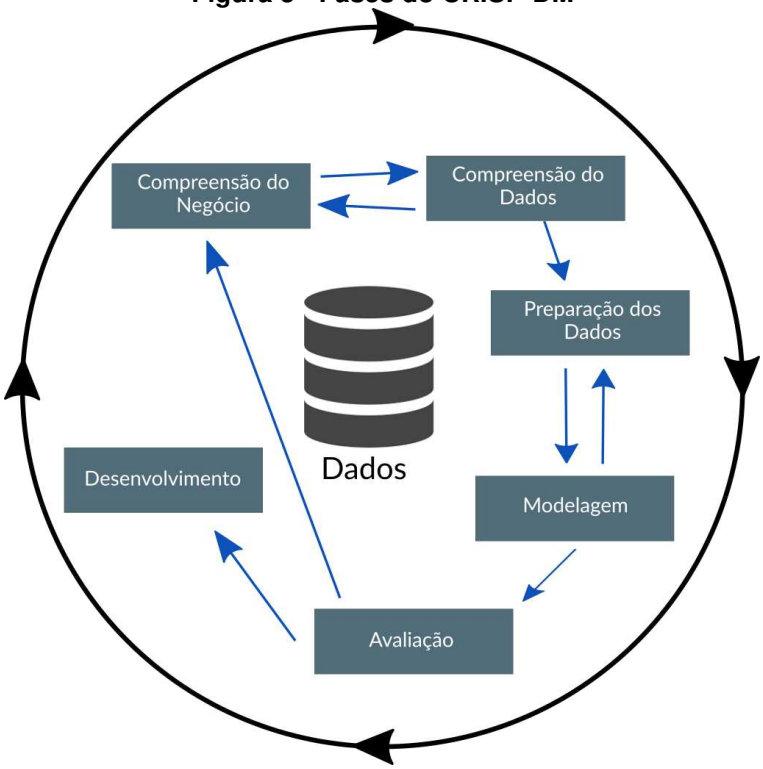
Shearer (2000) detalha as seis fases do modelo CRISP´-DM e as descreve como:

1. Compreensão do Negócio: A fase mais importante de qualquer projeto de mineração de dados, focadas na compreensão dos objetivos do projeto a partir de uma perspectiva de negócio, convertendo esse conhecimento torna-se a definição do problema de mineração de dados e cria um plano inicial projetado para atingir sua finalidade. As etapas incluem a determinação dos objetivos do negócio, a avaliação da situação, a determinação das metas de mineração de dados e a produção do plano do projeto.
2. Compreensão dos Dados: A fase de compreensão dos dados inicia com uma coleta inicial de dados. Para aumentar a familiaridade com os dados para identificar problemas de qualidade nos dados, obtendo uma visão inicial dos dados ou descobrir subconjuntos interessantes para formular hipóteses sobre informações ocultas.
3. Preparação dos Dados: A etapa de preparação dos dados engloba as atividades destinadas à criação do conjunto de dados definitivo, ou seja, os dados que serão utilizados como entrada nas ferramentas de modelagem a partir dos dados brutos iniciais. Essas atividades envolvem a seleção de tabelas, registros e atributos, além da transformação e limpeza dos dados para adequá-los às exigências das ferramentas de modelagem.
4. Modelagem: Nesta fase, são selecionadas e aplicadas várias técnicas de modelagem, e seus parâmetros são ajustados para obter valores otimizados. Como resultado, pode ser necessário retornar à etapa de preparação dos dados para realizar ajustes adicionais.
5. Avaliação: Antes de prosseguir para a implantação final do modelo construído, é importante avaliar mais detalhadamente o modelo e revisar sua construção para garantir que ele atinja adequadamente os objetivos de negócios, sendo crucial determinar se alguma questão comercial importante não foi suficientemente considerada. No final desta fase, o líder do projeto

deverá decidir exatamente como utilizar os resultados da mineração de dados. As principais etapas aqui são a avaliação dos resultados, a revisão do processo e a determinação dos próximos passos.

6. Implantação: Geralmente não é o fim do projeto, o conhecimento adquirido deve ser organizado e apresentado para que o cliente possa utilizá-lo. Dependendo dos requisitos, a fase de implantação pode ser tão simples como gerar um relatório ou tão complexa quanto implementar um processo de mineração de dados repetível em toda a empresa. É importante que o cliente entenda antecipadamente quais etapas devem ser executadas para realmente utilizar os modelos criados. As principais etapas são a implantação do plano, o monitoramento e a manutenção do plano, a produção do relatório final e a revisão do projeto.

Figura 3 - Fases do CRISP-DM



Fonte: Shearer (2000)

2.2.1 Artefatos das fases CRISP-DM

A comunicação eficaz dos resultados é essencial para manter todas as partes interessadas informadas e engajadas ao longo do projeto (Chapman et al. 2000). A documentação dos resultados não apenas registra o progresso e as descobertas alcançadas durante a execução do projeto, mas também fornece uma base sólida para a colaboração e a tomada de decisões informadas. Nesse sentido, os artefatos

em CRISP-DM desempenham um papel crucial. Conforme referenciado por Shearer (2000), esses artefatos, gerados em cada fase do processo de mineração de dados, são essenciais para garantir a transparência, a replicabilidade e a qualidade dos projetos de mineração de dados, além de facilitar a comunicação e a colaboração entre os membros da equipe e as partes interessadas no projeto.

No CRISP-DM, como descrito por Chapman et al. (2000) em seu manual *CRISP-DM 1.0 Step-by-step data mining guide*, possui os seguintes artefatos em suas seis fases distintas:

- Entendimento do Negócio:
  - Relatório de Contexto (*Background*)
  - Objetivos de Negócio e Critérios de Sucesso
  - Inventário de Recursos
  - Requisitos, Suposições e Restrições
  - Riscos e Contingências
  - Terminologia
  - Custos e Benefícios
  - Objetivos e Critérios de Sucesso em Mineração de Dados
  - Plano do Projeto
  - Avaliação Inicial de Ferramentas e Técnicas
- Entendimento dos Dados:
  - Relatório de Coleta Inicial de Dados
  - Relatório de Descrição dos Dados
  - Relatório de Exploração dos Dados
  - Relatório de Qualidade dos Dados
- Preparação dos Dados:
  - Relatório de Descrição do Conjunto de Dados
  - Relatório de Preparação dos Dados
- Modelagem:
  - Pressupostos de Modelagem
  - *Design* de Testes
  - Descrição do Modelo
  - Avaliação do Modelo
- Avaliação:

    - Avaliação dos Resultados de Mineração de Dados em relação aos Critérios de Sucesso de Negócio
    - Revisão do Processo
    - Lista de Possíveis Ações
- Implantação:
    - Plano de Implantação
    - Plano de Monitoramento e Manutenção
    - Relatório Final

Os documentos gerados ao longo das fases do CRISP-DM, conforme descritos por Chapman et al. (2000), estão detalhadamente apresentados no apêndice A deste Trabalho. Este apêndice não apenas contém os artefatos produzidos durante o processo de mineração de dados, mas também inclui instruções detalhadas sobre sua elaboração e propósito, conforme apresentado no manual de referência.

Esses documentos são essenciais não apenas para registrar o progresso e as descobertas alcançadas durante a execução do projeto, mas também para fornecer uma base sólida para a colaboração, a replicabilidade e a tomada de decisões informadas. Ao consultar este apêndice, os interessados terão acesso a uma documentação completa e transparente do processo de mineração de dados realizado, facilitando a compreensão e a revisão do trabalho realizado.

2.2.2 Conexão entre KDD e CRISP-DM

Ozçelik em sua publicação Process Models for Data Science Projects: CRISP-DM and KDD (2022) demonstra as principais diferenças entre os modelos, tendo como referência as fases dos modelos disponíveis na figura 4:

Quadro 1 - Comparação entre KDD e CRISP-DM

KDD	***CRISP-DM***
---	*Business Understanding*
Selection	*Data Understanding*
Pre-processing	
Transformation	*Data Preparation*
Data Mining	*Modeling*
Interpretation/Evaluation	*Evaluation*
---	*Deployment*

Fonte: Özçelik (2022), adaptação própria (2024)

Dentre as principais diferenças o autor destaca:

- CRISP-DM combina as etapas de Seleção e Pré-processamento na etapa de Compreensão de Dados.
- Os estágios do CRISP-DM são reversíveis. Dessa forma, quando um erro é cometido, é possível voltar atrás e corrigir o erro e fazer alterações sem completar todo o ciclo.
- O CRISP-DM difere do KDD com a fase de Entendimento de Negócios. Com a fase de Entendimento de Negócios, o CRISP-DM abrange todas as etapas da construção de um projeto confiável de ciência de dados.

Tendo conhecimento de ambas as metodologias de mineração de dados, fica evidente que o CRISP-DM seria um aperfeiçoamento visando produtos para cada fase de sua metodologia. Além disso, o CRISP-DM é mais flexível e adaptável a diferentes cenários de negócios. A sua estrutura cíclica permite que os cientistas de dados iterem sobre as etapas conforme necessário, tornando-o mais adequado para projetos de ciência de dados em constante evolução. Embora o KDD tenha sido pioneiro na formalização do processo de mineração de dados, o CRISP-DM aprimorou esse modelo ao adicionar uma camada de negócios e permitir maior flexibilidade no fluxo de trabalho.

No entanto, a escolha entre CRISP-DM e KDD dependerá em última análise das necessidades específicas do projeto e da organização. É importante lembrar que a melhor metodologia é aquela que se adapta bem ao problema em questão e fornece os resultados desejados.

2.2.3 Trabalhos que utilizaram CRISP-DM

Abaixo é possível encontrar alguns trabalhos recentes que utilizam esta metodologia, indo de acordo com a primeiro objetivo específico deste trabalho: Realizar uma revisão bibliográfica sobre artigos que utilizam o modelo de processos CRISP-DM.

Segundo Neto (2018) em seu artigo "O processo CRISP-DM aplicado na construção de uma solução para Análise de Risco de Crédito", demonstra como empresas utilizam dados para obter insights e melhorar serviços. Devido à quantidade de dados, a análise manual é inviável, o que requer a aplicação da Ciência dos Dados, que emprega técnicas estatísticas, computacionais e de aprendizado de máquina para sistematizar a análise. Neto (2018) detalha as etapas do processo CRISP-DM e sua

aplicação em problemas de classificação de risco de crédito, buscando altas taxas de aprovação com baixa inadimplência. Usando históricos de clientes para criar modelos preditivos, o estudo destaca a importância da preparação dos dados, incluindo limpeza e preenchimento de dados faltantes através de aprendizado de máquina. A técnica foi implementada em duas fases: avaliação do modelo preditivo e preenchimento de dados faltantes se o modelo fosse eficaz. Além disso, foram introduzidas técnicas não supervisionadas para identificar perfis de clientes e construir classificadores específicos, resultando em um índice KS de 31,7, considerado excelente para este contexto.

De acordo com Silva (2015) em sua publicação "Análise e mineração de dados dos cursos de pós-graduação do ensino à distância da UTFPR – Câmpus Medianeira", a vastidão de dados gerados pelo uso de sistemas de informação exige a implementação de sistemas adequados para a exploração desses dados. A mineração de dados serve a esse propósito, mas é essencial seguir uma metodologia completa e sistemática para garantir a credibilidade do conhecimento obtido. Diversas técnicas e métodos de mineração de dados foram experimentados e aplicados ao vasto volume de dados gerados pelos estudantes de pós-graduação na modalidade de ensino à distância da UTFPR, durante o período de agosto de 2014 a agosto de 2015, com o objetivo de prever a possível evasão de alunos que utilizam o sistema Moodle. Para esse caso específico, o algoritmo J48 foi identificado como o melhor classificador, alcançando uma precisão de 81,29% nas 1048 instâncias analisadas.

O estudo realizado por Peker e Kart (2023) em seu artigo "Transactional data-based customer segmentation applying CRISP-DM methodology: A systematic review", cujo acesso completo não foi possível por ser um material pago, o objetivo do estudo é fornecer uma revisão abrangente da literatura sobre segmentação de clientes baseada em dados transacionais. A revisão visa identificar diferentes características no campo, analisar a aplicação de técnicas de mineração de dados e destacar pontos importantes para pesquisas futuras. Foram utilizadas três grandes bases de dados online para revisar a literatura existente, resultando na seleção de 84 artigos relevantes publicados em periódicos de editoras renomadas. Esses artigos foram então completamente analisados com base nos diversos critérios das etapas da estrutura CRISP-DM, e os resultados foram relatados. Esta revisão sistemática é útil tanto para acadêmicos quanto para profissionais, oferecendo uma visão abrangente

do trabalho de pesquisa sobre segmentação de clientes utilizando mineração de dados e apresentando diretrizes para pesquisas futuras na área.

Na visão de Plotnikova, Dumas e Milani (2021) em seu artigo "Adapting the CRISP-DM Data Mining Process: A Case Study in the Financial Services Domain", no qual a consulta completa foi inviabilizada devido ser um material pago, relata um caso em uma organização de serviços financeiros com o objetivo de identificar lacunas percebidas no processo CRISP-DM e como ele é adaptado para abordar essas lacunas. O estudo de caso baseou-se na documentação de um portfólio de projetos de mineração de dados, complementada por entrevistas semiestruturadas com participantes do projeto. Os resultados identificaram 18 lacunas no CRISP-DM, além do impacto percebido e os mecanismos empregados para resolvê-las, agrupando essas lacunas em seis categorias. O estudo fornece aos profissionais um conjunto estruturado de lacunas a serem consideradas ao aplicar o CRISP-DM ou processos similares em serviços financeiros. Algumas dessas lacunas são genéricas e aplicáveis a outros setores com preocupações semelhantes, como telecomunicações e comércio eletrônico.

2.3 Diabetes Mellitus

O diabetes - Diabetes Mellitus, caracterizado por altos níveis de glicose no sangue devido à baixa secreção de insulina para a quebra de açúcar dos alimentos, representa atualmente um dos principais problemas de saúde, quanto quantidade de pessoas, afetando em incapacidade e em outros casos mortalidade, quanto também em um investimento governamental elevado para o controle e tratamento de complicações, segundo Brunner e Suddarth (2006) e já é a quarta maior causa de morte no Brasil, conforme referenciado por Pace e Nunes (2006).

As complicações decorrentes da progressão da doença, como alterações na intolerância à glicose sem diagnóstico e tratamento adequado, destacam a importância do diagnóstico precoce, implementação de tratamento, e mudanças de hábitos alimentares e estilo de vida saudável, como observado por Filho Rac et al. (2002).

Complementando, Michelutti (2006) ressalta que metade das pessoas com diabetes desconhecem sua condição até que sinais de complicações se manifestem. A escolha de abordar a diabetes surge em resposta ao alarmante aumento de casos.

Em 2021, a Fiocruz registrou 17 milhões de diagnósticos de diabetes no Brasil, colocando-o como o quinto país com o maior número de casos. Globalmente, a OMS relatou cerca de 422 milhões de casos de diabetes, resultando em 1,5 milhão de mortes anuais relacionadas à doença (OMS, 2023).

Utilizando como referência o site Kaggle (2024) que é definido como “a maior comunidade de ciência de dados do mundo, com ferramentas e recursos poderosos para ajudá-lo a atingir seus objetivos de ciência de dados”, obtemos um total de 1,424 Databases de Diabetes disponíveis para uso. Alguns destaques são:

Quadro 2 - Base de Dados Kaggle

Base de Dados	Votos da comunidade
Pima Indians Diabetes Database	4096
Heart Failure Prediction Dataset	2508
Diabetes Dataset	1125

Fonte: Autoria própria (2024)

2.4 Outras Metodologias para Mineração de Dados

Nos tópicos abaixo são apresentadas outras metodologias e ferramentas existentes que podem ser utilizadas para a mineração de dados, acompanhado da explicação de como funcionam alguns dos classificadores que são utilizados para essa finalidade.

2.4.1 SEMMA

De acordo com o SAS Institute (2017), o processo de mineração de dados é aplicável em uma variedade de setores e fornece metodologias para diversos problemas de negócios, como detecção de fraude, *marketing* de banco de dados, análise de risco, satisfação do cliente dentre outros, sendo destacado:

> O processo de Amostragem, Exploração, Modificação, Modelagem e Avaliação (SEMMA) de grandes quantidades de dados para descobrir padrões desconhecidos que podem ser utilizados como vantagem comercial (SAS Institute Inc., 2017).

O processo SEMMA, complementado por Santos e Azevedo (2005), fornece uma metodologia simples para compreensão do processo, permitindo o desenvolvimento e manutenção de projetos sistemáticos e apropriados de mineração de dados, fornecendo uma estrutura para a concepção, criação e evolução de

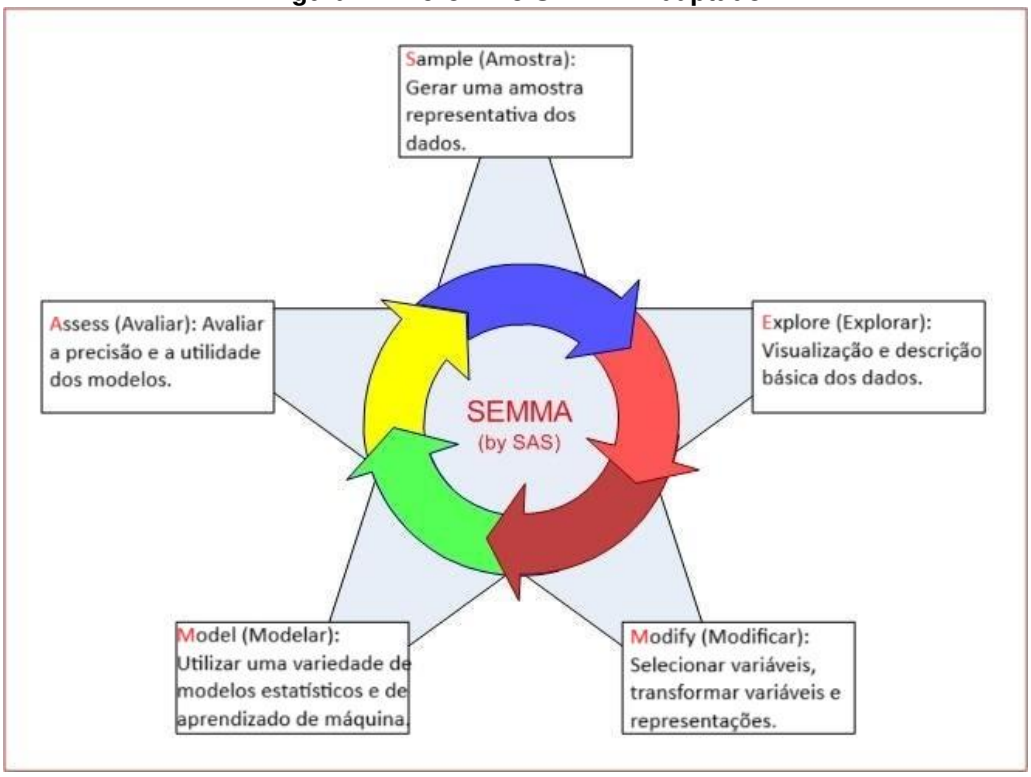
soluções, podendo ajudar a resolver problemas de negócios e também a atingir seus objetivos, mesmo sendo independente da ferramenta escolhida para mineração de dados, ele está vinculado ao SAS Enterprise Software Miner.

