UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
NOME DO DEPARTAMENTO OU PROGRAMA DE PÓS-GRADUAÇÃO
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E SUAS APLICAÇÕES

FERNANDA CRISTINA VIEIRA

ANÁLISE EXPLORATÓRIA DE DADOS: LIMPEZA, MANIPULAÇÃO E PRÉ PROCESSAMENTO APLICADO A DATASET DE PERFIL DE ATENDIMENTO NAS UNIDADES DE SAÚDE DA CIDADE DE CURITIBA

CURITIBA
2021

FERNANDA CRISTINA VIEIRA

ANÁLISE EXPLORATÓRIA DE DADOS: LIMPEZA, MANIPULAÇÃO E PRÉ PROCESSAMENTO APLICADO A DATASET DE PERFIL DE ATENDIMENTO NAS UNIDADES DE SAÚDE DA CIDADE DE CURITIBA

Exploratory data analysis: cleaning, manipulation and pre-processing applied to the care profile dataset in the Curitiba city health units

Monografia apresentada como requisito parcial à obtenção do título de Especialista em Ciência de Dados e suas aplicações, do Departamento Acadêmico de Informática, da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Leandro Batista de Almeida.

CURITIBA

2021

Ministério da Educação
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
UTFPR - CAMPUS CURITIBA
DIRETORIA-GERAL - CAMPUS CURITIBA
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO - CAMPUS CURITIBA
DEPARTAMENTO DE APOIO DAS ESPECIALIZAÇÕES LATO-SENSU DOS CURSOS DE INFORMÁTICA - CAMPUS CURITIBA
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E SUAS APLICAÇÕES

TERMO DE APROVAÇÃO

ANÁLISE EXPLORATÓRIA DE DADOS: LIMPEZA, MANIPULAÇÃO E PRÉ PROCESSAMENTO APLICADO A DATASET DE PERFIL DE ATENDIMENTO NAS UNIDADES DE SAÚDE DA CIDADE DE CURITIBA.

por

Fernanda Cristina Vieira

Este Trabalho de Conclusão de Curso foi apresentado às 19h00min do dia 09 de agosto de 2021 por videoconferência como requisito parcial à obtenção do grau de Especialista em Ciência de Dados e suas Aplicações na Universidade Tecnológica Federal do Paraná - UTFPR - Campus Curitiba. A aluna foi arguida pela Banca de Avaliação abaixo assinados. Após deliberação, a Banca de Avaliação considerou o trabalho aprovado.

_____

Prof. Dr. Leandro Batista de Almeida (Presidente/Orientador – DAINF-CT/ UTFPR-CT)

_____

Profa. Dra. Rita Cristina Galarraga Berardi (Avaliadora 1 – DAINF-CT/ UTFPR-CT)

_____

Prof. Msc. Christian Carlos de Souza Mendes (Avaliador 2 – DAELN-CT/ UTFPR-CT)

O Termo de Aprovação assinado encontra-se no sistema SEI- Processo nº 23064.031957/2021-91

RESUMO

O objeto de trabalho desse estudo abrange duas vertentes principais, a ciência de dados e o perfil de atendimento das unidades de saúde da cidade de Curitiba. Consiste na aplicação de análise exploratória de dados, focada na importância da qualidade das etapas iniciais como pré-requisito para posteriormente uma modelagem eficiente de Machine Learning. Ao analisar um dataset é possível obter informações importantes, que servirão para identificar padrões, características e associações entre os dados. Porém o grande volume de informações pode dificultar o processo, sendo assim, a combinação de análise exploratória e técnicas de visualização são essenciais para chegar a bons resultados. Este trabalho apresenta algumas técnicas de limpeza, manipulação e análise exploratória de dados.

Palavras-chave: Ciência de dados. Análise Exploratória. Atendimento de saúde.

ABSTRACT

The work object of this study encompasses two main aspects, data science and the service profile of healthcare units in the city of Curitiba. It consists of the application of exploratory data analysis, focused on the importance of quality in the initial stages as a prerequisite for an efficient modeling of Machine Learning later. When analyzing a dataset, it is possible to obtain important information, which will serve to identify patterns, characteristics and associations between the data. However, the large volume of information can make the process difficult, so the combination of exploratory analysis and visualization techniques are essential to achieve good results. This work presents some techniques for cleaning, manipulating and exploratory data analysis.

Key words: Data science. Exploratory Analysis. Health care.

SUMÁRIO

1 INTRODUÇÃO

Com o avanço da tecnologia da informação e o desenvolvimento de ferramentas para facilitar as atividades humanas nos mais diversos segmentos, uma infinidade de dados são gerados diariamente, repletos de informações relevantes para a tomada de decisão e que permitem a aquisição de experiências baseadas.

Os profissionais da computação por sua vez, tem trabalhado no sentido de atribuir valor a esses conteúdos, por meio da ciência de dados. Aplicações são desenvolvidas com potencial para processar grandes quantidades de dados e aplicação de técnicas de inteligência artificial (IA) e Machine Learning (ML). Sistemas computacionais capazes de aprender por si só, e tomar decisões sem que haja necessidade de interação humana. Conforme afirma Faceli et al. (2011), no início da utilização de IA para a solução de problemas reais, a aquisição do conhecimento de especialistas ocorria por meio de entrevistas que buscavam descobrir as regras que utilizariam para tomar decisões.

Atualmente, a análise exploratória de dados oferece aos cientistas uma série de técnicas, capazes de facilitar a interação inicial com os dados e prepara-los para melhor aplicar-se a futuros modelos de aprendizagem. Essa etapa permite aos profissionais compreender os dados com os quais irão trabalhar, sugerir hipóteses, extrair informações úteis dos dados e até mesmo substanciar demais dados para enriquecimento do dataset.

Com a ubiquidade do acesso à internet, dispositivos móveis e vestíveis, tem havido o desencadeamento de uma torrente de dados, bem como de empresas e máquinas que também geram enormes quantidades de dados (TAULLI, 2019).

Quando se trata de prestação de serviços públicos não é diferente, a grande quantidade de dados gerados tem incentivado o interesse pela interpretação e descoberta de informações com intuito de extrair conhecimento e apoiar a tomada de decisão dos poderes governamentais, no sentido de traçar estratégias para melhoria na qualidade dos atendimentos prestados à população.

O objetivo desse trabalho é apresentar o escopo a ser seguido para uma análise de qualidade, que possibilite a abstração de informações relevantes, ressaltando a importância dessa etapa inicial para posterior aplicação de modelos de aprendizagem de máquina. Aplicando técnicas de ciência de dados, desde a limpeza

e manipulação, até a análise exploratória de um dataset com dados do perfil de atendimento das unidades de saúde da cidade de Curitiba.

Este trabalho se justifica pela necessidade de descrever as técnicas de preparação e análise exploratória de dados, como também ferramentas para poder aplicá-las de forma eficiente e obter um conjunto de dados relevante.

O presente trabalho foi organizado em capítulos, descritos a seguir, e finalizando, foram apresentadas as referências utilizadas como embasamento ao referencial teórico. No capítulo 2, foram descritos os conceitos de Machine Learning, e na sequência os problemas ocasionados pela baixa qualidade dos dados. Foi apresentado o processo de preparação e análise exploratória de dados, bem como a aplicação de técnicas de ML na área da saúde, inclusive citando dois trabalhos correlatos. Ao final desse capítulo foram apresentadas as ferramentas, linguagens de programação e bibliotecas disponíveis e utilizados em ML, como também nesse trabalho.

No capítulo 3, são apresentados os passos para preparação e análise exploratória de dados no processo de ML num conjunto de dados referente ao perfil de atendimento dos profissionais de enfermagem na rede municipal de saúde de Curitiba. Descreveu-se o procedimento metodológico e a obtenção e carregamento dos dados.

O capítulo 4 apresenta a execução do experimento, comtemplando a preparação do conjunto de dados referente ao perfil de atendimento dos profissionais de enfermagem na rede municipal de saúde de Curitiba. Descreveu-se a limpeza dos dados, bem como as técnicas utilizadas para tal ação.

No capítulo 5, é apresentado os resultados do experimento, trazendo a análise exploratória do conjunto de dados referente ao perfil de atendimento dos profissionais de enfermagem na rede municipal de saúde de Curitiba. Descreve-se a análise dos dados, bem como as técnicas visualização dos dados.

O capítulo 6 apresenta as conclusões obtidas com a execução do trabalho, relacionando ao objetivo previamente definido, e sugere-se ideias para trabalhos futuros. Por fim, estão listados as referências utilizadas como base para este estudo e o apêndice, que contém os códigos gerados.

2 EMBASAMENTO TEÓRICO

Neste capítulo, foram descritos os conceitos relacionados ao estudo apresentado. Primeiramente, conceituou-se o termo Machine Learning, e na sequência o processo de preparação e análise exploratória de dados, bem como a aplicação de técnicas de ML na área da saúde, inclusive citando dois trabalhos correlatos. E também ferramentas, linguagens de programação e bibliotecas disponíveis e utilizados em ML.

2.1 MACHINE LEARNING

O aumento da quantidade de dados coletados referente às mais diversas atividades realizadas no cotidiano, e a dificuldade que permeia toda essa gama de informação gerada, conduz à procura por técnicas capazes de lidar com essa situação, bem como produzir conhecimento útil por meio dessas experiências. Assim, surge a necessidade da aplicação de Machine Learning, onde, segundo Taulli (2019), um computador pode aprender sem ser explicitamente programado. Em vez disso, ele vai ingerir e processar dados usando técnicas estatísticas sofisticadas.

> Para processar e obter informação útil a partir destes dados, é necessário automatizar diversas tarefas de coleta, processamento e análise de dados para tomada de decisão, uma vez que, devido ao grande volume de dados disponível, torna-se inviável realizar estas tarefas manualmente. Nesse contexto, surge a Inteligência Artificial, que visa simular o comportamento de um cérebro humano utilizando máquinas (ESCOVEDO e KOSHIYAMA, 2020).

ML como campo de estudo surgiu como um subcampo da IA, que se preocupava com métodos para melhorar o conhecimento ou desempenho de um agente inteligente ao longo do tempo, em resposta à experiência do agente no mundo, conforme Provost e Fawcett (2013). A evolução do poder de processamento dos computadores, aliado à possibilidade de armazenamento de enormes quantidades de dados possibilitaram o crescimento do ML. Avanços nessa área, bem como nas subáreas de IA resultaram em cada vez mais empresas, órgãos

públicos e instituições sem fins lucrativos empregarem IA e ML, como afirma Faceli et al. (2011).

No ramo empresarial, a aplicação de ML tem sido bem evidente, já que as empresas são capazes de gerar e armazenar grandes quantidades de dados que posteriormente poderão ser reutilizados para obter conhecimento. Machine Learning apresenta um aspecto fundamental para as organizações. Além do que conforme Taulli (2019), parece ser uma boa aposta que os volumes de dados continuarão a aumentar rapidamente.

Em ML, não é necessário que as características e padrões sejam encontrados por meio de rotinas manuais. Existem algoritmos capazes de executar essas tarefas com bastante precisão e muito rapidamente.

> Algoritmos de AM podem ser divididas em Preditivas e Descritivas. Em tarefas preditivas, algoritmos de AM são aplicados a conjuntos de dados de treinamento rotulados para induzir um modelo preditivo capaz de predizer, para um novo objeto representado pelos valores de seus atributos preditivos, o valor de seu atributo alvo. Modelos preditivos podem ser utilizados, por exemplo, para, a partir de seus sintomas, predizer o estado de saúde de um paciente. Nessas tarefas, em geral são utilizados algoritmos de AM que seguem o paradigma de aprendizado supervisionado. Em tarefas de descrição, ao invés de predizer um valor, algoritmos de AM extraem padrões dos valores preditivos de um conjunto de dados. Como não fazem uso do conhecimento do “supervisor externo”, esses algoritmos usam o paradigma de aprendizado não supervisionado (FACELI et al.,2011).

No que se refere a machine learning, o algoritmo é tipicamente diferente de um tradicional. A razão é que o primeiro passo é processar dados para, em seguida, o computador começar a aprender, assim afirma Taulli (2019). Além dos algoritmos, o ML possui também uma série de métodos, que permitem atingir o objetivo esperado.

Os modelos preditivos seguem o paradigma de aprendizagem supervisionada, onde destaca-se os métodos de regressão e classificação. E quanto aos modelos descritivos, a aprendizagem não supervisionada, onde destaca-se o agrupamento, associação e sumarização. Conforme Faceli et al. (2011), em tarefas de descrição ao invés de predizer um valor, algoritmos de ML extraem padrões dos

valores preditivos de um conjunto de dados, não fazem uso do conhecimento do “supervisor externo”.

Entre as descritivas o agrupamento de dados, que procura grupos de objetos similares entre si, e também regras de associação, que associam valores de um subconjunto de atributos preditivos a valores de outro subconjunto.

> As tarefas preditivas se distinguem pelo valor de rótulo a ser predito: discreto, no caso de tarefas de classificação; e contínuo, no caso de tarefas de regressão. As tarefas descritivas são genericamente divididas em: agrupamento, que dividem os dados em grupos de acordo com sua similaridade; sumarização, que buscam uma descrição simples e compacta para um conjunto de dados; e associação, que procuram padrões frequentes de associações entre os atributos de um conjunto de dados (FACELI et al.,2011).

Além de selecionar o algoritmo e o método adequado para a modelagem que se pretende realizar, outro ponto importante é a qualidade e quantidade do conjunto de dados disponíveis. Neste aspecto encaixam-se as etapas de preparação e análise exploratória de dados.

2.2 PROBLEMAS DE QUALIDADE DOS DADOS

Para a qualidade da aplicação de modelos de ML, e para garantia de um modelo eficiente, capaz de oferecer resultados pertinentes, é fundamental entender o problema a ser resolvido, bem como os dados com os quais se irá trabalhar. Segundo Faceli et al. (2011), apesar do crescente número de bases de dados disponíveis, na maioria das vezes não é possível aplicar algoritmos de ML diretamente sobre esses dados, é necessário aplicar técnicas para correção dos dados.

> Os bancos de dados do mundo real de hoje são altamente suscetíveis a dados ruidosos, ausentes e inconsistentes devido ao seu tamanho tipicamente enorme (geralmente vários gigabytes ou mais) e sua provável origem de fontes múltiplas e heterogêneas. Dados de baixa qualidade levarão a resultados de mineração de baixa qualidade (HAN; KAMBER; PEI, 2011).

Não somente para o ML, mas como para diferentes áreas de conhecimento, a qualidade de dados é um aspecto fundamental, desta forma tornou-se um importante objeto de estudo. O dado é considerado de qualidade se atende com precisão as expectativas para as quais será utilizado, no caso abordado nesse trabalho, a aplicação nos modelos de ML. Conforme Enap (2019), os dados são de alta qualidade, na medida em que atendem às expectativas e necessidades dos consumidores de dados e dependem, portanto, do contexto e das necessidades desses consumidores. Porém, esta não é a realidade.

> Os dados geralmente estão longe de ser perfeitos. Embora a maioria das técnicas de mineração de dados possa tolerar algum nível de imperfeição nos dados, o foco em entender e melhorar a qualidade dos dados geralmente melhora a qualidade da análise resultante. Os problemas de qualidade de dados que muitas vezes precisam ser resolvidos incluem a presença de ruído e valores discrepantes; dados ausentes, inconsistentes ou duplicados; e dados tendenciosos ou, de alguma outra forma, não representativos do fenômeno ou da população que os dados supostamente descrevem (TAN; STEINBACH; KUMAR, 2014).

Os padrões de qualidade de dados podem ser aplicados desde a coleta, como quando um dado é imputado em um sistema de computação, até a sua aplicação para análise da informação. Desta forma, a qualidade pode ocorrer por meio da prevenção, ou então da correção, portanto deve ser abordada em todo seu ciclo de vida. Os problemas de qualidade dos dados podem surgir em qualquer ponto do ciclo de vida dos dados, desde a criação até o descarte (ENAP, 2019).

Os problemas com a qualidade de dados podem ser observados de diferentes vertentes, e como já salientado, dependem de onde serão utilizados. Tan, Steinbach e Kumar (2014), afirmam que alguns dados começam a envelhecer assim que são coletados, em particular, se os dados fornecem um instantâneo de algum fenômeno ou processo contínuo, então esse instantâneo representa a realidade por apenas um tempo limitado.

Outro problema bastante comum é o viés de amostragem, que segundo Tan, Steinbach e Kumar (2014), ocorre quando uma amostra não contém diferentes tipos de objetos em proporção à sua ocorrência real na população. Por exemplo, os dados da pesquisa descrevem apenas aqueles que respondem à pesquisa.

O conhecimento sobre os dados também é um aspecto importante para a garantia de qualidade. Idealmente, os conjuntos de dados são acompanhados por documentação que descreve diferentes aspectos dos dados; a qualidade desta documentação pode auxiliar ou dificultar a análise subsequente (TAN; STEINBACH; KUMAR, 2014).

A má qualidade dos dados pode levar os modelos de aprendizagem a apresentar resultados ruins, que por sua vez irão influenciar a tomada de decisão. Outro problema é que esses dados podem implicar em gastos adicionais, já que os profissionais terão que dedicar algum tempo para a correção dos mesmos. E não obstante, a sujeira nos dados podem até mesmo apresentar prejuízos para as organizações, principalmente ao se tratar do setor privado.

> As organizações frequentemente superestimam a qualidade dos dados e subestimam as implicações de dados de baixa qualidade. As consequências de dados ruins podem variar de significativas a catastróficas. Os problemas de qualidade de dados podem fazer com que os projetos falhem, resultando em perda de receita e redução no relacionamento com o cliente, além da rotatividade do cliente. As organizações são rotineiramente multadas por não terem um processo de conformidade regulatória eficaz (GUDIVADA; APON; DING, 2017).

Jesilevska (2017), ressalta este fato dizendo que os dados de baixa qualidade podem implicar muitas consequências negativas, por exemplo, os custos operacionais, uma vez que tempo e outros recursos são gastos para detectar e corrigir erros.

Conforme Manjunath, et al. (2010) os fatores mais comuns que afetam a qualidade dos dados são:

- Procedimentos inadequados de manipulação de dados;
- Fraca relação com os dados originais;
- Erros na migração de dados;
- Dados não estruturados.

Dados sujos[1] podem causar confusão para o procedimento de mineração, resultando em uma saída não confiável. Além de que, como afirma HAN, KAMBER, PEI (2011), a maioria das ferramentas de mineração de dados precisa trabalhar em

[1] São conjuntos de dados com valores ausentes, com ruídos, fora do padrão comparado aos demais (HAN; KAMBER; PEI; 2011).

dados integrados, consistentes e limpos, o que requer limpeza de dados cara, integração de dados e transformação de dados como etapas de pré-processamento.

> Os dados sujos podem ser nomeados também como dados de baixa qualidade. Pode-se exemplificar como consequência de se utilizar dados de baixa qualidade dentro da empresa, como, a digitação incorreta de um nome que pode levar a duplicação de um cadastro da pessoa, com isto os relatórios passam a ter informações erradas, levando a uma tomada de decisão como o aumento do número de carteiras em uma sala de aula (MONTEIRO, 2017).

Em ML, a qualidade dos conjuntos de dados é essencial, pois os dados serão utilizados para treinar, validar e testar o modelo, e caso não estejam nos padrões esperados poderão interferir no resultado final. Alguns exemplos de problemas que podem ocorrer, citados por Gudivada, Apon, Ding (2017), valores discrepantes no conjunto de dados de treinamento podem causar instabilidade ou não convergência no aprendizado por conjunto. Dados incompletos, inconsistentes e ausentes podem levar a uma degradação drástica na previsão.

> As técnicas de pré-processamento de dados podem melhorar a qualidade dos dados, ajudando assim a melhorar a precisão e a eficiência do processo de mineração subsequente. O pré-processamento de dados é uma etapa importante no processo de descoberta de conhecimento, porque as decisões de qualidade devem ser baseadas em dados de qualidade. Detectar anomalias de dados, retificá-las antecipadamente e reduzir os dados a serem analisados pode gerar grandes recompensas para a tomada de decisões (HAN; KAMBER; PEI, 2011).