De acordo com o SAS Institute (2017) a metodologia SEMMA é composta considerando um ciclo com 5 etapas para o processo:

1. Amostragem – Demonstrar dados extraindo uma parte de um grande conjunto de dados, visando informações significativas, mas pequenas o suficiente para serem manipuladas rapidamente.
2. Explorar – Explorar dados para descobrir tendências e anomalias imprevistas para obter compreensão e *insights*.
3. Modificar – Modificação dos dados criando, selecionando e transformando variáveis para centralizar o processo de seleção do modelo.
4. Modelar – Envolve a modelagem dos dados para permitir que o software procure automaticamente combinações de dados que prevejam com segurança os resultados desejados.
5. Avaliar – Avaliar os dados e estimar o seu desempenho através da utilidade e confiabilidade dos resultados do processo de mineração de dados.

A metodologia SEMMA pode ser mais compreensível a partir da figura 5 – Explicação SEMMA:

Figura 4 – Acrônimo SEMMA Adaptado



Fonte: Vasconcellos (2017), adaptação própria (2024)

O SAS Institute (2017) também concluiu, enfatizando que pode ou não incluir todas as etapas do SEMMA em sua análise, e que uma ou mais etapas podem precisar ser repetidas múltiplas vezes para obter resultados satisfatórios.

2.4.2 TDSP

O Processo de Ciência de Dados de Equipe (TDSP), desenvolvido pela Microsoft, é uma abordagem ágil e iterativa para conduzir projetos de ciência de dados, comumente utilizado em conjunto com a plataforma *Azure Machine Learning*.

O TDSP visa melhorar a colaboração da equipe e oferecer soluções eficientes de análise preditiva, incorporando práticas recomendadas da Microsoft e de outros líderes do setor, além de poder ser adaptado a outros processos de ciência de dados, como CRISP-DM e KDD.

As cinco fases principais do TDSP são as seguintes:

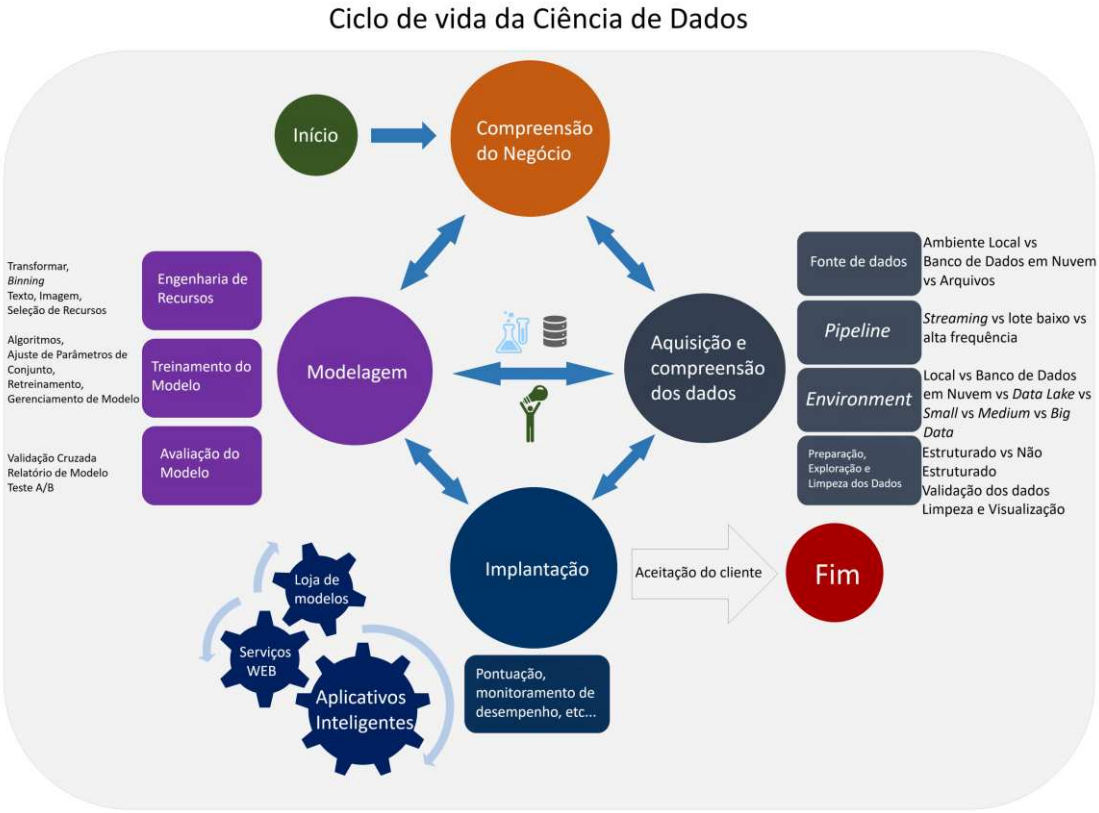
- Definição do Negócio: Nesta fase inicial, são estabelecidos os objetivos do projeto, as métricas de sucesso e são identificadas as fontes de dados relevantes. É crucial criar metas SMART (Específicas, Mensuráveis, Alcançáveis, Relevantes e com Tempo Definido) e garantir a precisão e relevância das fontes de dados selecionadas. Documentos como o

"Documento de Estatuto Dinâmico" e dicionários de dados detalhados são entregues nessa fase.

- Aquisição e Compreensão dos Dados: Aqui, os dados são preparados para análise e modelagem. Isso inclui atividades como a ingestão de dados nos ambientes de análise, exploração dos dados para determinar sua adequação para análise e configuração de um pipeline de dados para atualização regular e pontuação dos dados. Relatórios de qualidade de dados e arquiteturas de solução são entregues nessa fase.
- Modelagem: O foco nesta fase está no desenvolvimento de modelos de aprendizado de máquina precisos. Isso envolve a engenharia de recursos, o treinamento de modelos com diversos algoritmos e a avaliação da eficácia dos modelos. Um cientista de dados da equipe avalia os modelos quanto à sua adequação para produção, questionando sua confiança e considerando dados de teste.
- Implantação: Aqui, os modelos desenvolvidos são implementados em um ambiente de produção. Isso inclui a operacionalização dos modelos e *pipelines* de dados, bem como a integração de monitoramento para garantir que os modelos implantados estejam funcionando conforme o esperado. São entregues um painel de *status* e um relatório de modelagem final com detalhes da implantação nessa fase.
- Aceitação do Cliente: A fase final envolve garantir que o sistema implantado atenda aos objetivos do cliente. Isso inclui a validação do sistema para garantir sua precisão e implantação para uso. Um relatório de saída do projeto é fornecido ao cliente, detalhando aspectos técnicos do projeto para operação do sistema.

Seu ciclo de vida é definido como:

Figura 5 – Ciclo de vida do TDSP



Fonte: Microsoft (2024), adaptação própria (2024)

Cada uma dessas fases é iterativa, permitindo a revisão e o refinamento contínuos à medida que novos insights ou desafios surgem. Além disso, o TDSP enfatiza a importância da comunicação e colaboração eficazes dentro da equipe durante todas as fases do projeto. Em relação a estrutura é padronizada para organizar os artefatos do projeto, incluindo código, dados e documentação. Isso facilita a colaboração dentro da equipe e permite que os membros da equipe compartilhem facilmente seu trabalho uns com os outros, além disso, recomenda várias ferramentas e utilitários da Microsoft para apoiar cada fase do processo.

2.5 Classificadores

As tarefas são categorizadas por suas possíveis realizações, que consistem no que procurar nos dados ou em categorias de padrões com base na área de interesse (Larose 2005, apud JUNIOR, RODRIGUEZ 2022), em sua publicação *Conceitos de Mineração de dados na Web*, Santos (2009) descreve as técnicas básicas de Mineração de dados como:

- Classificação: Consiste na previsão de uma classe para um dado, com base em atributos previamente definidos. O algoritmo de classificação aprende com exemplos de dados já classificados, buscando identificar padrões que determinam a classe. É necessário um conjunto adequado de dados completos para cada classe considerada.
- Regressão: Similar à classificação, mas visa prever um valor numérico real em vez de uma classe. Por exemplo, pode ser usado para atribuir uma nota a um filme com base em seus atributos. Requer exemplos de dados com valores numéricos associados aos atributos.
- Agrupamento ou *Clustering*: Identifica grupos naturais de dados que compartilham similaridades entre si. Não requer classes ou valores predefinidos, mas sim algoritmos que formam grupos com base em métricas de similaridade. A maioria dos algoritmos considera apenas atributos numéricos, mas há extensões que lidam com dados não numéricos.
- Sumarização: Técnicas para identificar descrições concisas e compreensíveis dos dados ou de um subconjunto deles. A precisão não é necessariamente o foco; o objetivo é descrever os dados de forma inteligível. Pode ser feita através de regras, como no exemplo de documentos classificados como "alto" ou "médio" com base em atributos específicos.
- Modelagem de dependência: Identifica modelos que descrevem relações significativas entre valores de atributos ou entre valores nos dados. Inclui técnicas como busca de regras de associação, frequentemente assumindo atributos discretos.
- Detecção de mudança ou outliers: Identifica dados que não se ajustam a um modelo aceitável dos dados ou que mostram mudanças inesperadas.

Essas tarefas possuem classificadores próprios como:

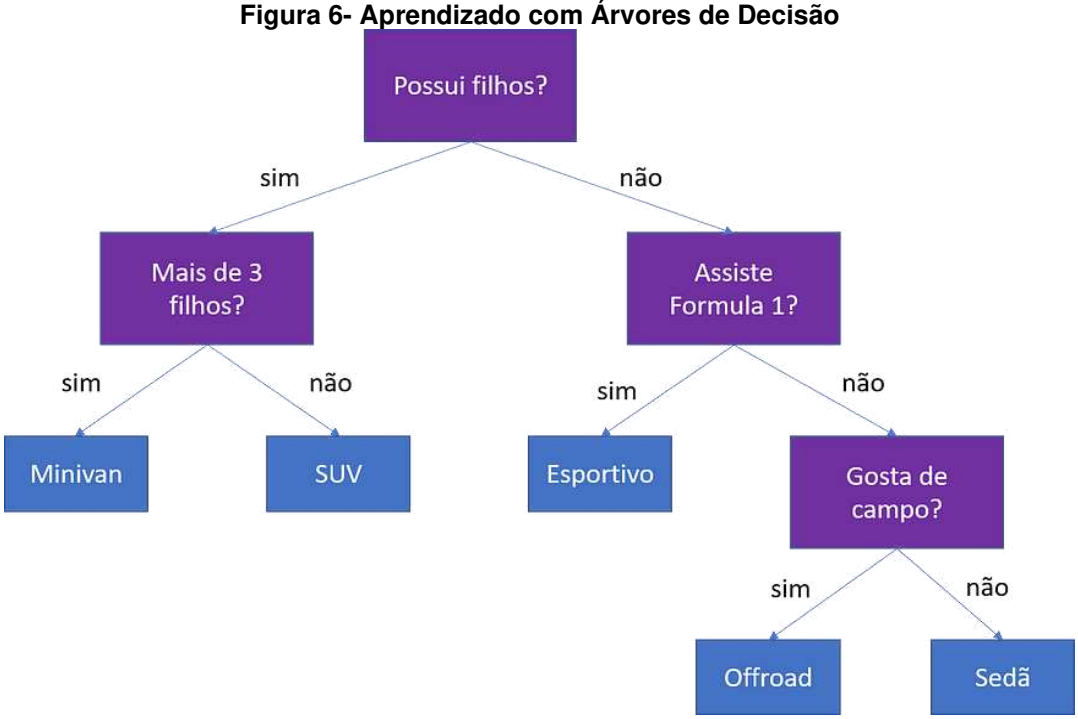
- Algoritmo OneR e Zero R: Saed Sayyad descreve os algoritmos de classificação como:
- OneR, abreviação de "One Rule", é um algoritmo de classificação simples, mas preciso, que gera uma regra para cada preditor nos dados e, em seguida, seleciona a regra com o menor erro total como sua "regra única".

Para criar uma regra para um preditor, construímos uma tabela de frequência para cada preditor em relação ao alvo. Foi demonstrado que o OneR produz regras apenas um pouco menos precisas do que os algoritmos de classificação de última geração, ao mesmo tempo que produz regras que são simples de serem interpretadas pelos humanos.

  - Para cada preditor,
  - Para cada valor desse preditor, faça uma regra como segue;
  - Conte com que frequência cada valor do alvo (classe) aparece
  - Encontre a aula mais frequente
  - Faça a regra atribuir essa classe a este valor do preditor
  - Calcule o erro total das regras de cada preditor
  - Escolha o preditor com o menor erro total.

- Zero R é o método de classificação mais simples que depende do alvo e ignora todos os preditores. O classificador ZeroR simplesmente prevê a categoria majoritária (classe). Embora não haja poder de previsibilidade no ZeroR, ele é útil para determinar um desempenho de linha de base como referência para outros métodos de classificação.
  - Construa uma tabela de frequência para o alvo e selecione seu valor mais frequente.
- Algoritmo de árvore de decisão: As árvores de decisão são uma técnica popular e surpreendentemente eficaz, especialmente para problemas de classificação (Vikas, 2024). São estruturas gráficas que representam decisões e seus resultados, semelhantes a fluxogramas. Cada nó na árvore representa uma pergunta ou condição, e cada ramo representa uma possível resposta ou resultado. O algoritmo C4.5, como discutido por Sumit Saha em seu artigo "*What is the* C4.5 *algorithm and how does it work*?", descreve como um método para montar de árvores de decisão. O algoritmo C4.5 baseia-se em dois princípios fundamentais: ganho de informação e entropia. O ganho de informação mede a redução na incerteza da classificação de um conjunto de dados após a divisão por uma determinada característica. Por outro lado, a entropia mede a incerteza inicial antes de fazer uma pergunta específica. Quando aplicado a conjuntos de dados, o algoritmo C4.5 busca maximizar o ganho de informação ao escolher as perguntas que resultam na maior redução da entropia.

Figura 6- Aprendizado com Árvores de Decisão



Fonte: Ricardo Araujo (2020)

- *Naive Bayes*: De acordo com Ray (2017), o algoritmo *Naive Bayes* é uma técnica de classificação baseada no Teorema de Bayes, que assume independência entre os preditores. É amplamente utilizado em problemas de classificação, como filtragem de *spam*, classificação de texto e sistemas de recomendação. *Naive Bayes* pertence a uma família de algoritmos de aprendizado generativo e é conhecido por sua eficiência e boa performance em diversas aplicações do mundo real.
- *Multilayer Perceptron*: De acordo com a publicação "*Multilayer Perceptron*" do *site* deepai (DEEPAI), o *perceptron* é uma ferramenta valiosa para classificar conjuntos de dados que são linearmente separáveis. No entanto, ele encontra limitações sérias quando os dados não seguem esse padrão, revelando que, para qualquer classificação de quatro pontos, existe um conjunto que não é linearmente separável. O *Multilayer Perceptron* (MLP) supera essa restrição, sendo capaz de classificar conjuntos de dados que não são linearmente separáveis. Ele alcança isso por meio de uma arquitetura mais robusta e complexa para aprender modelos de regressão e classificação para conjuntos de dados difíceis. O *Perceptron* consiste em uma camada de entrada e uma camada de saída, totalmente conectadas. Já

os MLPs possuem as mesmas camadas de entrada e saída, mas podem ter várias camadas ocultas entre elas.

- *Random Forest*: De acordo com a publicação sobre *Random Forest* no *site* da IBM, o *Random Forest* é um método de aprendizado de máquina que é eficaz tanto para problemas de classificação quanto para problemas de regressão. Ele opera construindo uma "floresta" de árvores de decisão durante o treinamento. Cada árvore na floresta é treinada de forma independente usando uma amostra aleatória do conjunto de dados, e a decisão final é tomada com base na votação das árvores individuais (no caso de classificação) ou na média das previsões (no caso de regressão). O *Random Forest* oferece várias vantagens, incluindo robustez contra *overfitting*, capacidade de lidar com conjuntos de dados grandes com muitas variáveis de entrada e capacidade de lidar com dados faltantes sem a necessidade de imputação. Além disso, ele pode fornecer uma estimativa da importância relativa das variáveis de entrada no processo de tomada de decisão.
- ADA BOOST: O algoritmo AdaBoost, abreviação de *Adaptive Boosting*, é uma técnica de *Boosting* usada como Método de *Ensemble* em Aprendizado de Máquina. Ele é chamado de *Adaptive Boosting* porque os pesos são reatribuídos a cada instância, com pesos mais altos atribuídos às instâncias classificadas incorretamente. O que esse algoritmo faz é construir um modelo e atribuir pesos iguais a todos os pontos de dados. Em seguida, atribui pesos mais altos aos pontos que são classificados incorretamente. Agora, todos os pontos com pesos mais altos recebem mais importância no próximo modelo. Ele continuará treinando modelos até obter um erro menor. Ao construir diferentes modelos no mesmo conjunto de dados, observamos variações na precisão. No entanto, aproveitando o poder do AdaBoost, podemos combinar esses algoritmos para aprimorar as previsões finais. Ao calcular a média dos resultados de modelos diversos, o AdaBoost nos permite alcançar uma precisão mais alta e reforçar as capacidades preditivas de forma eficaz.

2.6 Ferramentas para Mineração de Dados

Neste tópico é apresentado dois *softwares* que são utilizados para a mineração de dados, o IBM SPSS sendo uma opção paga e o Weka sendo um *software* gratuito e de código aberto.

2.6.1 SPSS

O IBM *Statistical Package for the Social Sciences* (SPSS) representa uma ferramenta crucial no campo da pesquisa quantitativa desde sua concepção em 1968 por Norman Nie, C. Hadlai Hull e Dale H. Bent. Amplamente utilizado por pesquisadores em diversas disciplinas, o SPSS facilita a coleta, análise e interpretação de dados. Sua *interface* intuitiva e extenso conjunto de módulos o tornam acessível tanto para profissionais das ciências humanas quanto das exatas. No entanto, é imperativo que os usuários possuam conhecimentos prévios em estatística descritiva e inferencial para explorar plenamente suas funcionalidades (Santos, 2018).

O SPSS é um pacote estatístico abrangente que permite não apenas a manipulação e análise de dados, mas também a preparação e validação deles. Compatível com diversos outros programas, como Excel, SAS e Stata, o SPSS facilita a abertura de arquivos sem a necessidade de conversões intermediárias. Seus principais recursos incluem a gestão de uma vasta quantidade de dados, a criação e modificação de variáveis, a realização de diversos tipos de análises estatísticas, como análise de variância e regressão, e a construção de gráficos representativos (IBM, 2016).

Segundo Santos (2018) a *interface* do SPSS é composta principalmente pelo SPSS *Data Editor*, onde os dados são introduzidos e manipulados. Nesta janela, cada coluna representa uma variável, enquanto as linhas correspondem aos casos ou indivíduos. Embora o SPSS seja relativamente autoexplicativo, oferecendo assistência por meio de várias opções de *menu*, os usuários podem acessar uma ampla variedade de análises estatísticas, como média, mediana, moda, desvio padrão e regressão linear, através da opção "estatística". Além disso, o SPSS oferece uma avaliação gratuita por meio de uma senha *trial* disponível no *site* da IBM, juntamente com suporte ao produto, manuais e tutoriais em vídeo para facilitar o entendimento da ferramenta.