Uma questão que afeta consideravelmente a qualidade dos dados e pode gerar problemas no resultado da aplicação de ML, é o fato de que muitas vezes os dados não foram inicialmente gerados para o fim que estão sendo empregados. Basicamente, o analista de dados utiliza uma base de dados já existente, para concluir algo sobre esses dados. Myatt (2007) diz que, a relevância e a qualidade dos dados afetarão diretamente a precisão dos resultados. Em uma situação ideal, os dados foram cuidadosamente coletados para responder às questões específicas definidas no início do projeto.

2.3 PREPARAÇÃO DE DADOS

A qualidade dos dados é o fator mais importante para influenciar a robustez dos resultados de qualquer análise. Os dados devem ser confiáveis e representar a população-alvo definida (Myatt, 2007). Ao realizar uma primeira análise pode-se observar alguns fatores que precisam ser tratados, a fim de preparar e limpar os dados. Alguns pontos de atenção nesta fase são os listados abaixo:

> Unidades de medida: é importante sempre observar a unidade das variáveis para fazer as devidas conversões, se necessário. Valores faltantes: se um atributo tem muitos registros faltantes, ele não deve ser utilizado como entrada de um modelo sem um tratamento apropriado. Valores inconsistentes: é importante entender se os valores inconsistentes são valores inválidos ou outliers. Um exemplo de valor inválido seria encontrar valores negativos em um campo que só pode ser positivo, ou um texto quando se espera um número. Já outliers são valores fora do intervalo esperado. Intervalo dos valores: preste atenção no intervalo dos valores das variáveis. Atributos como salário podem conter valores de 0 a mais de meio milhão de dólares, por exemplo. Este grande intervalo pode ser um problema para alguns modelos, e pode ser necessário transformar esta coluna usando uma transformação logarítmica para reduzir o seu intervalo de valores. Já variáveis com valores em intervalos pequenos (exemplo: idades entre 50 e 55) provavelmente não serão uma informação útil para os modelos. Vale lembrar de que o intervalo de valores adequado varia de acordo com o domínio da aplicação (ESCOVEDO e KOSHIYAMA, 2020).

Os dados, por vezes, não foram inicialmente coletados com um objetivo já definido, então é importante dedicar um tempo para sua limpeza e correção. Conjunto de dados problemáticos, onde não são aplicadas as devidas tratativas aos problemas, podem não contribuir para um desempenho de qualidade do modelo de ML, para tanto aplicam-se técnicas de pré-processamento de dados. As quais, segundo Faceli et al. (2011), são frequentemente utilizadas para melhorar a qualidade dos dados por meio da eliminação ou minimização dos problemas. Esse processo é conhecido como Data cleaning.

> Data cleaning é o processo de melhorar a qualidade dos dados modificando sua forma ou conteúdo, por exemplo, removendo ou corrigindo valores de dados que estão incorretos. Esta etapa geralmente precede a etapa de modelagem, embora uma passagem pelo processo de mineração de dados possa indicar que uma limpeza adicional é desejada e pode sugerir maneiras de melhorar a qualidade dos dados (PROVOST; FAWCETT, 2013).

Há várias maneiras de organizar os dados, dentre elas, afirma Faceli et al. (2011) dados estruturados, organizados em planilhas ou bancos de dados relacional. Dados não estruturados que não possuem formatação pré-definida e dados semiestruturados, como XML. Pode-se dizer que a principal fonte de dados utilizadas para ML, são os datasets, usualmente disponibilizados em formato csv[2].

O pré-processamento dos dados exige habilidades do analista de dados que vão além do conhecimento técnico. Segundo Batista (2003), essa fase é um processo semiautomático que depende da capacidade do analista em identificar os problemas presentes nos dados, além da natureza desses problemas, bem como os métodos para solucioná-los.

Para Taulli (2019), é necessário a aplicação de 3 etapas nessa fase, compreensão do negócio, compreensão dos dados e preparação dos dados.

> Compreensão do negócio: É preciso ter uma visão clara do problema de negócio a ser resolvido. Compreensão dos dados: Nessa etapa, serão examinadas as fontes de dados para o projeto. Preparação dos dados A primeira etapa no processo de preparação é decidir quais conjuntos de dados (datasets) usar (TAULLI, 2019).

Basicamente, o pré-processamento dos dados irá envolver desde a obtenção dos mesmos, os quais podem estar fragmentados e surge a necessidade de integrá-los, passa pela etapa de eliminação das informações que não serão relevantes para o objetivo esperado, até a limpeza dos dados propriamente dita. Com base em uma categorização inicial das variáveis, pode ser possível remover as variáveis não pertinentes. Segundo Myatt (2007), uma análise de correlações entre várias variáveis pode identificar dados que não fornecem nenhuma informação adicional para a análise e, portanto, podem ser removidas.

[2] Uma das formas mais comuns de armazenar dados em uma base é através do formato .csv (Comma Separated Value), que consiste em dados estruturados em um formato de tabela com cabeçalho: os dados são organizados em linhas e colunas, e primeira linha traz os nomes das colunas. Cada coluna representa um fator ou medida diferente e cada linha representa uma instância ou exemplo (ESCOVEDO e KOSHIYAMA, 2020).

2.4 ANÁLISE EXPLORATÓRIA DE DADOS

Após aplicar as técnicas de preparação dos dados, já com os dados prontos e nos formatos adequados, o analista de dados pode seguir para a fase de exploração dos dados, por meio da geração de gráficos, histogramas e estatísticas que permitirão uma melhor compreensão dos dados, como a definição de padrões e, extração de resultados e considerações referente ao conjunto de dados. Segundo Escovedo e Koshiyama (2020), para aprender mais sobre os dados de forma mais fácil e intuitiva, é possível usar a visualização de dados, ou seja, a criação de gráficos a partir dos dados brutos.

A análise exploratória de dados não tem por finalidade confirmar conceitos pré-definidos, mas sim descobrir padrões e formular hipóteses para estudos futuros, por meio de técnicas de visualização ou então estatística. Para Faceli et al. (2011), a análise das características presentes em um conjunto de dados permite a descoberta de padrões e tendências que podem fornecer informações valiosas para compreender o processo que gerou os dados.

> Uma grande quantidade de informações úteis pode ser extraída de um conjunto de dados por meio de sua análise ou exploração. Informações obtidas na exploração podem ajudar, por exemplo, na seleção da técnica mais apropriada para pré-processamento dos dados e para aprendizado. Uma das formas mais simples de explorar um conjunto de dados é a extração de medidas de uma área da estatística denominada estatística descritiva. A estatística descritiva resume de forma quantitativa as principais características de um conjunto de dados (FACELI et al., 2011).

A análise de dados é bastante utilizada como base para o processo de tomada de decisões, para tanto são realizadas sumarizações dos dados a fim de obter informações importantes e examinar o conjunto de dados de diferentes ângulos, para que posteriormente seja possível aplicar algoritmos de modelagem sobre eles. Pode-se dizer que esse processo contribui para a compreensão do status atual da área estudada. Conforme Myatt (2007), a análise exploratória de dados e a mineração de dados abrangem um amplo conjunto de técnicas para resumir os dados, encontrar relacionamentos ocultos e fazer previsões. Alguns dos métodos comumente usados incluem:

> Tabelas de resumo: as informações brutas podem ser resumidas de várias maneiras e apresentadas em tabelas. Gráficos: apresentar os dados graficamente permite que o olho identifique visualmente tendências e relacionamentos. Estatísticas descritivas: são descrições que resumem as informações sobre uma determinada coluna de dados, como o valor médio ou extremo valores. Estatística inferencial: métodos que permitem que sejam feitas alegações sobre o dado com confiança. Estatísticas de correlação: estatísticas que quantificam as relações dentro dos dados. Pesquisando: Fazer perguntas específicas sobre os dados pode ser útil se você entender a conclusão que está tentando chegar ou se quiser quantificar qualquer conclusão com mais informações. Agrupamento: métodos para organizar um conjunto de dados em grupos menores que potencialmente respondem a perguntas. Modelo matemático: uma equação matemática ou processo que pode fazer previsões (MYATT, 2007).

Como já citado anteriormente, uma forma bastante comum de organização dos dados é por meio de tabelas, onde cada linha representa um grupo e cada coluna uma variável. Porém, apesar da facilidade desse formato, devido ao grande volume de dados disponíveis, não é possível chegar a alguma conclusão olhando somente os dados em sua forma bruta. Para tanto, é possível segundo Myatt (2007), aplicar uma sumarização dos dados, utilizando algumas estatísticas como média, mediana, soma, mínimo, máximo e desvio padrão.

A abordagem de análise de dados voltada à estatística tem por finalidade resumir as principais características de um conjunto de dados. Faceli et al. (2011), diz que essas medidas permitem capturar informações como: frequência; Localização ou tendência central (por exemplo, a média); Dispersão ou espalhamento (por exemplo, o desvio padrão); Distribuição ou formato.

> A medida de frequência é a mais simples. Ela mede a proporção de vezes que um atributo assume um dado valor em um determinado conjunto de dados. Ela pode ser aplicada a valores tanto numéricos quanto simbólicos, e é muito utilizada nesses últimos. Um exemplo de seu uso seria: em um conjunto de dados médicos, 40% dos pacientes têm febre. As outras medidas diferem para os casos em que os dados apresentam apenas um atributo (dados univariados[3]) ou mais de um atributo (dados multivariados[4]). Elas são geralmente aplicadas a dados numéricos. Apesar de a maioria dos conjuntos de dados utilizados em AM apresentar mais de um atributo, análises realizadas em cada atributo podem oferecer informações valiosas sobre os dados (FACELI et al, 2011).

As visualizações por meio de gráficos são uma forma eficiente de análise de dados, que permitem identificar tendências, faixas, distribuições de frequência,

---

[3] Para os dados univariados, supor que um objeto x_i possua apenas um atributo (FACELI et al., 2011).

[4] Dados multivariados são aqueles que possuem mais de um atributo de entrada (FACELI et al., 2011).

relacionamentos, outliers e fazer comparações (Myatt, 2007). Há algumas técnicas para representar graficamente todas essas informações.

Segundo Myatt (2007), poligramas de frequência e histogramas são utilizados para representar informações de acordo com o número de observações relatadas para cada valor (ou intervalos de valores) para uma determinada variável. Um histograma é uma ótima ferramenta para visualizar dados numéricos, podemos ver quantos modos há, bem como observar a distribuição (Harrison, 2020). Outra forma de visualizar resumidamente dados de um dataset são os gráficos de dispersão, ou então, scatterplotts.

> Gráficos de dispersão podem ser usados para identificar se existe alguma relação entre duas variáveis contínuas com base nas escalas de razão ou intervalo. As duas variáveis são plotadas nos eixos x e y. Cada ponto exibido no gráfico de dispersão é uma única observação. A posição do ponto é determinada pelo valor das duas variáveis (MYATT, 2007).

Para visualizar a distribuição em um conjunto de dados, pode-se utilizar o gráfico de caixa, ou box plots, que conforme Myatt (2007), fornecem um resumo sucinto da distribuição geral de uma variável. Cinco pontos são exibidos: o valor do extremo inferior, o quartil inferior, a mediana, o quartil superior, o extremo superior e a média.

Algumas ferramentas permitem ainda ao analista de dados gerar múltiplos gráficos de uma vez. De acordo com Myatt (2007), essa técnica é utilizada quando se pretende exibir vários gráficos ao mesmo tempo em formato de tabela, muitas vezes referido como uma matriz. Isso dá uma visão geral dos dados de vários ângulos.

A correlação é uma forma bastante útil de análise dos dados, onde é possível estabelecer comparações entre as colunas do dataset, onde é possível obter informações bastante relevantes. Harrison (2020), diz que colunas com alto grau de correlação não agregam valor e podem prejudicar a interpretação da importância dos atributos e dos coeficientes de regressão.

Há diferentes métodos e técnicas que permitem ao analista avaliar seu conjunto de dados. Selecionar o que melhor se adequa à sua necessidade é uma tarefa que exige certo conhecimento tanto técnico quanto do negócio. Faceli et al. (2011), afirma que a análise dos dados pode ser realizada por técnicas estatísticas e

de visualização, que permitirão uma melhor compreensão da distribuição dos dados e poderá dar suporte à escolha de formas de abordar o problema.

2.5 MACHINE LEARNING APLICADA A SAÚDE

A utilização de ML tem crescido consideravelmente em diversos setores, na saúde não poderia ser diferente, considerada a imensidão de dados gerados diariamente, capazes de possibilitar riquíssimas aprendizagens sobre eles. Segundo Faceli et al. (2011), é uma das áreas de aplicação em que algoritmos de ML têm maior receptividade e são utilizadas para controlar epidemias, auxiliar exames clínicos, monitorar pacientes, dosar medicamentos e dar suporte ao diagnóstico médico.

> Os algoritmos de AM podem ser empregados para aprender com dados históricos dos pacientes e, posteriormente, realizar previsões sobre diagnósticos. Para isso, os dados do paciente normalmente são processados, relacionados com anotações clínicas, correlacionados, associados a sintomas, antecedentes familiares, hábitos e doenças. Os impactos de certos fatores biomédicos, como a estrutura do genoma ou variáveis clínicas, também podem ser considerados para prever a evolução de certas doenças. As aplicações mais comuns envolvem o prognóstico de progresso ou prevenção de doenças para reduzir o risco e resultados negativos (FACELI et al., 2011).

Ao extrair conhecimento dos dados por meio da ML, é possível obter inúmeros benefícios, visto a precisão com que os mesmos são capazes de atuar. Conforme Faceli et al. (2011), para diversas doenças, a capacidade preditiva dos métodos de AM já vem superando a de muitos especialistas, como na detecção de câncer de mama e doenças cardíacas.

> Recentemente, com o advento de dispositivos vestíveis, como os relógios e pulseiras inteligentes, os médicos podem analisar dados biométricos continuamente coletados dos pacientes, e as técnicas de ML podem ser usadas para obter resultados clinicamente relevantes a partir desses dados. Medições simples, como a frequência cardíaca média em repouso, podem proporcionar uma rápida compreensão da condição cardíaca do paciente. Indicadores mais avançados, como índice de estresse, podem ser usados para prever arritmias cardíacas com mais precisão. Nesse sentido, os algoritmos de ML dão suporte para que os médicos analisem um enorme volume de dados em um curto espaço de tempo e, desta forma, acompanhem com mais precisão a saúde dos seus pacientes (FACELI et al., 2011).

Nesse contexto, pesquisadores tem desenvolvido estudos voltados à aplicação do ML a área da saúde, que vão desde a identificação de padrões, até a extração de conhecimento desses conjuntos de dados, em consonância com a possibilidade de mais agilidade em chegar a diagnósticos e até mesmo a otimização do tempo dos profissionais da saúde.

Campos Neto (2016), teve por objetivo criar por meio da técnica Desire[5] um modelo descritivo que classifique os pacientes quanto ao risco de ocorrência de eventos cardíacos adversos maiores e indesejáveis, e avaliar objetivamente seu resultado. O autor deste trabalho chegou à conclusão que os modelos têm grande potencial quando aliados à experiência do profissional.

Bertozzo (2019), aplicou técnicas de ML e análise exploratória em um conjunto de dados de filas de consulta do SUS para visualização dos resultados obtidos e características relevantes. Como resultado dos classificadores apresentados, o autor concluiu que é possível prever em 80% o perfil de pacientes que faltam às consultas.

2.6 FERRAMENTAS PARA PREPARAÇÃO E ANÁLISE EXPLORATÓRIA DE DADOS

A área de ciência de dados possui uma infinidade de ferramentas para lidar com os problemas que se propõe a resolver. Suas abordagens vão desde a obtenção, limpeza, preparação, pré-processamento, até a descoberta de padrões, características e aplicação de modelos de aprendizagem. Para tanto existem ferramentas que abstraem a etapa de codificação, como também aquelas que permitem o desenvolvimento de algoritmos com linguagem de programação, como as utilizadas nesse trabalho.

Em se tratando de linguagem para análise de dados, destaca-se o Python, vista como uma linguagem poderosa para processamento de dados. Conforme McKinney (2012), para muitas pessoas, é fácil se apaixonar pela linguagem Python. Desde sua primeira aparição em 1991, Python se tornou uma das linguagens de programação dinâmicas mais populares.

---

[5] Registro Desire é o registro mais longevo de cardiologia intervencionista mundial, unicêntrico, e acompanha por mais de 13 anos 6.377 pacientes revascularizados unicamente pelo implante de stents farmacológicos (CAMPOS NETO, 2016).

> Para análise de dados e computação exploratória interativa e visualização de dados, o Python inevitavelmente fará comparações com muitas outras linguagens e ferramentas de código aberto e de programação comercial específicas de domínio em amplo uso, como R, MATLAB, SAS, Stata e outros. Nos últimos anos, o suporte aprimorado de biblioteca do Python (principalmente pandas) o tornou uma alternativa forte para tarefas de manipulação de dados. Combinado com a força do Python em programação de uso geral, é uma excelente escolha como uma linguagem única para a construção de aplicativos centrados em dados (MCKINNEY, 2012).

Uma ferramenta de bastante destaque para trabalhar com Python voltado a data Science é o pacote NumPy, abreviação de Numerical Python, que é fundamental para a computação científica (MCKINNEY, 2012).

> NumPy é uma biblioteca Python que fornece um objeto de matriz multidimensional, vários objetos derivados (como matrizes e matrizes mascaradas) e uma variedade de rotinas para operações rápidas em matrizes, incluindo matemática, lógica, manipulação de forma, classificação, seleção, E/S, álgebra linear básica, operações estatísticas básicas, simulação aleatória e muito mais (NUMPY, 2021).

Além do Numpy, outras bibliotecas apresentam considerável impacto para o trabalho com dados. Uma delas é o Pandas, que, segundo McKinney (2012), fornece estruturas de dados e funções ricas projetadas para tornar o trabalho com dados estruturados rápido, fácil e expressivo. E um dos ingredientes essenciais para que o Python seja um ambiente de análise de dados poderoso e produtivo.

Conforme sua página na internet, pandas é uma biblioteca de código aberto licenciada por BSD que fornece estruturas de dados de alto desempenho e fáceis de usar (PANDAS, 2021).

> O pandas combina os recursos de computação de matriz de alto desempenho do NumPy com os recursos flexíveis de manipulação de dados de planilhas e bancos de dados relacionais (como SQL). Ele fornece funcionalidade de indexação sofisticada para tornar mais fácil remodelar, fatiar e dividir, realizar agregações e selecionar subconjuntos de dados (MCKINNEY, 2012).

Já para a visualização de dados, plotagens de gráficos e publicação de resultados, a biblioteca matplolib tem muito a oferecer. McKinney (2012) diz que é a biblioteca Python mais popular para a produção de plotagens e outras visualizações de dados 2D. Originalmente criado por John D. Hunter e agora mantido por uma grande equipe de desenvolvedores. Como citado no próprio site, Matplotlib é uma

biblioteca abrangente para a criação de visualizações estáticas, animadas e interativas em Python (MATPLOTLIB, 2021).

Em consonância com as bibliotecas acima citadas, pode-se ainda aplicar em projetos de ML, para publicação das informações o seaborn, que permite fazer gráficos estatísticos em Python e se baseia no matplotlib e se integra estreitamente às estruturas de dados do pandas.

Ao selecionar a linguagem e bibliotecas que se pretende utilizar, o analista de dados passa à escolha do ambiente para trabalhar. Dentre os notebooks computacionais para desenvolvimento de algoritmos de ML, Harrisson (2020) cita o Jupyter, segundo ele uma ferramenta excelente para fazer análise de dados exploratória que é, um ambiente de open source que aceita Python e outras linguagens. Ele permite criar células de código ou conteúdo Markdown.

> Jupyter Notebook é um aplicativo da web de código aberto que permite criar e compartilhar documentos que contêm código, equações, visualizações e texto narrativo. Os usos incluem: limpeza e transformação de dados, simulação numérica, modelagem estatística, visualização de dados, aprendizado de máquina e entre outros (JUPYTER, 2021).

Com o avanço das aplicações de ML, houve o surgimento de diversas ferramentas voltadas para esse fim. Algumas, requerem do analista de dados um conhecimento mais técnico de linguagens de programação, já outras funcionam de forma mais automatizada, e não demandam tanta habilidade técnica. Essas ferramentas abstraem a programação, como também seus recursos.

3 APRESENTAÇÃO DO DATASET DO PROBLEMA DE PERFIL DE ATENDIMENTO DE ENFERMAGEM NAS UPAS

Neste capítulo, são apresentados os passos para preparação e análise exploratória de dados no processo de ML num conjunto de dados referente ao perfil de atendimento dos profissionais de enfermagem na rede municipal de saúde de Curitiba. Descreve-se o procedimento metodológico e a obtenção e carregamento dos dados.