2.6.2 Weka

O W*aikato Environment for Knowledge Analysis* (Weka) é uma *suite* de *software* de Mineração de Dados desenvolvida pela *University of Waikato*. Implementado em Java, é conhecido por sua portabilidade e orientação a objetos. O WEKA oferece uma variedade de algoritmos para diferentes tarefas de Mineração de Dados, incluindo classificação, predição numérica, agrupamento e associação. Alguns dos métodos incluem árvores de decisão, Naive Bayes, regressão linear, algoritmos de agrupamento como EM e SimpleKMeans, e métodos de associação como Apriori e FPGrowth. (Damasceno [s. d.])

Menotti (2017) destaca que o nome é inspirado na ave Weka da Nova Zelândia, que é curiosa e não voa, sendo um *software* de código aberto e é distribuído sob a GNU *General Public License*.

Devido sua gama de ferramentas para mineração de dados, ser de código aberto e disponibilidade de manuais *online* está foi a escolha de ferramenta para utilização nesse trabalho.

3 METODOLOGIA

A metodologia adotada neste estudo seguirá o modelo de processos CRISP-DM devido às suas características orientadas a processos e sua capacidade de proporcionar uma documentação abrangente, sendo aplicado em seis fases como descrito por SHEARER (2000).

3.1 Modelo de Processos CRISP-DM

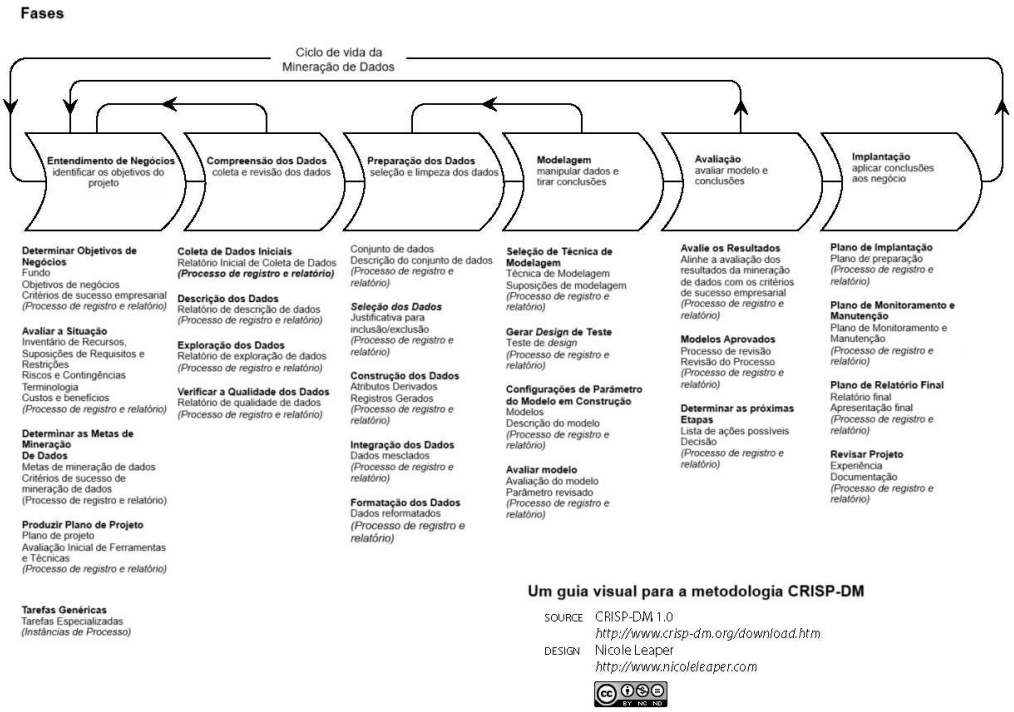
Busca-se seguir os passos do guia *Introdução ao CRISP-DM* (IBM, 2023) a fim de documentar as saídas existentes:

- Compreensão do Negócio: Na fase inicial, busca-se compreender profundamente os problemas, objetivos e requisitos de negócios. Neste primeiro momento, o foco está em definir claramente o problema de Ciência dos Dados a ser solucionado, onde a qualidade da definição desse problema é fundamental para o sucesso da solução.
    - Determinar os objetivos do negócio.
    - Avaliar a situação atual.
    - Estabelecer metas para o projeto de mineração de dados.
    - Definir critérios de sucesso.
- Compreensão dos Dados: Avaliar a qualidade dos dados, explorar seu conteúdo e começar a formular hipóteses iniciais, que incluem identificar bases de dados públicas relevantes na área de diabetes, realizar uma análise exploratória dos dados para compreender suas características, distribuições, valores faltantes etc.
    - Coletar dados iniciais.
    - Descrever os dados.
    - Explorar os dados para identificar padrões preliminares.
    - Verificar a qualidade dos dados e identificar problemas.
- Preparação dos Dados: Realizar a limpeza dos dados, tratando valores ausentes e inconsistências, realizar transformações nos dados, como definir etapas de pré-processamento, normalização de valores numéricos e opções de teste.
    - Selecionar dados relevantes.
    - Limpar os dados para corrigir erros e preencher lacunas.

    - Integrar dados de várias fontes, se necessário.
    - Transformar os dados em um formato adequado para análise.
- Modelagem: Selecionar diferentes técnicas de classificação encontradas, como Árvores de Decisão, Redes Neurais, K-NN, etc. Treinar os modelos com os dados de treinamento, ajustar seus parâmetros e avaliar o desempenho dos modelos utilizando métricas apropriadas.
    - Selecionar técnicas de modelagem.
    - Criar conjuntos de dados de treinamento e teste.
    - Construir e calibrar modelos.
    - Avaliar modelos e selecionar o melhor.
- Avaliação: Apresentação das métricas de avaliação, como precisão, estatística Kappa, matriz de confusão e comparação do desempenho dos diferentes modelos testados. Discutir as vantagens e limitações do CRISP-DM.
    - Avaliar os resultados em relação aos objetivos de negócios.
    - Revisar o processo e os passos tomados.
    - Determinar se os objetivos foram alcançados.
    - Identificar próximas etapas e recomendações.
- Implantação: Apresentar as conclusões do estudo, destacando a contribuição do CRISP-DM na construção de classificadores para a previsão de diabetes, discutir a documentação gerada ao longo do processo e sua utilidade para futuras pesquisas na área.
    - Planejar a implantação do modelo.
    - Preparar documentação e recursos necessários.
    - Implementar o modelo em ambiente de produção.
    - Monitorar o desempenho do modelo em produção e fazer ajustes conforme necessário.

As fases do modelo proposto por Chapman (2000) possuem tarefas e subtarefas a serem realizadas no decorrer das fases como pode se consultado a partir da Figura (8) abaixo:

Figura 7 – Um guia visual para a metodologia CRISP-DM



Fonte: Nicole Leaper (2009), adaptação própria (2024)

Para obter uma compreensão mais aprofundada da implementação prática do modelo CRISP-DM, incluindo detalhes sobre ferramentas e técnicas específicas utilizadas em cada fase, os leitores podem consultar o Apêndice A, onde estão disponíveis documentos e recursos adicionais.

4 DESENVOLVIMENTO

Este estudo segue as etapas do Modelo de Processos CRISP-DM conforme as diretrizes do Guia IBM SPSS *Modeler* (IBM), com um enfoque específico na análise de dados utilizando bases de dados públicas sobre Diabetes. Ao final de cada etapa, o Guia IBM SPSS Modeler (IBM) prevê uma série de perguntas a serem respondidas, a qual será dado ênfase no final de cada fase.

4.1 Compreensão de negócios

Inicialmente deve-se compreender o negócio para definir quais as necessidades do projeto e seus objetivos. A partir do contexto, são definidas as metas, critérios de sucesso e recursos disponíveis para a elaboração dos modelos.

4.1.1 Contexto

O principal objetivo é não apenas implementar os classificadores, mas também destacar a aplicação eficaz do Modelo CRISP-DM no contexto de Análise de Dados.

A seleção das bases de dados foi realizada após uma pesquisa de artigos científicos relacionados à utilização de classificadores para predição da doença Diabetes. Duas bases de dados foram escolhidas:

- *Early-Stage Diabetes Risk Prediction Dataset* (Faniqul, 2020).
    - O conjunto de dados foi alimentado originalmente por meio de questionários diretos de pacientes de um Hospital em Sylhet, Bangladesh.
- *Pima Indians Diabetes Database* (UCI MACHINE LEARNING, 2016).
    - A base de dados é originalmente do Instituto Nacional de Diabetes e Doenças Digestivas e Renais da Índia. Todos os pacientes são mulheres com pelo menos 21 anos de idade, de ascendência indígena “Pima”.

4.1.2 Objetivos de negócio e critérios de sucesso

Os objetivos deste projeto não se trata apenas de construir classificadores eficazes para predição da doença Diabetes, mas também documentar cada etapa do processo utilizando o Modelo CRISP-DM. Além disso, busca-se demonstrar como a

adoção desse modelo pode influenciar positivamente o processo de Mineração de Dados. Para garantir o sucesso do projeto, foram estabelecidos os seguintes critérios:

- Contribuição para a adoção de um Modelo de Processos:
    - Avaliar a contribuição do estudo na promoção da adoção do Modelo de Processos, demonstrando como ele pode beneficiar a orientação dos processos de descoberta do conhecimento e mineração de dados.
- Documentação do Processo de Mineração de Dados:
    - Verificar a qualidade e abrangência da documentação do processo de Mineração de Dados, garantindo que todas as etapas das 6 fases do CRISP-DM estejam claramente registradas e compreensíveis.
- Construção de Classificadores:
    - Avaliar o desempenho dos classificadores construídos, considerando métricas como precisão, *recall* e F1-*score.*
    - Garantir que os classificadores atinjam um nível satisfatório de desempenho para a tarefa de classificação da doença Diabetes.
- Utilização de Bases de Dados Públicas:
    - Demonstrar a decisão de utilizar apenas bases de dados públicas relacionadas à doença Diabetes, destacando a disponibilidade pública e a transparência dos dados utilizados no estudo.
- Contribuição para Futuras Pesquisas e Projetos:
    - Avaliar o potencial do estudo como recurso para futuras pesquisas e projetos relacionados à aplicação do modelo de processos CRISP-DM na classificação de dados.
    - Verificar a significativa contribuição para o conhecimento do modelo de processos, demonstrando sua capacidade de auxiliar e organizar os processos de Mineração e Análise de Dados.

4.1.3 Inventário de recursos

Nesta fase, é essencial realizar um inventário completo dos recursos utilizados no projeto. Isso inclui acesso às bases de dados públicas sobre diabetes, literatura relevante e ferramentas de mineração de dados. Optou-se pelo *software* Weka devido à sua vasta coleção de algoritmos de aprendizado de máquina, ser

código aberto e gratuito com funcionalidades para análise de dados. O Weka oferece ferramentas abrangentes para preparação de dados, classificação, regressão, agrupamento, regras de associação e visualização (Menotti, 2017).

As bases de dados selecionadas para aplicação do estudo junto com o Weka são as denominadas *Early-Stage Diabetes Risk Prediction Dataset* (Faniqul, 2020) e *Pima Indians Diabetes Database* (UCI MACHINE LEARNING, 2016).

- Recursos de *Hardware*:
    - O *hardware* proposto para suportar o projeto é um servidor composto por um processador Ryzen 5 2600, que oferece 6 núcleos e 12 *threads*, junto de 16 GB de memória RAM operando a 3200MHz, possuindo também um SSD NVME de 2300 mb/s de leitura. O servidor disponibilidade se mostrou mais do que adequado para suportar as atividades de análise de dados do projeto, proporcionando um ambiente de trabalho estável e eficiente para a equipe.
- Fontes de Dados e Depósitos de Conhecimento:
    - Identificar todas as fontes de dados disponíveis para a mineração, incluindo tipos e formatos de dados.
    - Avaliar a necessidade de adquirir mais dados externos, como informações demográficas, e identificar possíveis problemas de segurança que possam impedir o acesso aos dados necessários.
- Recursos da Equipe:
    - Equipe composta por dois analistas de dados, João Paulo Pereira Santa Clara e Murillo Iago Moreira, ambos estudantes do curso de Análise e Desenvolvimento de Sistemas da Universidade Tecnológica Federal do Paraná.

4.1.4 Requisitos, suposições e restrições

- Requisitos do projeto:
    - Definição dos Dados: As duas bases de dados selecionadas para aplicação do estudo junto com o Weka foram a *Early-Stage Diabetes Risk Prediction Dataset* (Faniqul, 2020) e *Pima Indians Diabetes Database* (UCI MACHINE LEARNING, 2016).
    - Escolha do *Software* e Técnicas de Classificação: O *Software* Weka foi escolhido como a principal ferramenta para conduzir as tarefas de

pré-processamento e modelagem. Será necessário a avaliação no decorrer do desenvolvimento do projeto sobre as capacidades do Weka em relação aos requisitos específicos do projeto.

- Suposições
    - Base de Dados *Pima Indians Diabetes*: Apesar de a base de dados oferecer uma ampla gama de variáveis preditoras médicas e uma variável alvo bem definida, não foram identificadas suposições específicas sobre os dados até o momento.
    - Base de Dados *Early Stage Diabetes Risk Prediction*: Até o momento, não há suposições específicas sobre os dados desta base. Uma exploração mais aprofundada será necessária para entender completamente sua estrutura e características.
- Restrições
    - Utilização Exclusiva do *Software* Weka: Inicialmente, o projeto está restrito à utilização do *Software* Weka para as tarefas de pré-processamento e modelagem. Esta decisão é fundamentada na necessidade de garantir consistência e eficácia nas análises realizadas. Durante a etapa de "Avaliação inicial de ferramentas e técnicas", será dada uma atenção especial à avaliação da adequação do Weka para atender aos requisitos do projeto.

4.1.5 Riscos e Planos de Contingência

Apesar de não identificarmos riscos sociais e econômicos diretos, reconhecemos que todo projeto carrega consigo fatores de risco inerentes. É fundamental que estejamos preparados para enfrentar e mitigar tais eventualidades.

- Riscos:
    - Impossibilidade de finalizar o estudo dentro do prazo estabelecido;
    - Integridade dos dados não condizentes com o esperado;
    - Indisponibilidade dos dados;
    - Possibilidade de os resultados do projeto não atenderem às expectativas iniciais.

- Planos de Contingência
    - Impossibilidade de finalizar o estudo dentro do prazo estabelecido:
        - Avaliar o progresso do projeto regularmente para identificar qualquer desvio em relação ao cronograma.
        - Se necessário, reavaliar o cronograma e priorizar as tarefas para garantir a conclusão dentro do prazo.
        - Comunicar quaisquer atrasos e avaliar uma extensão do prazo em último caso.
    - Integridade dos dados não condizentes com o esperado:
        - Realizar uma análise detalhada da qualidade dos dados assim que forem adquiridos.
        - Se os dados não atenderem aos critérios de qualidade esperados, considerar a seleção de bases de dados alternativas.
        - Implementar medidas de limpeza e pré-processamento adicionais, se necessário, para melhorar a qualidade dos dados.
    - Indisponibilidade dos dados:
        - Identificar possíveis fontes alternativas de dados que possam ser utilizadas em caso de indisponibilidade das que foram propostas originalmente como públicas.
        - Considerar procedimentos para lidar com interrupções inesperadas no acesso aos dados, como *backups* regulares e priorizar armazenamento local.
    - Possibilidade de os resultados do projeto não atenderem às expectativas iniciais:
        - Se os resultados preliminares não atenderem às expectativas, revisar a abordagem metodológica e considerar a seleção de outros algoritmos de classificação, outras bases de dados ou outras ferramentas de mineração de dados.

4.1.6 Terminologia

- CRISP-DM (*Cross-Industry Standard Process for Data Mining*): Um modelo de processo amplamente utilizado para guiar projetos de mineração de

dados, dividido em seis fases: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação.

- Modelo de Processos: Refere-se a uma estrutura organizada de atividades ou etapas a serem seguidas para atingir um objetivo específico, como o CRISP-DM para projetos de mineração de dados.
- Mineração de Dados: O processo de descoberta de padrões, tendências e informações úteis em conjuntos de dados grandes e complexos, geralmente com o uso de técnicas de análise estatística e de aprendizado de máquina.
- Precisão, *Recall*, *F1-score*, ROC Area: Métricas comuns de desempenho para avaliar a eficácia de classificadores em problemas de aprendizado supervisionado.
- *Pima Indians Diabetes Database*: Uma base de dados amplamente utilizada para pesquisa em diabetes, contendo informações médicas de pacientes, como número de gestações, IMC, nível de insulina, idade, entre outros, e uma variável alvo indicando o diagnóstico de diabetes.
- *Early-Stage Diabetes Risk Prediction Dataset*: Outra base de dados relacionada à previsão de risco de diabetes, embora não haja informações específicas fornecidas sobre suas características ou variáveis.
- *Software* Weka: Uma plataforma de *software* amplamente utilizada para mineração de dados e aprendizado de máquina, conhecida por sua vasta coleção de algoritmos de classificação, regressão, agrupamento e outras funcionalidades para análise de dados.
- Contingências: Planos ou medidas preparadas para lidar com possíveis riscos, eventos adversos ou imprevistos que possam surgir ao longo do projeto, visando garantir sua continuidade e sucesso.

4.1.7 Custos e benefícios

- Custos estimados:
  - Tempo e recursos de pessoal: Avaliar o tempo dedicado pela equipe ao projeto e os custos associados aos recursos humanos envolvidos.
- Benefícios previstos:

- Demonstração da aplicabilidade do modelo CRISP-DM: Avaliar como a implementação do modelo CRISP-DM pode impactar positivamente o processo de mineração de dados.
- Complemento no processo de mineração de dados na área da saúde: Identificar como o uso do CRISP-DM pode melhorar a eficiência e eficácia da análise de dados na área da saúde, especialmente no contexto da doença Diabetes.
- Contribuição significativa para a análise de dados: Estimar os benefícios potenciais da utilização do CRISP-DM na melhoria dos resultados e na geração de insights na análise de dados, independentemente das bases de dados utilizadas.

Os benefícios previstos incluem a demonstração da aplicabilidade do modelo CRISP-DM como um complemento no processo de mineração de dados, neste caso utilizado na área da saúde, especificamente sobre a doença Diabetes. Por mais que seja utilizado bases de dados sobre Saúde, o foco principal é contribuir significativamente na área de análise de dados propondo o uso do CRISP-DM.

4.1.8 Metas e critérios de sucesso da mineração de dados

- Metas de mineração de dados:
  - Descrição do Problema: Definir o problema de mineração de dados como sendo a construção de classificadores para prever a probabilidade de indivíduos possuírem Diabetes, utilizando métricas presentes no Weka, como instâncias classificadas corretamente, matriz de confusão, área sob a curva ROC, precisão, estatística Kappa, entre outros.
  - Documentação das Metas Técnicas: Registro das metas técnicas, especificando os resultados desejados, alinhados com o modelo de processos CRISP-DM, como a definição dos objetivos de negócio convertidos em objetivos de mineração de dados.
  - Estabelecimento de Resultados Desejados: Definição de números reais para os resultados desejados, como a porcentagem de instâncias classificadas corretamente, estatística Kappa e a área sob a curva ROC.