3.1 PROCEDIMENTOS METODOLÓGICOS

Foi aplicado o método de análise exploratória, o qual envolveu levantamento bibliográfico para maior conhecimento das técnicas de limpeza, manipulação e análise por meio de ML. Juntamente com uma abordagem quantitativa dos dados onde foram aplicadas algumas técnicas e apresentados os resultados por meio de estruturas de tabelas e gráficos.

Utilizou-se como fonte de informação o portal de dados abertos da cidade de Curitiba, onde são disponibilizados dados governamentais de domínio público para utilização e geração de conhecimento por parte da sociedade. Mais especificamente, a base utilizada contém informações de atendimentos de saúde. Ou seja, foi utilizada um conjunto de dados secundários, os quais já foram coletados e tabulados. Os dados abrangem o período de janeiro de 2020 a março de 2021.

Dados secundários são aqueles que já foram coletados, tabulados, ordenados, e, às vezes, até analisados, que estão catalogados à disposição dos interessados (MATTAR, 2005).

> Este sistema viabiliza o registro dos atendimentos prestados pela Secretaria Municipal de Saúde de Curitiba em sua rede de atenção. Esta rede é composta por Unidades Básicas de Saúde, Unidades de Pronto Atendimento, Centros de Especialidades Médicas e Odontológicas, entre outros. Os dados disponibilizados para consulta referem-se ao perfil de atendimento dos profissionais de enfermagem da rede municipal de saúde (CURITIBA, 2021).

Para este trabalho foi escolhida a linguagem de programação Python e foram utilizadas as bibliotecas Numpy, Pandas, Matplotlib e Seaborn, para preparação, análise exploratória e visualização dos dados por meio de gráficos. O

ambiente de desenvolvimento escolhido foi o Jupyter Notebook. Para execução das tarefas, foram consultados os comandos disponíveis em páginas da internet de cada uma das ferramentas. As bibliotecas já estavam disponíveis no Jupyter e foram importadas conforme necessidade.

3.2 COLETA DE DADOS

A coleta de dados em projetos de ML, é possível em diferentes fontes e formatos, a depender da situação que se deseja estudar. Em alguns casos, como na área empresarial, geralmente os registros são obtidos em suas próprias bases de dados. Porém, na internet há diversas fontes de dados, sobre os mais variados assuntos, como no caso do conjunto de dados utilizado neste trabalho. Foi realizado o download do dataset, disponível em formato CSV.

Juntamente ao arquivo com os dados, o portal de dados abertos da cidade de Curitiba, disponibiliza o dicionário de dados. Neste caso, o dataset conta com dados de atendimento de pacientes ocorridos em unidades de saúde da cidade. Consta nessa base, informações referentes ao problema de saúde apresentado, como também indicadores sociais. O conjunto de dados original é composto por 491.144 registros (linhas) e 42 variáveis (colunas), descritas no Quadro 1.

Quadro 1 – Dicionário de dados do Sistema E-Saude Enfermagem

Nome do Campo	Descrição	Tipo	Tamanho
Data do Atendimento	Data de Realização do Atendimento	DATE	
Data de Nascimento	Data de Nascimento do Paciente	DATE	
Sexo	Sexo do Paciente	VARCHAR2	1
Código do Tipo de Unidade	Código do Tipo de Unidade de Atendimento	NUMBER	5
Tipo de Unidade	Tipo de Unidade de Atendimento	VARCHAR2	50
Código da Unidade	Código da Unidade de Atendimento	VARCHAR2	150
Descrição da Unidade	Descrição da Unidade de Atendimento	VARCHAR2	80
Código do Procedimento	Código do Procedimento Realizado	VARCHAR2	12

Descrição do Procedimento	Descrição do Procedimento Realizado	VARCHAR2	255
Código do CBO	Código da Ocupação do Profissional	VARCHAR2	8
Descrição do CBO	Descrição da Ocupação do Profissional	VARCHAR2	200
Código do CID	Código do Diagnóstico	VARCHAR2	4
Descrição do CID	Descrição do Diagnóstico	VARCHAR2	150
Solicitação de Exames	Indica se ocorreu solicitação de Exames	VARCHAR2	3
Qtde Prescrita Farmácia Curitibana	Qtde de medicamentos prescritos na Farmácia Curitibana	NUMBER	10
Qtde Dispensada Farmácia Curitibana	Qtde de medicamentos dispensados na Farmácia Curitibana	NUMBER	10
Qtde de Medicamento Não Padronizado	Qtde de Medicamento Não Padronizado	NUMBER	10
Encaminhamento para Atendimento Especialista	Indica se houve encaminhamento para Atendimento de Especialista	VARCHAR2	3
Área de Atuação	Área de Atuação	VARCHAR2	255
Desencadeou Internamento	Indica se desencadeou Internamento	VARCHAR2	3
Data do Internamento	Data do Internamento do paciente	DATE	
Estabelecimento Solicitante	Estabelecimento que solicitou o internamento	VARCHAR2	80
Estabelecimento Destino	Estabelecimento que houve a internação	VARCHAR2	80
CID do Internamento	Código do diagnóstico do internamento	VARCHAR2	4
Tratamento no Domicílio	Tipo de Tratamento de Água no domicílio	VARCHAR2	30
Abastecimento	Tipo de Abastecimento de Água no domicílio	VARCHAR2	40
Energia Elétrica	Indica se há energia elétrica no domicílio	VARCHAR2	3
Tipo de Habitação	Tipo de habitação no domicílio	VARCHAR2	60
Destino Lixo	Destino do lixo no domicílio	VARCHAR2	30
Fezes/Urina	Destino das fezes/urina no domicílio	VARCHAR2	30

Cômodos	Qtde de Cômodos no domicílio	NUMBER	5
Em Caso de Doença	Serviços procurados em caso de doença	VARCHAR2	40
Grupo Comunitário	Grupo Comunitário em que o paciente participa	VARCHAR2	40
Meio de Comunicação	Meios de Comunicação utilizados no domicílio	VARCHAR2	40
Meio de Transporte	Meios de Transporte utilizados no domicílio	VARCHAR2	40
Município	Município do paciente	VARCHAR2	50
Bairro	Bairro do paciente	VARCHAR2	72
Nacionalidade	Nacionalidade do paciente (Brasileira, Naturalizado, Estrangeiro e Não informado)	VARCHAR2	20
cod_usuario	Código único do usuário	NUMBER	10
origem_usuario	1 - Residente no município 2 - não residente no município	NUMBER	1
residente	1 - Com cadastro definitivo na UBS 2 - sem cadastro definitivo na UBS	NUMBER	1
cod_profissional	Código único do profissional	NUMBER	10

Fonte: Curitiba (2021)

4 PREPARAÇÃO DO DATASET DO PROBLEMA PERFIL DE ATENDIMENTO DE ENFERMAGEM NAS UPAS

Neste capítulo, é apresentado a execução do experimento, comtemplando a preparação do conjunto de dados referente ao perfil de atendimento dos profissionais de enfermagem na rede municipal de saúde de Curitiba. Descreve-se a limpeza dos dados, bem como as técnicas utilizadas para tal ação.

4.1 PREPARAÇÃO E LIMPEZA DOS DADOS

Após as bibliotecas já importadas, realizou-se a leitura do dataset por meio do comando read_csv disponível no pandas, complementado pelos parâmetros sep, para definir o delimitados, no caso ponto e vírgula, e o encoding, que define a codificação a ser utilizada para leitura dos registros, nesse caso, latin-1. Os dados lidos foram convertidos em um dataframe para facilitar a manipulação.

Na sequência, foram aplicadas algumas funções do Python para compreender a situação do dataset, identificar dados com problemas e efetuar a preparação dos mesmos. O comando len foi utilizado para retornar a quantidade de linhas. Para identificação da quantidade de dados nulos em cada uma das variáveis usou-de o recurso isnull().sum(), onde foi identificado que algumas colunas possuíam uma grande quantidade de dados nulos. Para tanto aplicou-se o comando drop para excluir essas variáveis, já que pelo número elevado de dados faltantes não iriam beneficiar a análise. As variáveis excluídas foram as seguintes:

- Código do Tipo de Unidade
- Código da Unidade
- Código do Procedimento
- Código do CBO
- Qtde de Medicamento Não Padronizado
- Área de Atuação
- Desencadeou Internamento
- Data do Internamento
- Estabelecimento Solicitante
- Estabelecimento Destino

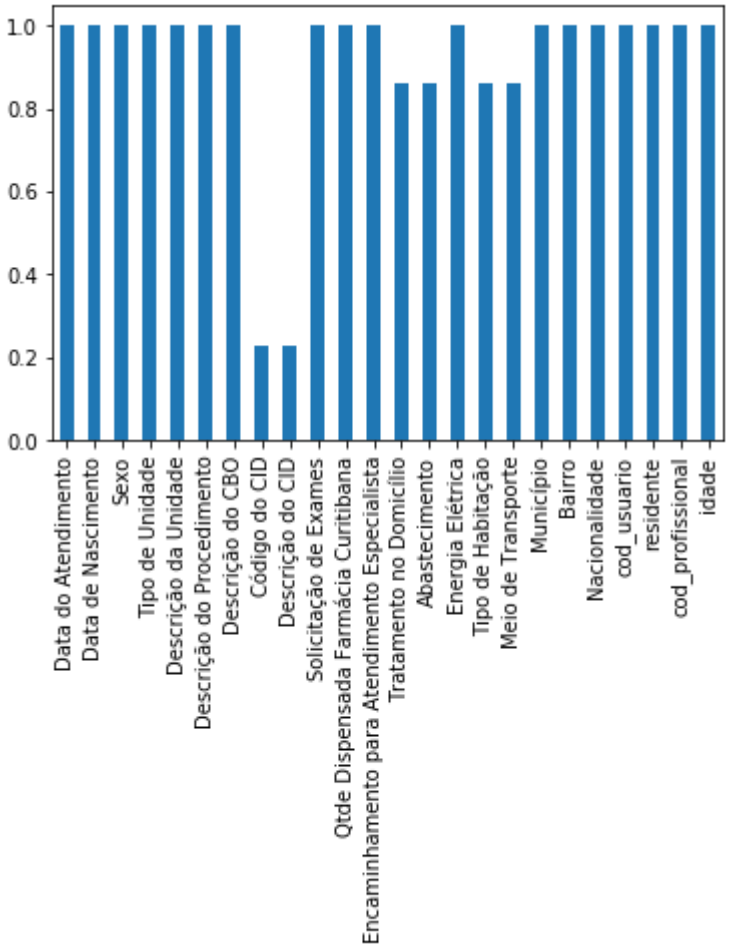
- CID do Internamento
- Fezes/Urina
- Em Caso de Doença
- Grupo Comunitário
- Meio de Comunicacao
- origem_usuario
- Código do Tipo de Unidade
- Qtde Prescrita Farmácia Curitibana
- Destino Lixo
- Cômodos

O método info foi utilizado para identificar se os tipos de cada coluna foram mantidos no dataframe conforme o dicionário de dados. Neste ponto, foi constatado que os campos de datas assumiram o tipo texto, para tanto foi aplicada a função to_datetime do pandas, que converteu os valores para o formato de data e hora, para posteriormente possibilitar a aplicação de técnicas de análise sobre esses dados. Criou-se uma coluna adicional no dataframe intitulada idade, onde realizou-se um cálculo de subtração data de atendimento e de nascimento do paciente. Utilizou-se o método rename para alterar o nome de algumas colunas, por exemplo, para corrigir a grafia do campo “Município”, onde constava “Municício”.

A função subplots do Matplotlib foi utilizada para apresentar o percentual de dados não ausentes em cada coluna, atrelado ao cálculo da média de campos nulos em cada variável. Como mostra o Gráfico 1, constatou-se que as variáveis Código do CID e Descrição do CID possuem um índice baixo de preenchimento. Como também as variáveis Tratamento no Domicílio, Abastecimento, Tipo de Habitação e Meio de Transporte, apresentaram percentual de preenchimento em torno de 80%.

Como foi observado que a taxa de preenchimento dos campos relacionados ao CID era baixa, foram criados 2 dataframes para possibilitar uma melhor análise dos dados. No dataframe1, os campos nulos para as colunas Código do CID e Descrição do CID foram preenchidos como “Desconhecido”, para que pudessem ser aplicadas técnicas de análise exploratória nos demais campos disponíveis. No dataframe2, foram excluídas as linhas cujo CID não estivesse informado, para que pudesse ser realizada uma análise sobre esse dado.

Gráfico 1 – Percentual de dados não ausentes por variável



Fonte: Elaborado pela autora (2021)

Depois disso, como ainda restaram algumas linhas com valores nulos no dataframe1, cerca de 20% dos registros, os quais foram excluídos, com o método dropna do pandas, utilizado para remover valores ausentes, e passado como parâmetro o comando any, o qual determina que se algum valor NA estiver presente, eliminar essa linha ou coluna.

Para possibilitar uma melhor visão dos atendimentos por idade, as mesmas foram agrupadas de 10 em 10 anos, e para as idades a partir de 90 foi criado um único grupo, visto que ocorrem com menor frequência no dataset. As idades iniciam em 0 e vão até 120 anos. Desta forma, usou-se o método cut do pandas para classificar as idades em segmentos, como parâmetro foram definidos os bins, definindo as bordas dos compartimentos, como abaixo:

- [0,10,20,30,40,50,60,70,80,90,120]

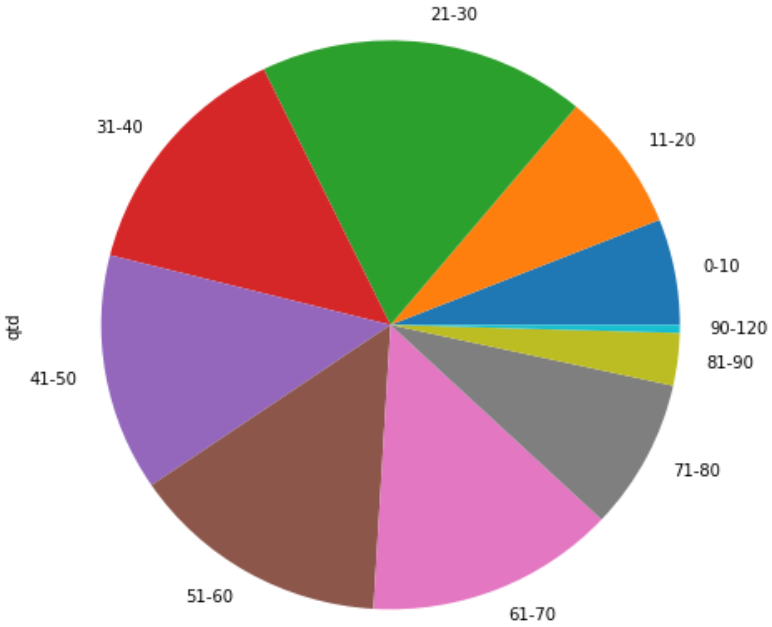
5 RESULTADO DA ANÁLISE EXPLORATÓRIA DE DADOS DO PROBLEMA DE PERFIL DE ATENDIMENTO DE ENFERMAGEM NAS UPAS

Neste capítulo, é apresentado os resultados do experimento, comtemplando a análise exploratória do conjunto de dados referente ao perfil de atendimento dos profissionais de enfermagem na rede municipal de saúde de Curitiba. Descreve-se a análise dos dados, bem como as técnicas visualização dos dados.

5.1 ANÁLISE EXPLORATÓRIA DOS DADOS

Após a fase de preparação dos dados, em que os mesmos já estão formatados e limpos, parte-se para o processo de exploração. Primeiramente, para representar as idades de forma visual realizou-se um group by por faixa etária, e então utilizou-se o comando plot.pie, que permite gerar gráficos em formato de pizza. Para tanto, aplicou-se a técnica de agrupamento, por meio da categorização[6] dividindo os dados em categorias. Como resultado, obteve-se o Gráfico 2, conforme abaixo.

Gráfico 2 – Distribuição de atendimento por faixa etária



Fonte: Elaborado pela autora (2021)

[6] A análise por categorização (clusterização), divide os dados em grupos (clusters) que são significativos, úteis ou ambos. Classes, ou grupos de objetos conceitualmente significativos que compartilham características comuns (TAN; STEINBACH; KUMAR, 2014).

No Quadro 2, é possível observar os agrupamentos efetuados, como também o total de atendimentos por faixa etária. Constatou-se que a faixa com mais frequência é de 21-30. E com menor frequência de 90-120. Os métodos mean, min e max, foram aplicados às idades e contatou-se que a idade média é 43,13 anos.

Quadro 2 – Quantidade de atendimentos por faixa etária

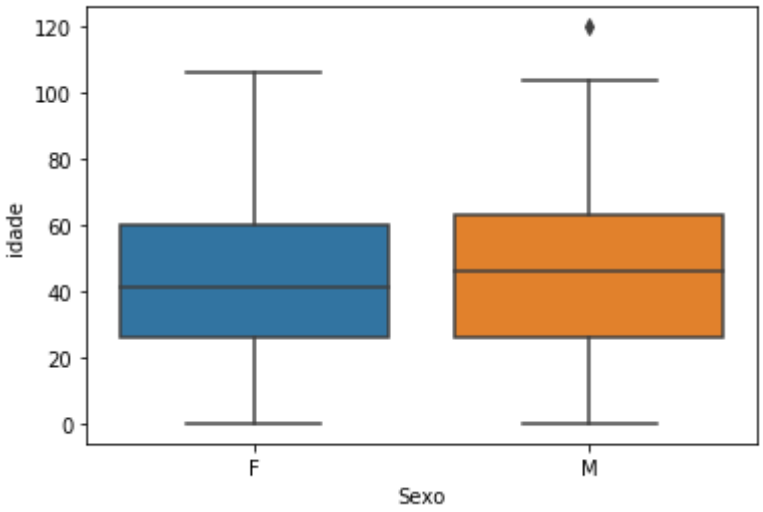
Faixa_etaria	Quantidade
0-10	20488
11-20	26862
21-30	62273
31-40	47072
41-50	45845
51-60	49430
61-70	47468
71-80	29130
81-90	10134
90-120	1542

Fonte: Elaborado pela autora (2021)

Para identificar idades fora do padrão, os chamados outliers, foi utilizado um boxplot, nesse caso, da biblioteca Searborn. Nota-se, que 25% dos dados, o primeiro quartil, encontra-se em torno de até 30 anos. Enquanto 50% variam na faixa de 30 a 60 anos, e os outros 25% acima de 60 anos. Nesse caso, foi observado que a mediana está em torno de 40 anos para o sexo feminino e um pouco acima para o sexo masculino. Encontrou-se outliers próximo do 120, como apresentado no Gráfico 3. Através do processo de categorização, por meio de agrupamento, foram identificados os outliers.

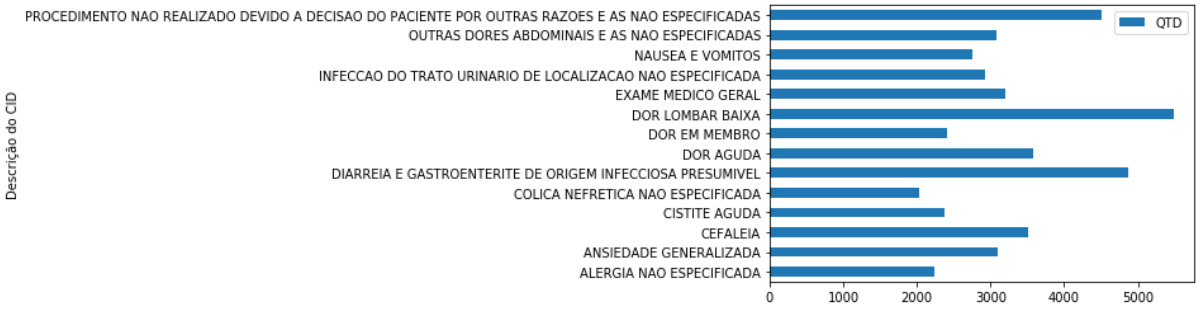
Realizou-se uma contagem de atendimento por CID, por meio do método group by, e considerados aqueles que mais apareceram, foi representada a frequência de atendimentos para cada CID, utilizando um recurso do pandas, o plot.barh, que apresenta os valores por meio de um gráfico horizontal, como é demonstrado no Gráfico 4Gráfico 4.