- Critérios de Sucesso da Mineração de Dados:
    - Métodos de Avaliação do Modelo: Métodos para avaliar o desempenho dos classificadores, incluindo métricas instâncias classificadas corretamente, estatística Kappa, precisão, *recall*, F1-*score*, área sob a curva ROC, entre outros.
    - Definição de Avaliações de Desempenho: Avaliações de desempenho específicas para determinar o sucesso do projeto, alinhadas com os objetivos do modelo CRISP-DM. Ficou definido que para ambas as métricas dos classificadores, uma pontuação acima de 80% do seu total é um critério de sucesso perante a mineração de dados.
    - Planejamento da Implementação: Considerar se a utilização eficaz do modelo de processos CRISP-DM é parte do sucesso da mineração de dados.

4.1.9 Plano de projeto

O objetivo do Plano de projeto é estabelecer uma compreensão sólida dos objetivos do projeto e das necessidades do negócio em relação à mineração de dados relacionada à diabetes. Esta etapa é fundamental para garantir que o projeto esteja alinhado com as metas e requisitos da organização, fornecendo uma base sólida para todo o processo de mineração de dados.

Quadro 3 - Plano de projeto

Fase	Tempo	Recursos	Riscos
Entendimento do negócio	4 semanas	Toda a equipe	Mudança de objetivos
Preparação dos dados	9 semanas	Toda a equipe	Indisponibilidade e inadequação de dados, tecnologia e ferramentas
Modelagem	4 semanas	Toda a equipe	Resultados insuficientes, problemas de tecnologia e ferramentas
Avaliação	3 semanas	Toda a equipe	Necessidade de preparar os dados e modelar novamente
Implementação	Em aberto	Ninguém	Nenhum

Fonte: Autoria própria (2024)

As etapas a serem executadas neste processo CRISP-DM consistem na execução de 5 fases, deixando a implementação como próximo passo do estudo. Dito isso, tem-se por finalidade demonstrar como a utilização do modelo de processos CRISP-DM pode vir a gerar um conteúdo relevante no âmbito de análise e processamento de dados.

4.1.10 Avaliação inicial de ferramentas e técnicas

Como citado anteriormente, o *software* Weka é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Ele contém ferramentas para preparação de dados, classificação, regressão, agrupamento, mineração de regras de associação e visualização (Menotti, 2017).

Por dispor de vários algoritmos úteis para a tarefa de classificação e tratamento dos dados, o Weka foi utilizado por ser bastante relevante.

As rotinas de pré-processamento no Weka são chamadas de filtros, onde existem filtros para Discretização, normalização, amostragem, seleção de atributos, entre outros (Pozo, 2015).

A partir da documentação do *software* Weka (Universidade de Waikato, 1993) e segundo Aurora Trinidad Ramirez Pozo do artigo “Mineração de Dados: Conceitos, Aplicações e Experimentos com Weka” da Universidade Federal do Paraná, alguns filtros/métodos existentes e relevantes no Weka são:

- Métodos supervisionados para tratamento de dados:
    - *Resample*: Faz uma amostragem estratificada com o *dataset* fornecido, onde o *dataset* deve ter um atributo nominal informando a classe. O filtro produz uma subamostra aleatória de um conjunto de dados usando amostragem com reposição ou sem reposição. O filtro pode ser feito para manter a distribuição de classes na subamostra ou para direcionar a distribuição de classes em direção a uma distribuição uniforme.
    - *StratifiedRemoveFolds*: Cria um *fold* estratificado para o *cross-validation*. Este filtro pega um conjunto de dados e gera uma dobra especificada para validação cruzada. A versão não supervisionada não estratifica as dobras.

    - *SpreadSubsample*: Produz uma amostra aleatória dos dados. Este filtro permite definir o máximo *spread* entre a classe mais rara e a classe mais comum. Por exemplo, 5:1.
- Métodos não-supervisionados para tratamento de dados:
    - *Resample* - amostragem aleatória (não estratificada) do *dataset*.
    - *Randomize* - embaralha conjunto de dados.
    - *RemoveFolds* – Define um *fold* para o *cross validation*.
    - *RemovePercentage* – Remove uma proporção do *dataset*.
    - *RemoveRange* - Remove um determinado intervalo de instâncias do *dataset*.
- Métodos supervisionados para tratamento de atributos:
    - *AttributeSelection* - Permite a combinação de vários métodos de avaliação e busca de atributos.
    - *Discretize*: O método discretiza um intervalo de atributos numéricos utilizando a técnica MDL (Fayyad & Irani's) ou MDL (Kononenko).
    - *NominalToBinary* - O método converte todos os atributos nominais para atributos binários numéricos.
- Métodos não-supervisionados para tratamento de atributos:
    - *Normalize*: valores no intervalo [0,1], exceto o atributo de classe.
    - *NumericTransform* - Aplica uma função matemática qualquer aos valores do atributo (classe Java).
    - *ReplaceMissingValues* – Preenche com a média (atrib. numérico) ou a moda (atrib.nominal).
    - *Standardize* – transformação dos valores para uma N (0,1).
    - *RemoveUseless* - Remove atributos nominais que variam muito (*threshold* definido pelo usuário, ex.: 95%) e atributos constantes (nme/nml).

Sendo assim, as duas bases de dados foram escolhidas para servirem de base para este estudo foram:

- *Early-Stage Diabetes Risk Prediction Dataset* (Faniqul, 2020).
    - Este conjunto de dados foi alimentado originalmente por meio de questionários diretos de pacientes de um Hospital em Sylhet, Bangladesh.

- O artigo onde a base de dados foi utilizada é o denominado *"Early-Stage Diabetes Prediction using Data Mining Algorithms"* (Moniruzzaman, 2022), disponível no Portal ACM *Digital Library*. O estudo citado foi realizado a fim de utilizar técnicas de Mineração de Dados para melhorar as métricas da predição, aplicando uma série de técnicas de aprendizado de máquina.

- *Pima Indians Diabetes Database* (UCI MACHINE LEARNING, 2016)
  - A base de dados é originalmente do Instituto Nacional de Diabetes e Doenças Digestivas e Renais. Todos os pacientes são mulheres com pelo menos 21 anos de idade, de ascendência indígena Pima.
  - O artigo onde a base de dados foi utilizada é o denominado "*Using Machine Learning Algorithms to Predict Diabetes Mellitus Based on PIMA Indians Diabetes Dataset*"(Miao, 2021) e tem como objetivo determinar a probabilidade de pacientes com a doença diabetes. Ele se concentra na construção de um modelo comparando uma série de algoritmos de aprendizado de máquina.

4.1.11 Resumo

O estudo visa não apenas implementar os classificadores, mas também destacar a aplicação eficaz do Modelo CRISP-DM no contexto da análise de dados de pacientes com Diabetes. O principal objetivo é documentar cada etapa do processo CRISP-DM e demonstrar sua contribuição para a Mineração de Dados na área da saúde. Os critérios de sucesso incluem contribuir para a adoção do Modelo de Processos, documentar adequadamente o processo de Mineração de Dados, construir classificadores com desempenho satisfatório e demonstrar a utilização de bases de dados públicas.

A equipe utiliza o *software* Weka devido à sua variedade de algoritmos de aprendizado de máquina e suas funcionalidades para análise de dados, onde os filtros e métodos relevantes do Weka são discutidos, assim como a justificativa para sua escolha. Os recursos de *hardware* incluem um servidor com processador Ryzen 5 2600 e 16 GB de RAM.

Os requisitos do projeto incluem a definição das bases de dados, a escolha do *software* Weka e a avaliação contínua de sua adequação. As suposições envolvem a

qualidade e características das bases de dados selecionadas, enquanto as restrições incluem a utilização exclusiva do software Weka.

Os riscos identificados incluem a impossibilidade de finalizar o estudo no prazo estabelecido, a integridade dos dados não atender às expectativas e a indisponibilidade dos dados. Planos de contingência foram estabelecidos para cada risco, incluindo monitoramento do progresso do projeto e avaliação de fontes alternativas de dados.

Os custos estimados incluem tempo e recursos de pessoal, enquanto os benefícios previstos incluem demonstrar a aplicabilidade do modelo CRISP-DM, complementar o processo de Mineração de Dados na área da saúde e contribuir significativamente para a análise de dados. As metas e critérios de sucesso da Mineração de Dados estão alinhados com os objetivos do modelo CRISP-DM e incluem avaliar o desempenho dos classificadores construídos.

O projeto segue as cinco fases do Modelo CRISP-DM, com a implementação como próximo passo. O objetivo é demonstrar como a utilização do modelo CRISP-DM pode gerar um conteúdo relevante na análise e processamento de dados.

4.1.12 Perguntas

- O que seu negócio espera alcançar com este projeto?
  - O objetivo principal do negócio é construir classificadores eficazes para prever a probabilidade de indivíduos possuírem Diabetes, utilizando métricas presentes no Weka. Além disso, o negócio busca documentar cada etapa do processo utilizando o Modelo CRISP-DM e demonstrar como a adoção desse modelo pode influenciar positivamente o processo de Mineração de Dados.
- Como você definirá a conclusão bem-sucedida de nossos esforços?
  - A conclusão bem-sucedida será definida pelo alcance dos critérios estabelecidos, como a contribuição para a adoção do Modelo de Processos, a qualidade da documentação do processo de Mineração de Dados, o desempenho dos classificadores construídos e a utilização eficaz do modelo CRISP-DM.
- Você tem o orçamento e os recursos necessários para atingir suas metas?

  - Os recursos necessários foram identificados, incluindo acesso às bases de dados públicas sobre diabetes, literatura relevante e ferramentas de mineração de dados. Além disso, foi especificado um servidor adequado para suportar as atividades de análise de dados do projeto.
- Você tem acesso a todos os dados necessários para este projeto?
  - Sim, foram selecionadas duas bases de dados públicas sobre diabetes para o projeto, *Early-Stage Diabetes Risk Prediction Dataset* e *Pima Indians Diabetes Database*.
- Você e sua equipe já discutiram os riscos e as contingências associados a este projeto?
  - Sim, foram identificados diversos riscos associados ao projeto, como a impossibilidade de finalizar o estudo dentro do prazo estabelecido e a integridade dos dados não condizentes com o esperado. Planos de contingência foram estabelecidos para lidar com esses riscos.
- Os resultados de sua análise de custo-benefício tornam este projeto vantajoso?
  - Sim, os benefícios previstos incluem a demonstração da aplicabilidade do modelo CRISP-DM como um complemento no processo de Análise de Dados, a fim de contribuir para estudos na área e melhorar a eficiência da análise de dados no contexto da doença Diabetes, não se limitando apenas a área da saúde, mas propondo a utilização em qualquer área.
- Especificamente, como a mineração de dados pode ajudá-lo a atingir suas metas de negócios?
  - Pode ajudar a atingir as metas de negócios construindo classificadores eficazes para prever a probabilidade de indivíduos possuírem Diabetes, utilizando métricas presentes no Weka.
- Você tem alguma ideia de qual técnica de mineração de dados pode produzir os melhores resultados?
  - Foram selecionadas técnicas de mineração de dados presentes no Weka para pré-processamento de dados e construção de classificadores.

- Como você saberá quando os resultados são precisos ou eficazes o suficiente?
    - Os resultados serão avaliados com base nas métricas de desempenho dos classificadores, como instâncias classificadas corretamente, estatística Kappa, precisão, *recall*, F1-*score*, área sob a curva ROC, entre outros. Uma pontuação acima de 80% do total em ambas as métricas são consideradas um critério de sucesso perante a mineração de dados.
- Como os resultados da modelagem foram implementados? Você levou em consideração a implementação em seu plano do projeto?
    - A implementação dos resultados da modelagem não foi especificamente mencionada, deixando em aberto a implementação dos resultados obtidos futuramente.
- O plano do projeto inclui todas as fases do CRISP-DM?
    - Não, o plano do projeto menciona a utilização do Modelo CRISP-DM e inclui várias etapas alinhadas com esse modelo. Porém, não inclui a Implementação, deixando em aberto a aplicação dos resultados obtidos com os classificadores em futuros trabalhos e sistemas como um próximo passo.
- Os riscos e as dependências foram considerados no plano?
    - Sim, os riscos e as dependências foram identificados e abordados no plano do projeto, com planos de contingência estabelecidos para lidar com essas eventualidades.

4.2 Compreensão dos dados

Para a realização do trabalho, a aquisição das bases de dados se deu por meio da seleção de artigos no portal ACM *Digital Library*, onde a pesquisa foi focada em técnicas de *Machine Learning* aplicadas especificamente ao Diabetes, buscando os que fossem voltados para a predição e classificação da doença, a fim de aplicar os dados públicos utilizados na metodologia de processos CRISP-DM.

A pesquisa de artigos ocorreu selecionando artigos completos e que atendessem aos requisitos das seguintes *strings*:

- (("*classification algorithm*" AND "diabetes") OR "*classification* diabetes" OR "*prediction* diabetes" OR "*classify* diabetes")

A *string* de busca retornou inicialmente cerca de 220 resultados, diante disto, os seguintes critérios de inclusão e exclusão foram aplicados:

Quadro 4 - Critérios de Seleção

Critérios de Inclusão	Critérios de Exclusão
CI01 - Trate sobre a diabetes e *Machine Learning*; CI02 - Que seja sobre predição ou classificação.	CE01 - Materiais bibliográficos que não apresentem seus resultados de utilização. CE02 - Trabalhos que não sejam direcionados ao diabetes; CE03 - Materiais que não utilizem técnicas e métodos de classificação e predição.

Fonte: Autoria própria (2024)

Com base nos critérios de seleção, foi dado prioridade a artigos que tratassem sobre a doença Diabetes e *Machine Learning*, que fossem sobre mineração de dados/classificação e artigos que apresentem seus resultados. Após aplicar os critérios de inclusão e exclusão, 2 artigos foram selecionados para compor este estudo diante dos melhores resultados dos classificadores com relação aos outros artigos.

4.2.1 Experimento 1 – *Early-Stage Diabetes Risk Prediction Dataset*

A fase de compreensão dos dados do Experimento 1 utilizando a base de dados "*Early-Stage Diabetes Risk Prediction Dataset*" consiste na exploração dos dados, a fim de identificar inconsistências para garantir que os dados estão prontos para as fases subsequentes de pré-processamento e modelagem.

4.2.1.1 Coleta inicial de dados

Disponibilizada em 7 de novembro de 2020 no site UC *Irvine Machine Learning Repository*, a base "*Early-Stage Diabetes Risk Prediction Dataset*" possui um total de 520 instâncias e 17 atributos, incluindo um para realizar testes para classificação, sendo eles:

Quadro 5 - Tipos De Dados em Early Stage Diabetes Risk Prediction Dataset

Nome do atributo	Tipo do atributo	Detalhes
Idade	Numérico	Idade (anos)
Gênero	Binário	Gênero
Poliúria	Binário	Grandes quantidades de urina liberadas
Polidipsia	Binário	Sede excessiva ou Consumo excessivo de líquidos
Perda repentina de peso	Binário	Perda de peso repentina
Polifagia	Binário	Fome em excesso
Fraqueza	Binário	Sente-se fisicamente fraco
Candidíase genital	Binário	Se o paciente possui candidíase genital
Desfoque visual	Binário	Se o paciente tem visão embaçada
Coceira	Binário	Coceira na pele
Irritabilidade	Binário	Se o paciente fica irritado com facilidade
Cicatrização lenta	Binário	Cicatrização retardada de feridas
Paresia parcial	Binário	Se o paciente tem musculatura fraca
Rigidez muscular	Binário	Dificuldade em mover os músculos
Alopecia	Binário	Queda de cabelo em pequenas quantidades
Obesidade	Binário	Gordura excessiva no corpo
Classificação	Binário	Se o paciente possui diabetes ou não

Fonte: Autoria própria (2024)

4.2.1.2 Exploração dos dados

Os dados da base de dados estão dispostos na seguinte maneira:

Quadro 6 - Ocorrências Early Stage Diabetes Risk Prediction Dataset

Nome do atributo	Sim (Pacientes)	Não (Pacientes)
Poliúria	258	262
Polidipsia	233	287
Perda repentina de peso	217	303
Polifagia	237	283
Fraqueza	305	215
Candidíase genital	116	404
Desfoque visual	233	287
Coceira	253	267
Irritabilidade	226	394
Cicatrização lenta	239	281
Paresia parcial	224	296
Rigidez muscular	195	325
Alopecia	179	341
Obesidade	88	432
Resultado	320	200

Fonte: Autoria própria (2024)

Os atributos que constam na base de dados são do tipo nominal, com valores YES(SIM) e NO(NÃO).

4.2.1.3 Relatório de Exploração de dados

- Relação entre Atributos e Diabetes: Observou-se que certos atributos, como obesidade, fraqueza e cicatrização lenta podem ter uma forte correlação com o desenvolvimento do diabetes.
- Relevância dos Atributos: Alguns atributos, como alopecia (queda de cabelo) e irritabilidade, parecem ter uma correlação menos direta com o diabetes e podem ser considerados menos relevantes para a análise preditiva.
- Identificação de Atributos Promissores: A partir das análises, identificamos os atributos mais promissores para a análise preditiva de diabetes, incluindo obesidade, fraqueza e cicatrização lenta.
- Revelação de Novas Características: As explorações revelaram insights sobre a relação entre os atributos e o diabetes, destacando a importância de certos fatores no processo de classificação.

4.2.1.4 Qualidade dos dados

Após uma análise inicial, verificou-se que o conjunto de dados "*Early-Stage Diabetes Risk Prediction Dataset"* está completo, sem valores faltantes ou ausentes.

4.2.1.5 Relatório de Verificação da qualidade dos dados

Integridade dos dados: Após uma análise inicial, verificou-se que não há evidências de corrupção ou problemas na integridade dos dados. Os dados parecem ser consistentes e confiáveis, pois todas as variáveis parecem ter sido preenchidas adequadamente, o que sugere uma alta completude dos dados.

4.2.2 Experimento 2 – *Pima Indians Diabetes Database*

A fase de compreensão dos dados do Experimento 2 envolve o mesmo processo realizado no Experimento 1, visando garantir a qualidade dos dados da base "*Pima Indians Diabetes Database*".

4.2.2.1 Coleta inicial de dados

Este conjunto de dados foi originalmente coletado pelo Instituto Nacional de Diabetes e Doenças Digestivas e Renais (NIDDK) e foi disponibilizado pelo repositório do UCI *Machine Learning*, e atualmente os dados já são utilizados no repositório default do Weka. A base de dados de diabetes dos índios Pima contém um total de 768 instâncias e 9 atributos, dispostos da seguinte forma:

Quadro 7 - Tipos de Dados Presentes em Pima Indians Diabetes Database

Nome do atributo	Tipo do atributo	Detalhes
Gravidezes	Numérico	O número de vezes que a pessoa esteve grávida.
Glicose	Numérico	A concentração de glicose no plasma sanguíneo.
Pressão arterial	Numérico	A pressão arterial diastólica da pessoa, medida em milímetros de mercúrio (mm Hg).
Espessura da Pele	Numérico	A espessura da dobra de pele no tríceps da pessoa, medida em milímetros (mm).
Insulina	Numérico	A concentração de insulina no sangue após 2 horas, medida em unidades internacionais por mililitro (mu U/ml).
Índice de Massa Corporal (BMI)	Numérico	O índice de massa corporal da pessoa, calculado como o peso em quilogramas dividido pelo quadrado da altura em metros (kg/m^2).
Pedigree de Diabetes	Numérico	A função de pedigree de diabetes que mede a predisposição genética para diabetes.
Idade	Numérico	A idade da pessoa em anos
Resultados	Numérico	Variável (0 ou 1) que determina se uma pessoa possui diabetes ou não.