Gráfico 3 – Gráfico de caixa de idades entre sexo feminino e masculino



Fonte: Elaborado pela autora (2021)

Gráfico 4 – Quantidade de atendimento por CID com maior frequência



Fonte: Elaborado pela autora (2021)

Para avaliar a relação de cada CID com as unidades de atendimento, foi aplicado o método de crosstab do pandas, onde foi calculada a tabulação entre essas duas variáveis, como demonstrado no Quadro 3Quadro 4. Constatou-se com esse relacionamento, alguns fatores, por exemplo:

- Nas UPAS Campo Comprido, Pinheirinho e Tatuquara, o CID mais comum é “dor lombar baixa”;
- Na UPA Sítio Cercado, o CID mais comum é o “exame médico geral”;

Quadro 3 - Relação entre CID e UPAs

Descrição da Unidade / Descrição do CID	UPA BOA VISTA	UPA CAJURU	UPA CAMPO COMPRIDO	UPA CIDADE INDUSTRIAL	UPA PINHEIRINHO	UPA SITIO CERCADO	UPA TATUQUARA
ALERGIA NAO ESPECIFICADA	272	376	232	457	269	327	308
ANSIEDADE GENERALIZADA	407	479	363	497	401	561	390
CEFALEIA	345	561	412	652	442	580	522
CISTITE AGUDA	265	268	328	396	325	415	377
COLICA NEFRETICA NAO ESPECIFICADA	304	337	250	223	317	361	233
DIARREIA E GASTROENTERITE DE ORIGEM INFECCIOSA PRESUMIVEL	540	663	494	932	525	844	865
DOR AGUDA	363	720	354	778	363	643	351
DOR EM MEMBRO	201	360	272	562	363	316	342
DOR LOMBAR BAIXA	691	686	550	920	729	899	1012
EXAME MEDICO GERAL	272	543	341	244	276	1320	205
INFECCAO DO TRATO URINARIO DE LOCALIZACAO NAO ESPECIFICADA	352	630	314	548	389	373	317
NAUSEA E VOMITOS	306	392	337	565	353	491	317
OUTRAS DORES ABDOMINAIS E AS NAO ESPECIFICADAS	525	453	505	223	527	502	347
PROCEDIMENTO NAO REALIZADO DEVIDO A DECISAO DO PACIENTE POR OUTRAS RAZOES E AS NAO ESPECIFICADAS	817	550	372	718	605	1055	389

Fonte: Elaborado pela autora (2021)

Com relação aos CIDs foi gerado uma matriz de correlação entre as UPAS, criada com o método corr do pandas que é demonstrado no Quadro 4. Nesse caso, as cores contribuem para a observação do resultado, visto que quanto mais próximo do azul escuro, mais alto o nível de correlação, por exemplo a UPA Pinheirinho e Boa Vista possui o grau de correlação mais próximo de 1.

Quadro 4 – Correlação entre os pares da coluna UPA em relação ao CID

Descrição da Unidade / Descrição da Unidade	UPA BOA VISTA	UPA CAJURU	UPA CAMPO COMPRIDO	UPA CIDADE INDUSTRIAL	UPA PINHEIRINHO	UPA SITIO CERCADO	UPA TATUQUARA
UPA BOA VISTA	1	0.515104	0.682636	0.508798	0.901939	0.522868	0.572164
UPA CAJURU	0.515104	1	0.577685	0.680751	0.531766	0.534891	0.537098
UPA CAMPO COMPRIDO	0.682636	0.577685	1	0.472006	0.831899	0.451267	0.772095
UPA CIDADE INDUSTRIAL	0.508798	0.680751	0.472006	1	0.607408	0.247789	0.774191
UPA PINHEIRINHO	0.901939	0.531766	0.831899	0.607408	1	0.383038	0.776027
UPA SITIO CERCADO	0.522868	0.534891	0.451267	0.247789	0.383038	1	0.283914
UPA TATUQUARA	0.572164	0.537098	0.772095	0.774191	0.776027	0.283914	1

Fonte: Elaborado pela autora (2021)

Alguns indicadores sociais são cadastrados junto aos dados de atendimento do paciente, como os meios de transporte utilizados, conforme demonstrado no Quadro 5. Foi levantado que a grande parte utiliza como meio de transporte somente o ônibus, 109.509 utilizam ônibus e carro, e, 14.771 utilizam carro.

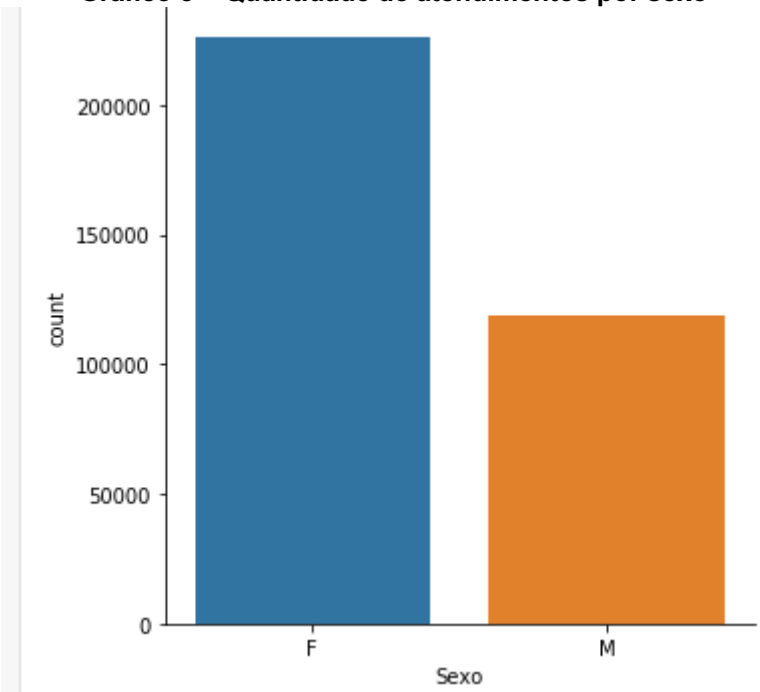
Quadro 5 – Contagem por meio de transporte

Meio de Transporte	Quantidade
ONIBUS	192144
ONIBUS,CARRO	109509
OUTROS	14878
CARRO	14771
OUTROS,ONIBUS,CARRO	6878
OUTROS,ONIBUS	2536
ONIBUS,CAMINHAO	1814

Fonte: Elaborado pela autora (2021)

Outro fator observado nos dados, foi a quantidade de atendimentos por sexo, neste caso utilizou-se um método de agrupamento, com o comando group by, onde foi constatado que do total de atendimentos analisados, 226.102 eram do sexo feminino, e 119.111 do sexo masculino. Para demonstrar de forma visual esses dados foi gerado um gráfico com o método factorplot do Seaborn, como mostrado no Gráfico 5. Utilizou-se um histograma de frequência agrupado por intervalos de classe, utilizando a técnica de categorização.

Gráfico 5 – Quantidade de atendimentos por sexo



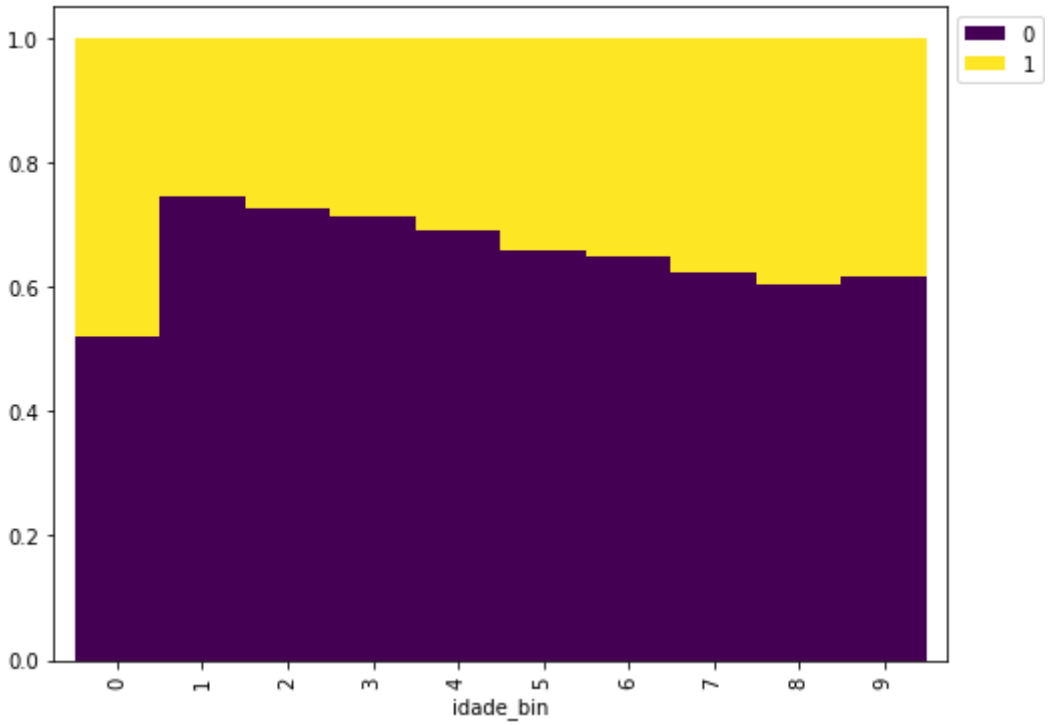
Fonte: Elaborado pela autora (2021)

A função subplots do Matplotlib foi utilizada, juntamente com o método cut do Pandas para comparar 2 categorias, no caso feminino e masculino. Para

aplicação desse método, foi necessário alterar o campo Sexo para numérico, utilizando o comando to_numeric. Foi aplicado um replace para alterar os valores, onde feminino assumiu valor 0, e masculino assumiu valor 1. As idades foram separadas em 10 quartis, para que se pudesse observar a distribuição dos atendimentos por sexo, para cada quartis, utilizando técnica de clusterização. O resultado é demonstrado no Gráfico 6.

É interessante observar que, por exemplo, quando se considera o atendimento de crianças, não ocorre grande variação entre sexo masculino e feminino. Fator comum no caso de atendimentos de jovens, adultos e idosos.