Fonte: Autoria própria (2024)

4.2.2.2 Exploração dos dados

Os valores dos dados estão dispostos na seguinte maneira:

Quadro 8 - Tipos de Dados Presentes em Pima Indians Diabetes Database

Nome do atributo	Média	Mínimo	Máximo
Gravidezes	3	0	17
Glicose	117 mg/dl	0 mg/dl	199 mg/dl
Pressão arterial	72 mm Hg	0 mm Hg	122 mm Hg
Espessura da Pele	23 mm	0 mm	99 mm
Insulina	32 um U/ml	0 um U/ml	84 um U/ml
Índice de Massa Corporal(BMI)	32 kg/m²	0 kg/m²	67.1 kg/m²
Pedigree de Diabetes	0.37	0.08	2.42
Idade	29 anos	21 anos	81 anos

Fonte: Autoria própria (2024)

Os atributos existentes na base de dados "*Pima Indians Diabetes Database*" são do tipo numérico, com valores de máximo, mínimo e média.

4.2.2.3 Relatório de exploração dos dados

- Relação entre Atributos e Diabetes: Observou-se que certos atributos, como glicose, pressão arterial e índice de massa corporal (IMC), podem ter uma forte correlação com o desenvolvimento do diabetes.
- Relevância dos Atributos: Alguns atributos, como alopecia (queda de cabelo) e irritabilidade, parecem ter uma correlação menos direta com o diabetes e podem ser considerados menos relevantes para a análise preditiva.
- Identificação de Atributos Promissores: A partir das análises, identificamos os atributos mais promissores para a análise preditiva de diabetes, incluindo glicose, pressão arterial e IMC.
- Revelação de Novas Características: As explorações revelaram *insights* sobre a relação entre os atributos e o diabetes, destacando a importância de certos fatores no processo de classificação.

4.2.2.4 Qualidade dos dados

A partir da base de dados *Pima Indians Diabetes Database* (UCI MACHINE LEARNING, 2016) foram identificados valores nulos (0) que podem vir a afetar a qualidade do conjunto de dados. Os valores nulos foram definidos como aqueles com um valor de 0 em atributos que normalmente não seriam 0, como glicose, pressão arterial, espessura da pele, insulina e índice de massa corporal (BMI). Os seguintes atributos com valor 0 foram encontrados em cada categoria de atributo:

- Glicose: 5 valores de 0.
- Pressão Arterial: 35 valores de 0.
- Índice de Massa Corporal (BMI): 11 valores de 0.
- Espessura da Pele: 227 valores de 0.
- Insulina: 395 valores de 0.

4.2.2.5 Relatório de verificação da qualidade dos dados

- No conjunto de dados foram identificados valores nulos em vários atributos, como glicose, pressão arterial e insulina. Esses valores podem afetar a precisão da análise e devem ser tratados adequadamente.

- Inconsistências nos Dados: Observou-se uma inconsistência nos valores zero para atributos como glicose, pressão arterial e insulina, que são fisicamente impossíveis. Esses dados podem ser erros de medição ou valores ausentes codificados incorretamente.
- Conclusões e Recomendações: Preencher valores nulos com estimativas apropriadas ou remover registros com valores ausentes, dependendo do impacto na análise preditiva.

4.2.3 Resumo

A etapa de compreensão dos dados consistiu em dois experimentos distintos para cada base de dados, focados na aplicação de técnicas de *Machine Learning* para a predição e classificação do diabetes. No primeiro experimento, denominado "Experimento 1 - *Early-Stage Diabetes Risk Prediction Database*", os dados foram adquiridos através da seleção de artigos no portal ACM *Digital Library*, resultando na escolha de uma base de dados do *UC Irvine Machine Learning Repository*. Esta base continha 520 instâncias e 17 atributos, incluindo características como idade, gênero e sintomas relacionados à diabetes. A análise revelou correlações significativas entre certos atributos, como obesidade, fraqueza e cicatrização lenta, e o desenvolvimento da doença. Além disso, não foram encontrados valores faltantes ou inconsistências nos dados, garantindo a integridade e confiabilidade do conjunto de dados.

No segundo experimento, intitulado "Experimento 2 - *Pima Indians Diabetes Database*", os dados foram coletados da mesma forma, resultando em uma base de dados com 768 instâncias e 9 atributos. A análise exploratória destacou a importância de atributos como glicose, pressão arterial e índice de massa corporal (IMC) na predição do diabetes. No entanto, foram identificados valores nulos em atributos cruciais, como glicose e pressão arterial, sugerindo a necessidade de uma revisão cuidadosa para corrigir possíveis inconsistências.

Em ambos os experimentos, foram identificados atributos promissores para a análise preditiva do diabetes, fornecendo *insights* valiosos sobre a relação entre os sintomas e a doença. Recomendações foram feitas para lidar com valores nulos e inconsistências nos dados, visando garantir a precisão das análises e a confiabilidade dos resultados. Este processo abrangente de compreensão dos dados estabeleceu uma base sólida para as etapas subsequentes de modelagem e avaliação dos algoritmos de *Machine Learning*.

4.2.4 Perguntas

- Todas as fontes de dados foram claramente identificadas e acessadas? Estou ciente de algum problema ou restrição?
    - Sim, todas as fontes de dados utilizadas no projeto foram identificadas, incluindo suas origens e quais foram selecionadas para análise. Além disso, foi mencionado a identificação dos problemas de valores nulos encontrados em uma das bases de dados e a consciência desses problemas.
- Foram identificados atributos-chave nos dados disponíveis?
    - Sim, foram identificados atributos-chave nos dados disponíveis. A análise revelou que certos atributos, como glicose, pressão arterial e índice de massa corporal (IMC), apresentam uma forte correlação com o desenvolvimento do diabetes. Esses atributos são considerados essenciais para a análise preditiva, pois há evidências científicas que os relacionam diretamente à doença.
- Esses atributos podem ajudar a formular hipóteses?
    - Sim, esses atributos podem ajudar a formular hipóteses sobre o desenvolvimento do diabetes. Por exemplo, considerando atributos como glicose, pressão arterial e índice de massa corporal (IMC), podemos formular hipóteses sobre como esses fatores estão relacionados à probabilidade de uma pessoa desenvolver diabetes.
    - A partir desses atributos, é possível investigar questões como:
        - Qual é a relação entre os níveis de glicose no sangue e o risco de diabetes?
        - Como a pressão arterial elevada está associada ao desenvolvimento da doença?
        - Existe uma correlação entre o IMC e a predisposição ao diabetes?
- Foi observado o tamanho de todas as fontes de dados?
    - Sim, foi mencionado o número total de instâncias em cada conjunto de dados, fornecendo uma visão geral do tamanho dos dados.
- É possível usar um subconjunto de dados onde apropriado?

  - Sim, é possível utilizar um subconjunto de dados onde apropriado. Se durante a análise dos dados e o desenvolvimento dos modelos preditivos, as métricas não estiverem alcançando os objetivos desejados, é viável considerar a utilização de um subconjunto dos dados. Nesse caso, podem ser selecionados os dados mais relevantes e informativos para o problema em questão, concentrando-se nos atributos que demonstraram maior correlação com o diabetes ou que mostraram ser mais preditivos durante a exploração dos dados.
- Quais são os problemas de qualidade de dados deste projeto? Existe um plano para abordar esses problemas?
  - Os problemas de qualidade de dados identificados incluem valores nulos em uma das bases de dados.
  - Para tratamento dos valores nulos ou omissos, dois métodos não-supervisionados para tratamento de atributos foram selecionados, *RemoveUseless* e *ReplaceMissingValues*. Caso não sejam efetivos, a exclusão das instâncias pode ser considerada.
- Os passos de preparação de dados estão claros? Por exemplo, sabe-se quais origens de dados mesclar e quais atributos filtrar ou selecionar?
  - Sim, com base nos atributos-chave identificados, como glicose, pressão arterial e índice de massa corporal (IMC), podemos determinar quais origens de dados mesclar e quais atributos filtrar ou selecionar. Esses atributos foram considerados essenciais para a análise preditiva do diabetes devido à sua forte correlação com a doença, conforme revelado pela análise. Portanto, pode-se priorizar esses atributos durante a preparação dos dados.

4.3 Preparação dos dados

Nesta etapa é previsto uma série de processos que visam melhorar a qualidade dos dados, seguido da definição das etapas de pré-processamento que podem vir a melhorar as métricas e resultados da mineração de dados. As bases de

dados foram divididas em dois experimentos distintos, Experimento 1 – *Early-Stage Diabetes Risk Prediction Dataset* e Experimento 2 – *Pima Indians Diabetes Database.*

4.3.1 Experimento 1 – *Early-Stage Diabetes Risk Prediction Dataset*

Nesta etapa do Experimento 1 os dados são preparados e formatados para utilização na etapa 4 Modelagem, conforme o relatório de qualidade gerado na etapa anterior “Compreensão dos dados”.

4.3.1.1 Seleção de dados

- Seleção de itens (linhas): As instâncias presentes no conjunto de dados "*Early-Stage Diabetes Risk Prediction*" que possuírem valores nulos ou algum tipo de ruído foram descartadas na análise.
- Seleção de atributos (colunas): Foram selecionados atributos como idade, glicose, pressão arterial, índice de massa corporal (BMI), como sendo os atributos de mais relevância com base na sua importância para o risco de diabetes, onde podem vir a ser concentrados em um subconjunto de dados para análise.

4.3.1.2 Inclusão e exclusão de dados

Documentou-se que não houve a necessidade de incluir ou excluir dados no conjunto "*Early-Stage Diabetes Risk Prediction*" devido à ausência de valores nulos ou faltantes.

4.3.1.3 Limpeza e formatação dos dados

O conjunto de dados foi preparado de acordo com os requisitos de entrada do *software* Weka, garantindo consistência e conformidade com os algoritmos de classificação disponíveis.

4.3.2 Experimento 2 – *Pima Indians Diabetes Database*

Esta etapa busca seguir o mesmo processo realizado no Experimento 1 Preparação dos dados, visando manter a consistência e integridade dos dados para prosseguir com as próximas etapas da metologia.

4.3.2.1 Seleção de dados

- Seleção de itens (linhas): As instâncias presentes no conjunto de dados "*Pima Indians Diabetes Database*" que possuírem valores nulos ou algum tipo de ruído foram descartadas na análise.
- Seleção de atributos (colunas): Foram selecionados atributos como obesidade, fraqueza e cicatrização lenta, como sendo os atributos de mais relevância com base na sua importância para o risco de diabetes, onde podem vir a ser concentrados em um subconjunto de dados para análise.

4.3.2.2 Inclusão e exclusão de dados

Foram identificados valores nulos no conjunto "*Pima Indians Diabetes Database*" em atributos como glicose, pressão arterial, BMI, espessura da pele e insulina. Os valores nulos foram definidos como aqueles com um valor de 0 em atributos que normalmente não seriam 0, como glicose, pressão arterial, espessura da pele, insulina e índice de massa corporal (BMI). Os seguintes atributos com valor 0 foram encontrados em cada categoria de atributo:

- Glicose: 5 valores de 0.
- Pressão Arterial: 35 valores de 0.
- Índice de Massa Corporal (BMI): 11 valores de 0.
- Espessura da Pele: 227 valores de 0.
- Insulina: 395 valores de 0.

Esses valores podem afetar a precisão da análise e devem ser tratados adequadamente. Para tratamento dos valores nulos ou omissos, dois métodos não-supervisionados para tratamento de atributos foram selecionados, *RemoveUseless* e *ReplaceMissingValues*. Caso não sejam efetivos, a exclusão das instâncias pode ser considerada.

4.3.2.3 Limpeza e formatação dos dados

- Dois métodos não-supervisionados para tratamento de atributos foram selecionados para o conjunto de dados "*Pima Indians Diabetes Database*", *RemoveUseless* e *ReplaceMissingValues*. Caso não sejam efetivos, a exclusão das instâncias pode ser considerada.
- O filtro de pré-processamento não-supervisionado "*NumericToNominal*" para tratamento de atributos foi aplicado no conjunto "*Pima Indians Diabetes Database*", com o intuito de utilizar com êxito os classificadores propostos na próxima etapa, Modelagem.
- O conjunto de dados foi preparado de acordo com os requisitos de entrada do *software* Weka, garantindo consistência e conformidade com os algoritmos de classificação disponíveis.

4.3.3 Etapas de pré-processamento

Como citado anteriormente no tópico avaliação inicial de ferramentas e técnicas, alguns métodos serão utilizados a fim de tratar os dados e diminuir o balanceamento entre as classes.

Alguns métodos/filtros do Weka foram selecionados, onde segundo o artigo Mineração de Dados: Conceitos, Aplicações e Experimentos com Weka (Pozo, 2015), os seguintes métodos são relevantes para tratar de instâncias e atributos:

- Métodos supervisionados para tratamento de instâncias:
  - *Resample*, *StratifiedRemoveFolds* e *SpreadSubsample*.
- Métodos não-supervisionados para tratamento de instâncias:
  - *Resample*, *Randomize*, *RemoveFolds*, *RemovePercentage* e *RemoveRange*.
- Métodos supervisionados para tratamento de atributos:
  - *AttributeSelection*, *Discretize*, *NominalToBinary*.
- Métodos não-supervisionados para tratamento de atributos:
  - *Normalize*, *NumericTransform*, *Standardize*, *RemoveUseless* e *ReplaceMissingValues*.

4.3.4 Resumo

Nesta etapa do projeto, foi iniciado a Seleção de Dados, onde os conjuntos "*Early-Stage Diabetes Risk Prediction*" e "*Pima Indians Diabetes Database*" foram escolhidos devido à sua relevância para o problema em questão. Foram selecionados subconjuntos relevantes de dados, com base nos atributos mais importantes para a previsão do risco de diabetes, como idade, glicose, pressão arterial e índice de massa corporal (BMI), conforme discutido em nossa análise inicial.

Em seguida, na questão de Incluir ou Excluir Dados, foi constatado que não houve necessidade de excluir dados do conjunto "*Early-Stage Diabetes Risk Prediction*", pois não foram encontrados valores nulos ou faltantes. No entanto, identificou-se valores nulos no conjunto "*Pima Indians Diabetes Database*". Optou-se por aplicar métodos não-supervisionados, como *RemoveUseless* e *ReplaceMissingValues*, não descartando a remoção das instâncias para lidar com esses valores nulos e garantir a integridade dos dados.

Após a limpeza dos dados, foi aplicada a formatação e preparação adequada dos conjuntos de dados para o *software* Weka. Foram definidos métodos de pré-processamento, tanto supervisionados quanto não supervisionados, para tratar instâncias e atributos, conforme detalhado na seção de Etapas de Pré-processamento. Estas etapas visam garantir que os dados estejam consistentes e em conformidade com os requisitos de entrada do Weka, preparando-os para a análise e modelagem subsequentes.

Pode-se definir quais os métodos de tratamento e pré-processamento de dados serão testados, com o intuito de extrair o máximo de cada um dos classificadores. Os filtros de pré-processamento incluem:

- Métodos supervisionados para tratamento de instâncias:
    - *Resample*, *StratifiedRemoveFolds* e *SpreadSubsample*.
- Métodos não-supervisionados para tratamento de instâncias:
    - *Resample*, *Randomize*, *RemoveFolds*, *RemovePercentage* e *RemoveRange*.
- Métodos supervisionados para tratamento de atributos:
    - *AttributeSelection*, *Discretize*, *NominalToBinary*.
- Métodos não-supervisionados para tratamento de atributos:

- *Normalize*, *NumericTransform*, *Standardize*, *RemoveUseless* e *ReplaceMissingValues*.

4.3.5 Perguntas

- Com base em sua exploração e entendimento iniciais, você conseguiu selecionar subconjuntos relevantes de dados?
  - Sim, foram identificados os atributos mais relevantes, como idade, glicose, pressão arterial, BMI, entre outros, para análise. Dado uma primeira análise, optou-se por não dividir o trabalho em subconjuntos de dados onde poderão vir a ser separados caso surja a necessidade.
- Você limpou os dados de forma efetiva ou removeu os dados irrecuperáveis? Documente qualquer decisão no relatório final.
  - Sim, os dados foram limpos de maneira eficaz. No conjunto "*Pima Indians Diabetes Database*", foram identificados valores nulos em atributos cruciais, como glicose, pressão arterial, BMI, entre outros, e foram tratados utilizando métodos não-supervisionados, como "*RemoveUseless*" e "*ReplaceMissingValues*".
- Os diversos conjuntos de dados estão integrados adequadamente? Ocorreu algum problema de mesclagem que deva ser documentado?
  - Não houve integração de diversos conjuntos de dados neste estágio do projeto. Cada conjunto de dados foi tratado separadamente, de acordo com suas características específicas.
- Você pesquisou os requisitos das ferramentas de modelagem que planeja usar?
  - Sim, foram identificados os filtros de pré-processamento e classificadores relevantes disponíveis no *software* Weka.
- Há problemas de formatação que possam ser abordados antes da modelagem? Isso inclui questões de formatação necessária, bem como tarefas que possam reduzir o tempo de modelagem.
  - Os conjuntos de dados foram preparados de acordo com os requisitos de entrada do *software* Weka, garantindo que estivessem em conformidade com os algoritmos de classificação disponíveis. Não

foram identificados problemas de formatação que pudessem afetar a modelagem.

4.4 Modelagem

Na etapa de modelagem é onde são empregados as ferramentas, técnicas e métodos de pré-processamento junto aos dados anteriormente preparados, utilizando dos processos definidos nas três primeiras etapas do CRISP-DM (Compreensão de negócios, Compreensão dos dados e Preparação dos dados).

4.4.1 Seleção de técnicas de modelagem

Nesta fase, serão considerados diversos algoritmos de classificação para construir e avaliar o modelo de classificação utilizando dados referentes à doença Diabetes. Os algoritmos a serem utilizados incluem:

- *ZeroR*: É o método de classificação mais simples que depende do alvo e ignora todos os preditores. O classificador *ZeroR* simplesmente prevê a categoria majoritária (classe). Embora não haja poder de previsibilidade no *ZeroR*, ele é útil para determinar um desempenho de linha de base como referência para outros métodos de classificação (Sayad, 2010).
- *One R*: Abreviação de "*One Rule*", é um algoritmo de classificação simples, mas preciso, que gera uma regra para cada preditor nos dados e, em seguida, seleciona a regra com o menor erro total como sua "regra única" (Sayad, 2010).
- *C4.5 J48* (Árvore de Decisão): C4.5 é um dos algoritmos de classificação populares que gera uma árvore de decisão. Ele divide o conjunto de dados em subconjuntos com base nas características, facilitando a interpretação (Saha, 2018).
- *Naive Bayes*: O algoritmo *Naive Bayes* é eficaz para tarefas de classificação e é baseado no Teorema de *Bayes*. Ele assume independência entre os atributos, o que pode funcionar bem para alguns conjuntos de dados (Ray, 2017).

- *Multilayer Perceptron*: Uma rede neural de múltiplas camadas que pode aprender relações complexas nos dados. Isso pode ser benéfico quando as relações entre os atributos não são lineares (DeepAI, 2021).
- *Random Forest*: Uma floresta aleatória é um estimador de meta que ajusta vários classificadores de árvore de decisão em várias subamostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o ajuste excessivo (IBM, 2018).
- *AdaBoostM1*: É mais bem usado para aumentar o desempenho de árvores de decisão em problemas de classificação binária. Pode ser usado para aumentar o desempenho de qualquer algoritmo de aprendizado de máquina (Isaini, 2023).

O principal objetivo é empregar técnicas de classificação para prever a presença de diabetes em pacientes com base em dados clínicos e demográficos. Serão exploradas diferentes técnicas, como regressão logística, *Random Forest*, entre outras, conhecidas por sua eficácia na classificação de dados binários.

4.4.2 Suposições de modelagem

Devido ao formato numérico dos dados da base de dados *Pima Indians Diabetes database*, foi necessário aplicar o filtro de pré-processamento "*NumericToNominal*" do *software* Weka. Esse filtro converte os atributos numéricos em atributos nominais, garantindo que os algoritmos de classificação propostos possam ser aplicados de forma adequada.

4.4.3 *Design* de teste

Os critérios do modelo serão definidos com base nas métricas de desempenho dos algoritmos de classificação. Serão consideradas métricas como TP *Rate*, FP *Rate*, Precisão, Revocação e ROC Area.

O *software* Weka propõe quatro opções de teste para avaliar o desempenho dos métodos de classificação, onde de acordo com a documentação do Weka (Universidade de Waikato, 1993), são eles:

- "*Use training set*": Utiliza o mesmo *dataset* para treinar e testar o modelo.
- "*Supplied test set*": O usuário fornece um outro arquivo para testes.

- “*Cross-validation*”: Realiza validação cruzada utilizando o número de *folds* indicado pelo usuário.
- “*Percentage split*”: Separa uma partição para treinamento e outra para teste.

Serão utilizadas três opções de teste oferecidas pelo software Weka: "*Use training set*", "*Cross-validation*" e "*Percentage split*". Cada uma dessas opções será explorada para determinar a mais adequada para avaliar o desempenho dos modelos de classificação. A opção "*Supplied test set*" será desconsiderada neste momento, pois o foco está em não utilizar dados externos para os testes.

4.4.4 Construção dos modelos

Seleção de Algoritmos de Classificação: Os algoritmos de classificação foram escolhidos com base nos resultados da fase anterior do projeto, levando em consideração os métodos de pré-processamento propostos e a adequação aos dados das bases *Early Stage Diabetes Risk Prediction Dataset* e *Pima Indians Diabetes Database*.

Cada algoritmo de classificação será testado utilizando os diferentes métodos de pré-processamento definidos na fase de seleção de dados. Serão realizadas várias iterações para ajustar os parâmetros dos modelos, utilizando as opções de teste definidas anteriormente. Durante esse processo, serão registrados as configurações e os resultados de desempenho de cada modelo.

Registro e Comparação dos Resultados: Os resultados de desempenho de cada modelo serão registrados e comparados entre si. Serão identificados os modelos que apresentam o melhor desempenho com base nos critérios de excelência definidos anteriormente.

Seleção do Modelo Final: Com base na análise dos resultados, será selecionado o modelo de classificação que melhor atende aos objetivos do projeto e às métricas de desempenho estabelecidas.

Ao seguir essas etapas, será possível construir e avaliar adequadamente os modelos de classificação propostos, garantindo a escolha do modelo mais eficaz para prever a presença de diabetes nas bases de dados selecionadas.

4.4.5 Experimento 1 – Early-Stage Diabetes Risk Prediction Dataset

Nesta seção, vários modelos são construídos seguindo a metodologia CRISP-DM, utilizando a base de dados *Early-Stage Diabetes Risk Prediction Dataset* com o objetivo de testar diversos classificadores para determinar quais modelos apresentam melhor desempenho na predição do risco de Diabetes.

4.4.5.1 Modelo 1 - *ZeroR*

- Base de dados: *Early-Stage Diabetes Risk Prediction Dataset*
  - Configurações específicas do modelo:
    - Algoritmo: *ZeroR;* Opção de teste: *Percentage split* 66%; Filtro de pré-processamento: *StratifiedRemoveFolds*.
  - Resultados da aplicação do modelo:
    - Instâncias classificadas corretamente: 14 (77.7778%);
    - Estatística Kappa: 0;
    - Matriz de confusão: *True Positive*: 14, *False Positive*: 0, *False Negative*: 4, *True Negative*: 0,
    - ROC Area: 0.500.

O classificador *ZeroR* não apresentou um desempenho satisfatório, indicando que não forneceu informações significativas para a classificação das instâncias. Os modelos apenas conseguiram prever a classe majoritária para todas as instâncias, não sendo capazes de identificar outros padrões nos dados.

4.4.5.2 Modelo 2 - *OneR*

- Base de dados: *Early-Stage Diabetes Risk Prediction Dataset*
  - Configurações específicas do modelo:
    - Algoritmo: *OneR*;
    - Opção de teste: *Training* set;
    - Filtro de pré-processamento: *StratifiedRemoveFolds*.
  - Resultados da aplicação do modelo:
    - Instâncias classificadas corretamente: 44 (84.6154%);
    - Estatística Kappa: 0.6977;

- Matriz de confusão: *True Positive*: 24, *False Positive*: 8, *False Negative*: 0, *True Negative*: 20;
- ROC Area: 0.875.

O modelo apresentou um resultado significativo, demonstrando uma eficiência maior que o *ZeroR* para ambas as bases de dados.

4.4.5.3 Modelo 3 – C4.5(J48)

- Base de dados: *Early-Stage Diabetes Risk Prediction Dataset*
  - Configurações específicas do modelo:
    - Algoritmo: C4.5(J48);
    - Opção de teste: *Training set*;
    - Filtro de pré-processamento: Sem filtro.
  - Resultados da aplicação do modelo
    - Instâncias classificadas corretamente: 413 (98.6538%);
    - Estatística Kappa: 0.9716;
    - Matriz de confusão: *True Positive*: 316, *False Positive*: 4, *False Negative*: 3, *True Negative*: 197;
    - ROC Area: 0.995.

O modelo apresentou uma alta taxa de instâncias classificadas corretamente (98.6538%), indicando um desempenho promissor na predição das classes. A Estatística Kappa de 0.9716 sugere uma concordância substancial entre as previsões do modelo e as observações reais.

4.4.5.4 Modelo 4 – *Naive Bayes*

- Base de dados: *Early-Stage Diabetes Risk Prediction Dataset*
  - Configurações específicas do modelo:
    - Algoritmo: *Naive Bayes;*
    - Opção de teste: Training set;
  - Filtro de pré-processamento: *Stratified Remove Folds*.
  - Resultados da aplicação do modelo:
    - Instâncias classificadas corretamente: 50 (96.1538%);
    - Estatística Kappa: 0.9188;

- Matriz de confusão: *True Positive*: 31*, False Positive*: 1, *False Negative*: 1, *True Negative*: 19;
- ROC Area: 0.994.

O modelo apresentou uma alta taxa de instâncias classificadas corretamente, com Estatística Kappa próxima a 1 e pontuação ROC Area próxima a 1. Isso sugere que o modelo é eficaz na classificação das instâncias e pode fornecer previsões confiáveis.

4.4.5.5 Modelo 5 – *Multilayer Perceptron*

- Base de dados: *Early-Stage Diabetes Risk Prediction Dataset*
  - Configurações específicas do modelo:
    - Algoritmo: *Multilayer Perceptron*;
    - Opção de teste: *Training set*;
    - Filtro de pré-processamento: Sem filtro.
  - Resultados da aplicação do modelo:
    - Instâncias classificadas corretamente: 514 (98.8462%);
    - Estatística Kappa: 1;
    - Matriz de confusão: *True Positive*: 317, *False Positive*: 3, *False Negative*: 3, *True Negative*: 197;
    - ROC Area: 0.992.

Os resultados mostraram que o modelo foi capaz de classificar corretamente 514 instâncias, representando uma taxa de acerto de 98.8462%. Além disso, a estatística Kappa foi de 1, indicando uma concordância perfeita entre as classificações do modelo e as classificações reais. A área sob a curva ROC (ROC Area) foi de 0.992, indicando um excelente desempenho do modelo na discriminação entre as classes.

4.4.5.6 Modelo 6 - *RandomForest*

- Base de dados: *Early-Stage Diabetes Risk Prediction Dataset*
  - Configurações específicas do modelo:
    - Algoritmo: *RandomForest*;
    - Opção de teste: *Training set*;
    - Filtro de pré-processamento: Sem filtro.
  - Resultados da aplicação do modelo:

- Instâncias classificadas corretamente: 520 (100%);
- Estatística Kappa: 1;
- Matriz de confusão: *True Positive*: 320, *False Positive*: 0, *False Negative*: 0, *True Negative*: 200;
- ROC Area: 1.000.

O algoritmo *RandomForest* obteve resultados expressivos, alcançando uma taxa de classificação correta de 100%, uma Estatística Kappa de 1 e uma pontuação da área sob a curva ROC de 1.000, indicando um desempenho excelente na classificação.

4.4.5.7 Modelo 7 – *AdaBoostM1*

- Base de dados: *Early-Stage Diabetes Risk Prediction Dataset*
  - Configurações específicas do modelo:
    - Algoritmo: *AdaBoostM1*;
    - Opção de teste: *Training set*;
    - Filtro de pré-processamento: *StratifiedRemoveFolds*.
  - Resultados da aplicação do modelo:
    - Instâncias classificadas corretamente: 50 (96.1538%);
    - Estatística Kappa: 0.9202;
    - Matriz de confusão: *True Positive*: 30, *False Positive*: 2, *False Negative*: 0, *True Negative*: 20;
    - ROC Area: 1.000.

Esses resultados sugerem que o *AdaBoostM1* pode ser uma escolha eficaz para a base de dados *Early Stage Diabetes Risk Prediction*.

4.4.6 Experimento 2 – *Pima Indians Diabetes Database*

Seguindo os mesmos passos do Experimento 1, vários modelos são construídos no contexto da base de dados *Pima Indians Diabetes Database* com o objetivo de encontrar o classificador mais adequado na predição do risco de Diabetes.

4.4.6.1 Modelo 1 - *ZeroR*

- Base de dados: *Pima Indians Diabetes Database*
    - Configurações específicas do modelo:
        - Algoritmo: ZeroR;
        - Opção de teste: *Percentage split* 66%;
        - Filtro de pré-processamento: *NumericToNominal*.
    - Resultados da aplicação do modelo:
        - Instâncias classificadas corretamente: 176 (68.1992%);
        - Estatística Kappa: 0;
        - Matriz de confusão: *True Positive*: 178, *False Positive*: 0, *False Negative*: 83, *True Negative*: 0;
        - ROC Area: 0.500.

O modelo *ZeroR* não apresentou um desempenho satisfatório, indicando que não forneceu informações significativas para a classificação das instâncias. Não foram reveladas novas percepções ou padrões incomuns pelo modelo *ZeroR*.

4.4.6.2 Modelo 2 - *OneR*

- Base de dados: *Pima Indians Diabetes Database*
    - Configurações específicas do modelo:
        - Algoritmo: *OneR*;
        - Opção de teste: *Training set*;
        - Filtro de pré-processamento: *Stratified Remove Fold*s e *NumericToNominal*.
    - Resultados da aplicação do modelo:
        - Instâncias classificadas corretamente: 77 (100%);
        - Estatística Kappa: 1;
        - Matriz de confusão: *True Positive*: 50, *False Positive*: 0, *False Negative*: 0, *True Negative*: 27;
        - ROC Area: 1.000.
- Ambos os modelos apresentaram um resultado significativo, demonstrando uma eficiência maior que o *ZeroR* para ambas as bases de dados.

4.4.6.3 Modelo 3 – C4.5(J48)

- Base de dados: *Pima Indians Diabetes Database*
    - Configurações específicas do modelo:
        - Algoritmo: C4.5(J48);
        - Opção de teste: *Training set*;
        - Filtro de pré-processamento: *Stratified Remove Folds* e *NumericToNominal*.
    - Resultados da aplicação do modelo:
        - Instâncias classificadas corretamente: 500 (65.1042%);
        - Estatística Kappa: 0;
        - Matriz de confusão: *True Positive*: 500, *False Positive*: 0, *False Negative*: 268, *True Negative*: 0;
        - ROC Area: 0.500.

A taxa de instâncias corretamente classificadas foi significativamente abaixo do esperado (65.1042%), sugerindo um desempenho menos satisfatório do modelo.

4.4.6.4 Modelo 4 - *NaiveBayes*

- Base de dados: *Pima Indians Diabetes Database*
    - Configurações específicas do modelo:
        - Algoritmo: *Naive Bayes*;
        - Opção de teste: *Training set*;
        - Filtro de pré-processamento: *Stratified Remove Folds* e *NumericToNominal*.
    - Resultados da aplicação do modelo:
        - Instâncias classificadas corretamente: 77 (100%);
        - Estatística Kappa: 1;
        - Matriz de confusão: *True Positive*: 50, *False Positive*: 0, *False Negative*: 0, *True Negative*: 27;
        - ROC Area: 1.000.
- O modelo apresentou uma alta taxa de instâncias classificadas corretamente, com Estatística Kappa próxima a 1 e pontuação ROC Area próxima a 1. Isso sugere que o modelo é eficaz na classificação das instâncias e pode fornecer classificações confiáveis.

4.4.6.5 Modelo 5 - *RandomForest*

- Base de dados: *Pima Indians Diabetes Database*
  - Configurações específicas do modelo:
    - Algoritmo: *RandomForest*;
    - Opção de teste: *Training set*;
    - Filtro de pré-processamento: *NumericToNominal*.
  - Resultados da aplicação do modelo:
    - Instâncias classificadas corretamente: 768 (100%);
    - Estatística Kappa: 1;
    - Matriz de confusão: *True Positive*: 500, *False Positive*: 0, *False Negative*: 0, *True Negative*: 268;
    - ROC Area: 1.000.
- O algoritmo *RandomForest* obteve resultados expressivos em ambas as bases de dados, alcançando uma taxa de classificação correta de 100%, uma Estatística Kappa de 1 e uma pontuação da área sob a curva ROC de 1.000, indicando um desempenho excelente na classificação de ambas as bases de dados.

4.4.6.6 Modelo 6 – *AdaBoostM1*

- Base de dados: *Pima Indians Diabetes Database*
  - Configurações específicas do modelo
    - Algoritmo: *AdaBoostM1*
    - Opção de teste: *Training set*;
    - Filtro de pré-processamento: *NumericToNominal*.
  - Resultados da aplicação do modelo
    - Instâncias classificadas corretamente: 500 (65.1042%);
    - Estatística Kappa: 0;
    - Matriz de confusão: *True Positive*: 30, *False Positive*: 2, *False Negative*: 0, *True Negative*: 20;
    - ROC Area: 0.601.

Para a base de dados *Pima Indians Diabetes Data*base, o algoritmo *AdaBoostM1* não conseguiu atingir um desempenho satisfatório, obtendo uma

Estatística Kappa de 0 e uma pontuação da área sob a curva ROC (ROC Area) de 0.601.

4.4.7 Avaliação dos Resultados

Com base na avaliação dos resultados obtidos nos conjuntos de dados, o algoritmo *Random Forest* se mostrou como o modelo mais consistente e eficaz para ambas as bases de dados.

Na base de dados *Early-Stage Diabetes Risk Prediction Dataset* o algoritmo *Random Forest* demonstrou um desempenho superior em termos de precisão e estatística Kappa em comparação com os outros classificadores testados, como *Naive Bayes*, *Multilayer Perceptron* e *AdaBoostM1*.

Quadro 9 - Resultados Early-Stage Diabetes Risk Prediction Dataset

Classificador	Instâncias classificadas corretamente	Kappa	Matriz de confusão	ROC Area	Opções de teste e pré-processamento
ZeroR	14(77.7778%)	0	*TPositive*:14 *FPositive*: 0 *FNegative*: 4 *TNegative*: 0	0,500	*Percentage split* 66% com *Stratified Remove Folds.*
One R	44(84.6154%)	0.6977	*TPositive*: 24 *FPositive*: 8 *FNegative*: 0 *TNegative*: 20	0,875	*Training set* com *Stratified Remove Folds.*
Naive Bayes	50(96.1538%)	0.9188	*TPositive*: 31 *FPositive*: 1 *FNegative*: 1 *TNegative*: 19	0,994	*Training set* com *Stratified Remove Folds.*
AdaBoostM1	50(96.1538%)	0.9202	*TPositive*: 30 *FPositive*: 2 *FNegative*: 0 *TNegative*: 20	1,000	*Training set* com *Stratified Remove Folds.*
C4.5(J48)	513(98.6538%)	0.9716	*TPositive*: 316 *FPositive*: 4 *FNegative*: 3 *TNegative*: 197	0,995	*Training set* sem filtro.
Multilayer Perceptron	514(98.8462%)	1	*TPositive*: 317 *FPositive*: 3 *FNegative*: 3 *TNegative*: 197	0,992	*Training set* sem filtro.
Random Forest	520(100%)	1	*TPositive*: 320 *FPositive*: 0 *FNegative*: 0 *TNegative*: 200	1,000	*Training set* sem filtro.

Fonte: Autoria própria (2024)

Dado os resultados acima dos modelos desenvolvidos referentes a base de dados *Early-Stage Diabetes Risk Prediction Dataset*, vale ressaltar que para a maioria

dos classificadores testados o filtro *Stratified Remove Folds* se mostrou mais eficiente como etapa de pré-processamento de dados, melhorando significativamente as métricas dos modelos construídos com os algoritmos *OneR*, *ZeroR*, *Naive Bayes* e *AdaBoostM1,* porém o método de pré-processamento diminui drasticamente o número de instancias da base de dados. Os classificadores C4.5(J48), *Multilayer Perceptron* e *Random Forest* não resultaram em melhorias utilizando o filtro *Stratified Remove Folds*, porém obtiveram os melhores resultados. Dado os resultados analisados, o *Random Forest* se mostrou mais eficiente e foi selecionado como o modelo a ser utilizado na base de dados em questão.

Na base de dados *Pima Indians Diabetes Database*, o classificador *Random Forest* novamente demonstrou um desempenho superior em termos de precisão e estatística Kappa em comparação com os outros classificadores testados, como *Naive Bayes*, *ZeroR, C4.5(J48)* e *AdaBoostM1*.

Quadro 10 - Resultados Pima Indians Diabetes Database

Classificador	Instâncias classificadas corretamente	Kappa	Matriz de confusão	ROC Area	Opções de teste e pré-processamento
AdaBoostM1	500(65.1042%)	0	*TPositive*: 30 *FPositive*: 2 *FNegative*: 0 *TNegative*: 20	0,601	*Training set* e *NumericToNominal*.
C4.5(J48)	500(65.1042%)	0	*TPositive*: 500 *FPositive*: 0 *FNegative*: 268 *TNegative*: 0	0,500	*Training* e *NumericToNominal*.
ZeroR	178(68.1992%)	0	*TPositive*: 178 *FPositive*: 0 *FNegative*: 83 *TNegative*: 0	0,500	*Percentage split* 66% e *NumericToNominal*.
One R	77(100%)	1	*TPositive*: 50 *FPositive*: 0 *FNegative*: 0 *TNegative*: 27	1,000	*Training set* com *Stratified Remove Folds* e *NumericToNominal*.
Naive Bayes	77(100%)	1	*TPositive*: 50 *FPositive*: 0 *FNegative*: 0 *TNegative*: 27	1,000	*Training set* com *Stratified Remove Folds* e *NumericToNominal*.
Random Forest	768(100%)	1	*TPositive*: 500 *FPositive*: 0 *FNegative*: 0 *TNegative*: 268	1,000	Training set e *NumericToNominal*.

Fonte: Autoria própria (2024)

Para a base de dados *Pima Indians Diabetes Database*, o filtro de pré-processamento *Stratified Remove Folds* resultou em melhoras significativas nos

modelos *OneR* e *Naive Bayes*, porém houve uma redução drástica no número de instâncias. Vale ressaltar que dado o tipo numérico dos dados, foi necessário aplicar o filtro *NumericToNominal,* onde ele converte os atributos de tipo numérico para nominal, a fim de utilizar corretamente os classificadores propostos. O *Random Forest* novamente se mostrou o mais eficiente nesta base de dados, sendo considerado o modelo mais adequado para a tarefa de classificação.

Uma característica interessante do *Random Forest* é sua capacidade de lidar com ambas as bases de dados sem a necessidade de filtro de pré-processamento de instância. Enquanto outros algoritmos exigiram o uso de técnicas de pré-processamento, como o *StratifiedRemoveFolds*, o *Random Forest* mostrou-se robusto o suficiente para lidar diretamente com a complexidade dos dados, mantendo a integridade das instâncias originais. Vale ressaltar que para a base de dados "*Pima Indians Diabetes Database*" foi aplicado o filtro de tratamento de atributos "*NumericToNominal*" para utilizar os classificadores propostos.

Ao selecionar o *Random Forest* como o modelo final para este projeto, levamos em consideração não apenas suas métricas de desempenho, mas também sua capacidade de se adaptar a diferentes características dos conjuntos de dados e sua interpretabilidade.

O próximo passo envolverá a otimização e ajuste do modelo *Random Forest*, se necessário, para garantir um desempenho ainda melhor. Isso pode incluir ajustes nos hiperparâmetros do modelo.

Em resumo, a escolha do *Random Forest* como o modelo final foi baseada em uma análise abrangente dos resultados e é fundamentada na sua superioridade em termos de desempenho e capacidade de generalização.

4.4.8 Resumo

Na fase de modelagem, foram considerados diversos algoritmos de classificação para prever a presença de diabetes com base em dados clínicos. Algoritmos como *ZeroR*, *OneR*, C4.5 (J48), *Naive Bayes*, *Multilayer Perceptron*, *RandomForest* e *AdaBoostM1* foram escolhidos devido à sua eficácia na classificação de dados binários.

O pré-processamento incluiu a conversão de atributos numéricos em nominais para garantir a aplicabilidade dos algoritmos. Métricas como TP *Rate*, FP *Rate*,

Precisão, Revocação e Área ROC foram utilizadas para avaliar o desempenho dos modelos.

Foram exploradas quatro opções de teste: "*Use training set*", "*Supplied test set*", "*Cross-validation*" e "*Percentage split*", com o objetivo de determinar a mais adequada para avaliar o desempenho dos modelos de classificação.

Os resultados foram registrados para cada modelo testado em duas bases de dados: *Early Stage Diabetes Risk Prediction* e *Pima Indians Diabetes Database*. O algoritmo *RandomForest* destacou-se como o mais eficaz, apresentando uma taxa de classificação correta de 100% em ambas as bases de dados, sem a necessidade de pré-processamento adicional.

A escolha do *RandomForest* como modelo final foi baseada em sua consistência, capacidade de lidar com diferentes características dos conjuntos de dados e interpretabilidade. O próximo passo envolverá a otimização do modelo, se necessário, para garantir um desempenho ainda melhor.

Em resumo, o estudo propôs e avaliou diversos modelos de classificação para prever o diabetes com base em dados clínicos, destacando o *RandomForest* como a melhor opção devido ao seu desempenho superior e capacidade de generalização.

4.4.9 Perguntas

- Você está apto a compreender os resultados dos modelos?
    - Sim, os resultados incluem métricas de desempenho como taxa de instâncias corretamente classificadas, estatística Kappa e área sob a curva ROC.
- Os resultados do modelo fazem sentido para você sob a perspectiva puramente lógica? Existem inconsistências aparentes que precisam de maior exploração?
    - Os resultados do modelo fazem sentido sob uma perspectiva lógica, pois foram obtidos utilizando métodos estatísticos e algoritmos de aprendizado de máquina bem estabelecidos. Não há inconsistências aparentes nos resultados apresentados.
- A partir de uma olhada inicial, os resultados parecem abordar as questões de negócios da organização?

- Os modelos foram construídos para prever a presença de diabetes com base em dados clínicos e demográficos, o que é relevante para o contexto do problema.

- Você explorou mais de um tipo de modelo e comparou os resultados?
  - Sim, foram explorados diversos tipos de modelos, incluindo *ZeroR*, *OneR*, C4.5(J48), *Naive Bayes*, *Multilayer Perceptron*, *RandomForest* e *AdaBoostM1*. Os resultados de cada modelo foram comparados para determinar o mais eficaz.
- Os resultados de seu modelo são implementáveis?
  - Os resultados dos modelos são implementáveis, pois foram obtidos a partir de algoritmos de aprendizado de máquina comumente utilizados na indústria. O modelo selecionado como final, *Random Forest*, mostrou-se capaz de lidar com ambas as bases de dados sem a necessidade de pré-processamento adicional, o que facilita sua implementação.

4.5 Avaliação

Ao longo das diferentes etapas do projeto, os resultados foram avaliados em relação aos objetivos de negócio e aos critérios de sucesso estabelecidos. A aplicação do modelo CRISP-DM na construção de classificadores para análise de dados relacionados à doença Diabetes proporciona *insights* relevantes.

Os classificadores desenvolvidos apresentaram resultados consistentes em termos de precisão e desempenho, o que indica uma aderência aos critérios de sucesso estabelecidos. A eficácia do modelo CRISP-DM na organização e execução do processo de mineração de dados foi evidenciada pelos resultados obtidos. A precisão dos modelos, juntamente com outras métricas de avaliação, demonstrou a capacidade do CRISP-DM em orientar a construção de modelos preditivos eficazes, em que cada modelo foi avaliado em termos de sua capacidade de classificar corretamente as instâncias e sua adequação aos objetivos do projeto.

As descobertas mais significativas incluem a eficácia do algoritmo *Random Forest* na classificação precisa dos dados relacionados à doença Diabetes em ambas as bases de dados. Além disso, a identificação de padrões e relações nos dados que

levaram à seleção do *Random Forest* como o modelo final é uma descoberta importante.

Os modelos foram classificados de acordo com sua aplicabilidade aos objetivos de negócio, com o *Random Forest* destacando-se como o mais aplicável devido à sua capacidade de lidar com ambas as bases de dados sem a necessidade de pré-processamento adicional, com 100% de instâncias classificadas corretamente. A alta precisão e consistência dos modelos confirmam sua utilidade para a organização na tomada de decisões relacionadas à saúde e ao tratamento da diabetes.

Além disso, a escolha do *software* Weka se mostrou adequada para as tarefas de pré-processamento e modelagem. A vasta gama de algoritmos e ferramentas disponíveis no Weka facilitou a implementação dos modelos e permitiu a exploração de diferentes abordagens de análise de dados. A flexibilidade e a eficiência do Weka contribuíram significativamente para o sucesso do projeto.

4.5.1 Revisão do processo

Durante a revisão do processo de mineração de dados, foram identificadas áreas de melhoria e sugestões para atividades adicionais. Uma análise mais detalhada dos resultados obtidos em cada fase do projeto pode revelar oportunidades de otimização e aprimoramento dos modelos. Próximos passos:

Validação Externa: Realizar uma validação externa dos modelos desenvolvidos, utilizando conjuntos de dados adicionais, para verificar sua eficácia em diferentes contextos.

Exploração de Técnicas Adicionais: Explorar outras técnicas de pré-processamento e modelagem de dados, a fim de adequar novos conjuntos de dados ao interesse do projeto.

Documentação e Disseminação: Documentar e disseminar os resultados obtidos, contribuindo para o conhecimento na área de mineração de dados e auxiliando em futuras pesquisas e projetos relacionados.

4.5.2 Próximos passos

Com base na avaliação dos resultados e na revisão do processo, os próximos passos do projeto incluem:

- Realizar uma validação externa dos modelos desenvolvidos, utilizando conjuntos de dados adicionais, para verificar sua eficácia em diferentes contextos.
- Explorar outras técnicas de pré-processamento e modelagem de dados, a fim de adequar novos conjuntos de dados ao interesse do projeto.
- Documentar e disseminar os resultados obtidos, contribuindo para o conhecimento na área de mineração de dados e auxiliando em futuras pesquisas e projetos relacionados.

4.6 Implantação

A fase de implantação do CRISP-DM é onde os resultados dos modelos são preparados para serem utilizados de maneira prática. No contexto deste projeto, a implantação não envolveu a integração com sistemas operacionais ou a automação de processos. Em vez disso, a ênfase está na aplicação do modelo de processos CRISP-DM e na sua documentação.

4.6.1 Planejamento

- Seleção dos Modelos: Considerando os resultados da avaliação, que destacaram o desempenho superior do algoritmo *Random Forest* em termos de precisão e estatística Kappa em ambas as bases de dados escolhidas, decidimos adotá-lo como o modelo principal para a aplicação. Além disso, conforme sugerido pelo guia IBM SPSS *Modeler* CRISP-DM(IBM), é crucial elaborar um plano passo a passo para a integração dos modelos com os sistemas existentes. Isso inclui especificações técnicas detalhadas, como requisitos de formato de saída do modelo, garantindo a compatibilidade e a eficácia dos modelos dentro do ambiente de negócios.
- Monitoramento Contínuo: Estabelecer um sistema de monitoramento contínuo, a fim de acompanhar o desempenho dos modelos implantados.
- Avaliação de Impacto: Conduzir avaliações regulares para medir o impacto dos modelos implantados nos processos do negócio e nos resultados organizacionais. Isso permitirá uma análise contínua do valor agregado pelos modelos e a identificação de áreas para melhorias adicionais.

- Atualização e Manutenção: Plano de atualização e manutenção para garantir que os modelos permaneçam relevantes e precisos ao longo do tempo.

Em resumo, o planejamento de implantação visa garantir uma implementação bem-sucedida e sustentável dos modelos de classificação desenvolvidos, contribuindo para a melhoria dos processos de negócios e o alcance dos objetivos organizacionais relacionados à análise de dados, neste caso, sobre o conjunto de dados referentes à doença Diabetes.

4.6.2 Relatório Final

O projeto de mineração de dados foi conduzido com o propósito principal de utilizar o modelo de processos CRISP-DM norteando o processo de construção de classificadores. Sendo assim, buscou-se explorar conjuntos de dados relacionados à doença Diabetes, visando fornecer *insights* para a previsão, diagnóstico e possível tratamento da condição. Os objetivos foram estabelecidos não apenas na contribuição sobre como a adoção de um Modelo de Processos pode ser benéfico norteando processos de descoberta do conhecimento e mineração de dados, mas também na construção bem-sucedida de classificadores e a necessidade de compreender os fatores de risco associados à Diabetes, a fim de desenvolver modelos preditivos que possam auxiliar na classificação dos dados.

A metodologia CRISP-DM foi adotada como estrutura para o projeto, dividindo-o em seis fases distintas: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação. Cada fase foi planejada e executada com ênfase na seleção e preparação adequada dos dados, na escolha e avaliação de modelos de classificação e na identificação de estratégias de implantação e posterior manutenção.

O projeto utilizou duas bases de dados públicas: *Early Stage Diabetes Risk Prediction Dataset* e *Pima Indians Diabetes Database*. Diversos algoritmos de classificação foram testados, juntamente com etapas/processos de pré-processamento para encontrar a forma mais eficaz de manejo dos dados. Alguns algoritmos de classificação incluem: *ZeroR*, *One R*, C4.5 (J48), *Naive Bayes*, *Multilayer Perceptron*, *Random Forest* e *AdaBoostM1*. Após uma análise abrangente, o algoritmo *Random Forest* mostrou-se como o mais eficaz em ambas as bases de

dados, demonstrando uma precisão consistente e uma capacidade superior de lidar com a complexidade dos dados.

Dado os resultados apresentados na seção 4.4.7 - Avaliação dos resultados, fica evidente que o classificador *Random Forest* demonstrou superioridade em termos de desempenho e capacidade de generalização para ambos os experimentos, Experimento 1 – Early-Stage Diabetes Risk Prediction Dataset e Experimento 2 – Pima Indians Diabetes Database, onde obteve 100% das instancias classificadas corretamente.

Em resumo, a escolha do *Random Forest* como o modelo final foi baseada em uma análise abrangente dos resultados. Os próximos passos envolvem a otimização e ajuste do modelo *Random Forest*, se necessário, para garantir um desempenho ainda melhor. Isso pode incluir ajustes nos hiperparâmetros do modelo.

A avaliação dos resultados confirmou a eficácia dos modelos desenvolvidos em relação aos critérios de sucesso estabelecidos. Os classificadores demonstraram consistência em termos de precisão e desempenho, validando a metodologia CRISP-DM como uma abordagem eficaz para a construção de modelos preditivos. Além disso, a escolha do *software* Weka facilitou a implementação dos modelos, enquanto a análise detalhada dos resultados em relação aos critérios de sucesso empresarial demonstrou o valor agregado dos modelos construídos para o contexto da doença Diabetes.

A análise de custo/benefício considera os recursos investidos no projeto em relação aos benefícios obtidos. Embora o investimento inicial em coleta, preparação e análise dos resultados possa ser significativo, os benefícios potenciais incluem a melhor compreensão dos fatores de risco da doença, junto da demonstração de como um Modelo de Processos pode ser benéfico, propondo assim o CRISP-DM como um complemento ao processo de KDD.

O projeto de mineração de dados sobre Diabetes forneceu *insights* que podem auxiliar no aprimoramento dos processos de prevenção, diagnóstico e tratamento da doença. Os modelos desenvolvidos representam ferramentas poderosas para identificar padrões e tendências nos dados, contribuindo para uma abordagem mais proativa e personalizada sobre Diabetes. A implementação bem-sucedida dos modelos tem o potencial de impactar positivamente os resultados organizacionais.

A conclusão deste projeto destaca a importância da metodologia CRISP-DM na condução de projetos de mineração de dados. Recomenda-se a continuação da

exploração de diferentes técnicas de modelagem e algoritmos de aprendizado de máquina, bem como a realização de validações externas dos modelos desenvolvidos, a fim de garantir sua robustez e generalização para diferentes contextos. Além disso, a disseminação dos resultados obtidos pode contribuir significativamente para o avanço do conhecimento na área de mineração de dados e para futuras pesquisas e projetos relacionados.

5 CONSIDERAÇÕES FINAIS

O estudo realizado utilizando a metodologia CRISP-DM para a construção de classificadores relacionados à doença Diabetes proporcionou *insights* relevantes e contribuiu significativamente para estudos de Mineração de Dados. Ao finalizar este projeto, algumas considerações perante os objetivos surgem:

A análise de projetos que incorporaram o CRISP-DM contribuiu significativamente para a compreensão e planejamento do trabalho, guiando a implementação do CRISP-DM às necessidades específicas do projeto, permitindo compreender os benefícios e desafios associados ao seu uso em diferentes contextos. A revisão bibliográfica foi essencial, capacitando todos os envolvidos a executar este projeto.

O uso do CRISP-DM como estrutura para o projeto demonstrou sua eficácia e relevância na condução de projetos de mineração de dados. A abordagem hierárquica e iterativa permitiu uma compreensão detalhada de cada fase do processo, desde a compreensão do negócio até a fase de implantação dos modelos, fornecendo uma estrutura sólida para a realização do estudo.

A análise dos dados relacionados à diabetes e a construção de modelos preditivos não apenas ofereceram *insights* sobre os fatores de risco e padrões associados à doença, mas também contribuíram para o entendimento mais amplo sobre como é possível e benéfico a utilização de um Modelo de Processos na construção de classificadores como um complemento ao KDD. Os modelos desenvolvidos representam ferramentas valiosas que podem nortear profissionais a implementar estratégias de análise de dados em quaisquer contextos.

A elaboração de uma documentação abrangente das fases do processo de mineração de dados, adaptada às necessidades e requisitos do estudo de caso, proporcionou uma visão clara e acessível do processo. Esta documentação não apenas facilitou a compreensão dos detalhes técnicos do estudo, mas também serviu

como uma ferramenta educacional para aqueles interessados em se aprofundar no estudo sobre mineração de dados e sua aplicação.

Dado o objetivo de implementar algoritmos de classificação utilizando ferramentas do WEKA, o *software* se mostrou uma ferramenta robusta e eficaz para a implementação, oferecendo uma ampla variedade de algoritmos de classificação e recursos de pré-processamento de dados que foram utilizados no projeto.

Os modelos desenvolvidos representam ferramentas que podem ajudar profissionais a implementar estratégias de análise de dados em quaisquer contextos. A avaliação e comparação dos modelos permitiram identificar os mais eficazes, demonstrando a aplicação de diferentes algoritmos de classificação em problemas, não apenas relacionados a diabetes.

Em resumo, o estudo do CRISP-DM para a construção de classificadores relacionados à diabetes foi uma jornada que teve resultados interessantes, contribuindo tanto para o avanço do conhecimento na área da saúde quanto para a compreensão mais profunda dos processos de mineração de dados. O sucesso deste projeto valida a importância da metodologia CRISP-DM como uma abordagem eficaz e abrangente para a análise de dados em diversos contextos e destaca seu potencial para impulsionar futuras pesquisas e inovações.

5.1 Trabalhos futuros

Como em qualquer projeto de pesquisa, há espaço para expansão e aprimoramento. Futuras pesquisas podem explorar ainda mais os dados relacionados à diabetes, considerando diferentes fontes de informação e incorporando novas técnicas de análise de dados. Além disso, a aplicação do CRISP-DM em outras áreas da saúde e em diferentes contextos organizacionais pode fornecer *insights* adicionais e contribuir para avanços significativos no campo da mineração de dados.

REFERÊNCIAS

AVELAR, Cátia Fabíola Parreira; ROCHA, Thiago Augusto Hernandes; CRUZ, Flávia Juliesse Soares. MINERAÇÃO DE DADOS**:** uma revisão da literatura em Administração. Revista das Faculdades Integradas Vianna Júnior – Vianna Sapiens, Juiz de Fora, ISSN 2177 3726, 25p, dezembro 2017.

AZEVEDO, Ana Isabel Rojão Lourenço; SANTOS, Manuel Filipe**. KDD, SEMMA and CRISP-DM**: ***a parallel overview***. 2008. 6p. IADIS *European Conference on Data Mining* - Amsterdam, 2008.

CHAPMAN, P. et al. **CRISP-DM 1.0** ***Step-by-step data mining guide***, kde.cs.uni-kassel.de, 2000. Disponível em: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf. Acesso em: 29 de fevereiro de 2024.

CHAUDHARY, Bivek. ***Importance of Data****: (A Term Paper)*. 2023. 8p. *Data Analysis – Tribhuvan University*, Nepal, 2023.

DAMASCENO, Marcelo. **Introdução a Mineração de Dados Utilizando o WEKA** [s. d.]. 14p. Conferência – Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte/Campus Macau, [s.d.].

DeepAI. ***What are multilayer perceptrons?*****.** DeepAI, [s.d.] Disponível em: https://deepai.org/machine-learning-glossary-and-terms/multilayer-perceptron. Acesso em: 17 de outubro de 2023.

FANIQUIL, Islam Et al. ***Early Stage Diabetes Risk Prediction Dataset*****.** Kaggle, 2020. Disponível em: https://www.kaggle.com/ishandutta/early-stage-diabetes-risk-prediction-dataset. Acesso em: 19 de junho de 2023.

FAYYAD, et al. ***Knowledge Discovery and Data Mining: Towards a Unifying Framework***, 1996. 7p. KDD'96 Conferência Internacional – AAAI Press, Portland Oregon, 1996.

FAYYAD, et al. ***From Data Mining to Knowledge Discovery in Databases****, American Association for Artificial Intelligence* – Portland Oregon. 1996.

Filho RAC, Corrêa LL, Ehrhardt AO, Cardoso GP, Barbosa GM. **O papel da glicemia capilar de jejum no diagnóstico precoce do diabetes mellitus**: Correlação com fatores de risco cardiovascular. Scielo Brasil, 2002. Disponível em: https://www.scielo.br/j/abem/a/sHHCMhTKHM6HsfG4dPHR55q/?lang=pt. Acesso em: 01 de março de 2024.

FREITAS, Natália Emília Pereira De. **ANÁLISE DE DADOS EPIDEMIOLÓGICOS E CLÍNICOS EM PACIENTES COM DIABETES MELLITUS TIPO 2**. 2019. 55p. Análise de Dados – Universidade Federal do Rio Grande do Norte, Natal, 2019.

IBM - **Introdução ao CRISP-DM**, IBM, 2023. Disponível em: https://www.ibm.com/docs/pt-br/spss-modeler/18.4.0?topic=guide-introduction-crisp-dm. Acesso em: 25 de abril de 2024.

IBM. ***What is Random Forest?***, IBM, 2023. Disponível em: https://www.ibm.com/topics/random-forest. Acesso em: 17 de outubro de 2023.

ISAINI, Anshul. ***AdaBoost Algorithm: Understand, Implement and Master AdaBoost***, *Analytics* Vidhya, 2023. Disponível em: https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/#What_Is_the_AdaBoost_Algorithm. Acesso em: 17 de outubro de 2023.

JUNIOR, Roberto Rosa da Silveira; RODRIGUEZ, Daniel Lins, **MINERAÇÃO DE DADOS**: UM OLHAR DE POSSIBILIDADES E APLICAÇÕES PARA ÓRGÃOS DA ADMINISTRAÇÃO PÚBLICA FEDERAL, 2022, 28p. Mineração de Dados – Universidade de Brasília, Distrito Federal, 2021.

MENOTTI, David. ***Data Mining with Open Source Machine Learning Software in Java***. 2017. 44p – *Data Mining* – Universidade Federal do Paraná, Curitiba, 2017.

MICHELUTTI, Marilene Miguel. **DIABETES MELLITUS**: A IMPORTÂNCIA DO DIAGNÓSTICO E DO TRATAMENTO. 2006. 8p. Bioquímica-Clínica – Academia de Ciencia e Tecnologia de São José do Rio Preto, São José do Rio Preto, 2006.

MICROSOFT. **O que é o Processo de Ciência de Dados de Equipe?** Microsoft, 2024. Disponível em: https://learn.microsoft.com/pt-br/azure/architecture/data-science-process/overview. Acesso em 15 de março de 2024.

NETO, Milton Vasconcelos da Gama. **O processo CRISP-DM aplicado na construção de uma solução para Análise de Risco de Crédito**. 2018. Análise de Dados – Universidade Federal de Pernambuco, Recife, 2018.

Organização Mundial Da Saúde (OMS). **Diabetes**. *World Health Organization*, 2024. Disponível em: https://www.who.int/health-topics/diabetes#tab=tab_1. Acesso em: 02 de março de 2024.

Özçelik, Fatih Eren. ***Process Models for Data Science Projects: CRISP-DM and KDD***, *Medium*, 2022. Disponível em: https://medium.com/kodluyoruz/process-models-for-data-science-projects-crisp-dm-and-kdd-172b631dcac1. Acesso em: 03 de junho de 2024.

PACE. Ana Emilia; NUNES. Polyana Duckur; OCHOA-VIGO. Katia. **O Conhecimento dos Familiares Acerca da Problemática do Portador de Diabetes Mellitus**. 2003. 8p, Enfermagem – Universidade de São Paulo, São Paulo, 2003.

FIOCRUZ. **Diabetes**. Fiocruz, [s.d.] Disponível em: https://portal.fiocruz.br/diabetes. Acesso em: 01 de março de 2024.

PEKER, S; Kart Ö. ***Transactional data-based customer segmentation applying CRISP-DM methodology*: *A systematic review***. 2023. 21p. *Data Mining – Journal of Data, Information and Management*, Suíça, 2023.

PLOTNIKOVA V; Dumas M; Milani F; ***Adapting the CRISP-DM Data Mining Process*: *A Case Study in the Financial Services Domain***. 2021. 16p. *Data Mining – Research Challenges in Information Science*, Suíça, 2021.

POZO, Aurora Trinidad Ramirez. **Mineração de Dados: Conceitos, Aplicações e Experimentos com Weka**, UFPR, 2015. Disponível em: https://www.inf.ufpr.br/aurora/disciplinas/ERBD/ERBD10.pptx. Acesso em: 20 de maio de 2024.

RAY, Sunil. ***Naive Bayes Classifier Explained: Applications and Practice Problems of Naive Bayes Classifier***. *Analytics* Vidhya, 2017. Disponível em: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/. Acesso em: 17 de outubro de 2023.

SAHA, Sumit. ***What is the C4.5 algorithm and how does it work? - Towards Data Science***. *Medium*, 2018. Disponível em: https://towardsdatascience.com/what-is-the-c4-5-algorithm-and-how-does-it-work-2b971a9e7db0. Acesso em: 17 de outubro de 2023.

SANTOS, Alexandra – **IBM SPSS como Ferramenta de Pesquisa Quantitativa**, 2018. 5p. Análise de Dados – Pontifícia Universidade Católica de São Paulo, São Paulo, 2018.

SANTOS, Rafael. **Conceitos de Mineração de dados na Web**, 2009. 40p. Mineração de Dados – Instituto Nacional de Pesquisas Espaciais, São Paulo, 2009.

SAP, **ROC Area**, SAP, 2018. Disponível em: https://help.sap.com/docs/SAP_ANALYTICS_CLOUD/00f68c2e08b941f081002fd3691d86a7/235c79933a7b4f398369e23a04520a3e.html?locale=pt-BR. Acesso em: 19 de junho de 2023.

SAS *Institute*, ***Introduction to SEMMA***, SAS, 2017. Disponível em: https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm. Acesso em 12 de março de 2024.

SAYAD, Saed, ***An Introduction to Data Science ONER E ZEROR***, Saedsayad, 2010. Disponível em: https://saedsayad.com/data_mining_map.htm. Acesso em: 17 de outubro de 2023.

SENNA, Ayrton. PENSADOR: **Seja você quem for, seja qual for a... Ayrton Senna**, Pensador, 1994. Disponível em: https://www.pensador.com/frase/MTE4ODIx/. Acesso em: 01 de janeiro de 2024.

SHEARER, Golin. *The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing*, Los Angeles, Volume 5, 15p, Dezembro 2000.

SILVA, Valdeir Luiz Da Silva. **Análise e mineração de dados dos cursos de pós-graduação do ensino à distância da UTFPR – Câmpus Medianeira**, 2015. 50p. Mineração de Dados – Universidade Tecnológica Federal do Paraná Campus Medianeira. Medianeira, 2015.

SRIDHARAN, Mithun. ***CRISP-DM - A Framework For Data Mining And Analysis****, Think Insights*, 2018. Disponível em https://thinkinsights.net/data/crisp-dm/. Acesso em 14 de março de 2024.

SMELTZER. Suzanne C; BARE. Brenda G. Brunner & Suddarth. Manual de Enfermagem Médico-Cirúrgica. Rio de Janeiro, 13 edição, 1.152p, dezembro, 2006.

UCI *Machine Learning Repository.* ***Welcome to the UC Irvine Machine Learning Repository***, UCI *Machine Learning Repository*, 2023. Disponível em: https://archive.ics.uci.edu/ml/index.php. Acesso em 24 de maio de 2023.

International Diabetes Federation, ***Facts & Figures,*** *International Diabetes Federation*, 2023. Disponível em: https://idf.org/about-diabetes/facts-figures. Acesso em 15 de junho de 2023.

UCI *Machine Learning,* ***Pima Indians Diabetes Database***, Kaggle, 2016. Disponível em: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Acesso em: 19 de junho de 2023.

Universidade de Waikato, **Pacotes de recursos do software Weka**, *SourceForge*, 2023. Disponível em: https://weka.sourceforge.io/doc.dev/. Acesso em: 21 de junho de 2023.

VIKAS, Nehra. ***Decision Tree,*** *Medium*, 2024. Disponível em: https://medium.com/@nehra.vikas95/decision-tree-db1d7f37187f. Acesso em 28 de abril de 2024.

WILD, Sarah. FISCHBACHER, Colin. MCKNIGHT, John. ***Using Large Diabetes Databases for Research****. 2016. 6p. Database research - Journal of Diabetes Science and Technology*, *United Kingdom*, 2016.

Apêndice A – Documentos gerados pelo CRISP-DM

Esta seção contém breves descrições do objetivo e do conteúdo dos relatórios mais importantes. Aqui, nos concentramos em relatórios que visam comunicar os resultados de uma fase a pessoas não envolvidas nesta fase (e possivelmente não envolvidas neste projeto).

Estes não são necessariamente idênticos aos resultados descritos no modelo de referência e o guia do usuário. O objetivo dos resultados é principalmente documentar os resultados durante a execução do projeto.

- Compreensão empresarial: Os resultados da fase de Entendimento do Negócio podem ser resumidos em um relatório. Sugerimos as seguintes seções:
    1. Fundo: O Contexto fornece uma visão geral básica do contexto do projeto. Isso lista qual área em que o projeto está funcionando, quais problemas foram identificados e porque a mineração de dados parece fornecer uma solução.
    2. Objetivos de negócios e critérios de sucesso: Os Objetivos de Negócios descrevem quais são os objetivos do projeto em termos de negócios. Para cada objetivo, Critérios de Sucesso Empresarial, ou seja, medidas explícitas para determinar se ou não, o projeto teve sucesso em seus objetivos, deverá ser fornecido. Esta seção deve também liste os objetivos que foram considerados, mas rejeitados. A justificativa para a seleção deve ser dados os objetivos.
    3. Inventário de recursos: O Inventário de Recursos visa identificar pessoal, fontes de dados, instalações técnicas e outros recursos que possam ser úteis na execução do projeto.
    4. Requisitos, suposições e restrições: Este resultado lista requisitos gerais sobre como o projeto é executado, tipo de resultados do projeto, suposições feitas sobre a natureza do problema e os dados que estão sendo usados e restrições impostas ao projeto.
    5. Riscos e contingências: Este resultado identifica problemas que podem ocorrer no projeto, descreve as consequências e declara quais ações podem ser tomadas para minimizar o efeito.

6. Terminologia: A Terminologia permite que pessoas não familiarizadas com os problemas abordados pelo projeto para se familiarizar mais com eles.
7. Custos e benefícios: Descreve os custos do projeto e os benefícios comerciais previstos se o projeto for bem-sucedido (por exemplo, retorno do investimento). Outros benefícios menos tangíveis (por exemplo, a satisfação do cliente) também devem ser realçados.
8. Metas de mineração de dados e critérios de sucesso: As metas de mineração de dados declaram os resultados do projeto que permitem o alcance de os objetivos do negócio. Além de listar as prováveis abordagens de mineração de dados, os critérios de sucesso para os resultados também devem ser listados em termos de mineração de dados.
9. Plano de projeto: Lista as etapas a serem executadas no projeto, juntamente com duração, recursos necessários, entradas, saídas e dependências. Sempre que possível, deverá tornar-se explícitas as iterações em grande escala no processo de prospecção de dados, por exemplo, repetições do fases de modelagem e avaliação.
10. Avaliação inicial de ferramentas e técnicas: Esta seção dá uma visão inicial de quais ferramentas e técnicas provavelmente serão usadas e como. Descreve os requisitos para ferramentas e técnicas, lista as ferramentas disponíveis e técnicas e combina-as com os requisitos.

- Compreensão dos dados: Os resultados da fase de compreensão dos dados são geralmente documentados em vários relatórios. Idealmente, estes relatórios devem ser escritos durante a execução das respetivas tarefas. Os relatórios descrever os conjuntos de dados que são explorados durante a compreensão dos dados. Para o relatório final, um resumo das partes mais relevantes é suficiente.
    1. Relatório inicial de coleta de dados: Este relatório descreve como as diferentes fontes de dados identificadas no inventário foram capturados e extraídos.
    2. Tópicos a serem abordados:
        - Contexto dos dados

- Lista de fontes de dados com ampla área de dados necessários coberta por cada uma.
- Para cada fonte de dados, método de aquisição ou extração.
- Problemas encontrados na aquisição ou extração de dados.

3 Relatório de descrição de dados: Cada conjunto de dados adquirido é descrito.

4 Tópicos a serem abordados

- Cada fonte de dados descrita detalhadamente
- Lista de tabelas (pode ser apenas uma) ou outros objetos de banco de dados.
- Descrição de cada campo incluindo unidades, códigos utilizados etc.
- Relatório de exploração de dados: Descreve a exploração de dados e seus resultados

5 Para cada área de exploração realizada:

- Regularidades ou padrões esperados.
- Método de detecção.
- Regularidades ou padrões encontrados, esperados e inesperados.
- Conclusões para transformação de dados, limpeza de dados e qualquer outro pré-processamento.
- Conclusões relacionadas às metas de mineração de dados ou objetivos de negócios.
- Resumo das conclusões.
- Contexto incluindo expectativas amplas sobre a qualidade dos dados.

- Preparação de dados: Os relatórios da fase de preparação dos dados concentram-se nas etapas de pré-processamento que produzem os dados a serem extraídos.

1 Relatório de descrição do conjunto de dados: Fornece uma descrição do conjunto de dados (após o pré-processamento) e o processo pelo qual ele foi produzido.

2 Tópicos a serem abordados:

- Contexto incluindo objetivos gerais e plano para pré-processamento. Justificativa para inclusão/exclusão de conjuntos de dados.
- Descrição do pré-processamento, incluindo as ações necessárias para resolver quaisquer problemas de qualidade dos dados
- Descrição detalhada do conjunto de dados resultante, tabela por tabela e campo por campo.
- Justificativa para inclusão/exclusão de atributos.
- Descobertas feitas durante o pré-processamento e quaisquer implicações para trabalhos futuros.
- Resumo e conclusões.

- Modelagem: Os resultados produzidos durante a fase de Modelagem podem ser combinados em um relatório. Sugerimos as seguintes seções:
  1 Suposições de modelagem: Esta seção define explicitamente quaisquer suposições feitas sobre os dados e quaisquer suposições implícitas na técnica de modelagem a ser usada.
  2 *Design* de teste: Esta seção descreve como os modelos são construídos, testados e avaliados.
  3 Tópicos a serem abordados: Contexto, descrição ampla do tipo de modelo e dos dados de treinamento a serem usados, explicação de como o modelo será testado ou avaliado, planejar produção de dados de teste, resumo do plano de teste, descrição do modelo, resumo e conclusões.
  4 Os resultados produzidos durante a fase de Modelagem podem ser combinados em um relatório. Sugerimos as seguintes seções:
    - Avaliação do modelo: Esta seção descreve os resultados dos testes dos modelos de acordo com o design do teste.
    - Tópicos a serem abordados: Visão geral do processo de avaliação e resultados, incluindo quaisquer desvios do plano de teste, avaliação detalhada do modelo, incluindo medições como precisão e interpretação do comportamento, comentários sobre

modelos feitos por especialistas no domínio ou em dados, *Insights* sobre porque uma determinada técnica de modelagem e determinadas configurações de parâmetros levaram a resultados bons/ruins e avaliação resumida do conjunto completo de modelos.

- Avaliação: Resultados da mineração de dados em relação aos critérios de sucesso empresarial Este relatório compara os resultados da mineração de dados com os objetivos empresariais e os critérios de sucesso empresarial:
  1 Tópicos a serem abordados:
    - Revisão dos Objetivos de Negócios e Critérios de Sucesso de Negócios (que podem ter mudado durante e/ou como resultado da mineração de dados).
    - Para cada critério de sucesso empresarial: Comparação detalhada entre critérios de sucesso e resultados de mineração de dados, conclusões sobre a viabilidade do critério de sucesso e adequação dos dados, processo de mineração de dados, revisão do sucesso do projeto, se existem novos objetivos de negócios a serem abordados posteriormente no projeto ou em novos projetos, conclusões para futuros projetos de mineração de dados.

  2 Revisão do processo: Esta seção avalia a eficácia do projeto e identifica quaisquer fatores que possam ter sido negligenciados e que devem ser levados em consideração se o projeto for repetido.

  3 Lista de ações possíveis: Esta seção faz recomendações sobre os próximos passos do projeto.

- Implantação: Esta seção especifica a implantação dos resultados da mineração de dados.
  1 Tópicos a serem abordados:
    - Resumo dos resultados implementáveis (derivado do relatório Próximas etapas).
    - Descrição do plano de implantação.

2 Plano de monitoramento e manutenção: O plano de monitoramento e manutenção específica como os resultados implantados devem ser mantidos.

- Visão geral da distribuição dos resultados e indicação de quais resultados podem exigir atualização (e por quê).
- Para cada resultado implantado.
- Descrição de como a atualização será acionada (atualizações regulares, acionamento, evento, monitoramento de desempenho).
- Descrição de como será realizada a atualização.
- Resumo do processo de atualização de resultados.

3 Relatório final: O relatório final é utilizado para resumir o projeto e seus resultados.

- Conteúdo: Resumo do Entendimento do Negócio: antecedentes, objetivos e critérios de sucesso, resumo do processo de mineração de dados, resumo dos resultados da mineração de dados, resumo da avaliação dos resultados, resumo dos planos de implantação e manutenção, análise de custo/benefício, conclusões para o negócio, conclusões para futuras pesquisas de mineração de dados.