Gráfico 6 – Comparação por categoria sexo x idade



Fonte: Elaborado pela autora (2021)

5.2 RESULTADOS

Para que fosse possível realizar a análise exploratório do conjunto de dados de perfil de atendimento de enfermagem da cidade de Curitiba, foi de fundamental importância a execução das tarefas de limpeza e preparação dos dados. À primeira vista, o dataset utilizado não parecia conter muitos problemas com relação às informações contidas, no entanto, numa primeira análise, foram observados fatores que precisavam ser melhorados. Como os campos de CID e descrição CID, que não

estão preenchidos em alguns casos, cuja informação pode ser bastante relevante para a análise.

Após tratativa desses, foi possível aplicar técnicas de análise que permitiram obter informações relevantes quanto a padrões e características dos dados. O dataset original possui mais de 400.000 linhas, porém devido às sujeiras, dados, ausentes e outliers, algumas amostras precisaram ser descartadas, fato este que por sua vez poderia impactar no resultado final da aplicação de um modelo de ML.

O dataset escolhido é bastante rico, porém possui vários dados sociais do indivíduo, como meios de transporte, meios de comunicação, moradia, dentre outros. Algumas dessas colunas foram excluídas e em outras foram aplicadas técnicas de agrupamento, para se possibilitar uma melhor visão desses dados. Para aplicação de um modelo de ML com intuito de prever algum tipo de informação com relação aos atendimentos, o dataset necessita ser aprimorado, para resultados mais precisos.

6 CONCLUSÃO

A quantidade de dados gerados atualmente tem incentivado a crescente aplicação de técnicas de ML, já que essas informações podem ser utilizadas tanto para melhoria da qualidade de vida das pessoas, como por exemplo, quando aplicada à área saúde, ou então, no caso de organizações empresariais, onde os modelos de aprendizagem são usados com o intuito de apoiar a tomada de decisão, aumento de produtividade e campanhas de marketing mais eficientes.

Ter em mãos uma infinidade de dados não é suficiente, é importante que esses dados apresentem qualidade, para que possam ser aplicados nos algoritmos de ML. Porém, isso nem sempre é considerado desde o momento da obtenção da informação, até mesmo pelo fato de que muitas vezes o dado armazenado não apresenta o padrão esperado. Para tanto, as etapas iniciais de limpeza e análise são fundamentais para o processo, como se propôs neste trabalho.

O dataset utilizado embora já estivesse em um formato previamente adequado, apresentou alguns problemas que precisaram ser tratados, como os casos de dados faltantes, dados pessoais e limpeza de campos. Para tanto, fez-se uso de técnicas disponíveis no Python, baseado nos manuais disponíveis nas próprias página web das ferramentas utilizadas.

Os resultados obtidos na etapa de análise exploratória dos dados, demonstraram a importância da fase anterior de preparação do conjunto de dados. Onde foi possível obter informações relevantes sobre os registros disponíveis.

Para sequência da ideia abordada neste trabalho, sugere-se a aplicação de demais técnicas, como também em datasets maiores, desenvolvendo o código para tanto. Aconselha-se ainda o aprimoramento da base de dados utilizada, na possibilidade de incluir mais informações referentes aos atendimentos e agregar demais informações dos usuários atendidos pelo serviço.

REFERÊNCIAS

BATISTA, Gustavo Enrique de Almeida Prado Alves. **Pré-processamento de dados em aprendizado de máquina supervisionado**. 2003. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2003. Disponível em: <https://teses.usp.br/teses/disponiveis/55/55134/tde-06102003-160219/pt-br.php>. Acesso em: 29 mai. 2021.

BERTOZZO, Richard Junior. **Aplicação de Machine Learning em dataset de consultas médicas do SUS**. 2019. TCC (Graduação em Sistemas de Informação) - Universidade Federal de Santa Catarina. Centro Tecnológico, Florianópolis, Santa Catarina, 2019. Disponível em: < https://repositorio.ufsc.br/handle/123456789/202663>. Acesso em: 30 mai. 2021.

CAMPOS NETO, Cantídio de Moura. **Análise inteligente de dados em um banco de dados de procedimentos em cardiologia intervencionista**. 2016. Tese (Doutorado em Medicina/Tecnologia e Intervenção em Cardiologia) - Instituto Dante Pazzanese de Cardiologia, Universidade de São Paulo, São Paulo, 2016. Disponível em: <https://www.teses.usp.br/teses/disponiveis/98/98131/tde-18102016-085650/pt-br.php>. Acesso em: 29 mai. 2021.

CURITIBA. **Consulta de bases**: Sistema E-Saude - Perfil de atendimento de Enfermagem nas Unidades Municipais de Saúde de Curitiba. 2021. Disponível em: <https://www.curitiba.pr.gov.br/dadosabertos/busca/?pagina=7>. Acesso em: 16 mai. 2021.

ENAP. **Gerenciamento de Metadados e da qualidade de Dados.** Governança de Dados. Brasília, 2019. Disponível em: <https://repositorio.enap.gov.br/bitstream/1/5008/4/M%C3%B3dulo%204%20-%20Gerenciamento%20de%20Metadados%20e%20da%20qualidade%20de%20Dados.pdf>. Acesso em: 20 out. 2021.

ESCOVEDO, Tatiana; KOSHIYAMA, Adriano. **Introdução a Data Science**: Algoritmos de Machine Learning e métodos de análise. Casa do Código. 2020

FACELI, Katti; LORENA, Ana Carolina; GAMA, João; CARVALHO, André Carlos Ponce de Leon Ferreira de. **Inteligência artificial:** uma abordagem de aprendizado de máquina. 2. ed. Rio de Janeiro: LTC, 2021.

GUDIVADA, Venkat N.; APON, Amy; DING, Junhua. **Data Quality Considerations for Big Data and Machine Learning**: Going Beyond Data Cleaning and Transformations. International Journal on Advances in Software 10.1, 2017, p. 1 - 20. Disponível em: <https://www.researchgate.net/publication/318432363_Data_Quality_Considerations_for_Big_Data_and_Machine_Learning_Going_Beyond_Data_Cleaning_and_Transformations>. Acesso em: 30 mai. 2021.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data mining**: concepts and techniques. 3. ed. Waltham, MA, 2011.

HARRISON, Matt. **Machine Learning**: Guia de Referência Rápida. São Paulo: Novatec Editora, 2020.

JESILEVSKA, Svetlana. **Data Quality Dimensions to Ensure Optimal Data Quality**. The Romanian Economic Journal. v. 20, n. 63, mar. 2017. p. 89-103. Disponível em: <http://www.rejournal.eu/sites/rejournal.versatech.ro/files/articole/2017-04-02/3443/6jesilevska.pdf>. Acesso em: 29 mai. 2021.

JUPYTER. **Documentation**. 2021. Disponível em: <https://jupyter.org/documentation>. Acesso em: 06 jun. 2021.

MANJUNATH T.N, R. S., Ravindra S, & Ravikumar G.K. (2010). **Analysis of Data Quality Aspects in DataWarehouse Systems.** International Journal of Computer Science and Information Technologies, 2. Disponível em: <https://www.researchgate.net/publication/230639906_Analysis_of_Data_Quality_Aspects_in_DataWarehouse_Systems>. Acesso em: 20 out. 2021.

MATTAR. F. N. **Pesquisa de Marketing**: Metodologia e Planejamento. 6 ed. São Paulo: Atlas, 2005.

MATPLOLIB. **Overview**. 2021. Disponível em: <https://matplotlib.org/stable/contents.html>. Acesso em: 06 jun. 2021.

MCKINNEY, Wes. **Python for Data Analysis**. Gravenstein Highway North, Sebastopol: O'Reilly Media, 2012.

MONTEIRO, Cláudio Lopes. **Modelo para classificação de dados sujos em uma base de dados estruturada.** Instituto doctum de educação e tecnologia. Faculdades

Integradas de Caratinga. Caratinga. 2017. Disponível em: < https://dspace.doctum.edu.br/handle/123456789/410>. Acesso em: 20 out. 2021.

MYATT, Glenn J**. Making Sense of Data I**: A Practical Guide to Exploratory Data Analysis and Data Mining. 2 ed. New Jersey: Wiley, 2007.

NUMPY. **NumPy v1.21 Manual**. 2021. Disponível em: <https://numpy.org/doc/stable/>. Acesso em: 05 jun. 2021.

PANDAS. **Pandas documentation**. 2021. Disponível em: <https://pandas.pydata.org/docs/#pandas-documentation>. Acesso em: 05 jun. 2021.

PROVOST, Foster; FAWCETT, Tom. **Data Science for Business**. Gravenstein Highway North, Sebastopol: O'Reilly Media, 2013.

PYTHON. **Python 3.9.5 documentation**. 2021. Disponível em: <https://docs.python.org/3/>. Acesso em: 06 jun. 2021.

SEARBORN. **User guide and tutorial**. 2020. Disponível em: <https://matplotlib.org/stable/contents.html>. Acesso em: 06 jun. 2021.

TAN, Pang-ning; STEINBACH, Michael; KUMAR, Vipin. **Introduction to Data Mining**. Edinburgh Gate, Harlow: Pearson Education Limited, 2014.

TAULLI, Tom. **Introdução à Inteligência Artificial:** Uma abordagem não técnica. São Paulo: Novatec Editora, 2019.

APÊNDICE A - Códigos utilizados para preparação e análise exploratória dos dados

```
#importação das bibliotecas utilizadas
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

#leitura do dataset em formarto CSV
df = pd.read_csv('../dados/2021-05-06_Sistema_E-Saude_Enfermagem_-_Base_de_Dados.csv', sep=';', encoding='latin-1')

#verificar o tamanho do dataset
len(df)

#verificar quantidade de dados nulos em cada coluna
df.isnull().sum()

#exclusão de colunas com muitos dados faltantes
df = df.drop(['Código do Tipo de Unidade',
              'Código da Unidade',
              'Código do Procedimento',
              'Código do CBO',
              'Qtde de Medicamento Não Padronizado',
              'Área de Atuação',
              'Desencadeou Internamento',
              'Data do Internamento',
              'Estabelecimento Solicitante',
              'Estabelecimento Destino',
              'CID do Internamento',
              'Fezes/Urina',
              'Em Caso de Doença',
              'Grupo Comunitário',
              'Meio de Comunicacao',
              'origem_usuario',
              'Código do Tipo de Unidade',
              'Qtde Prescrita Farmácia Curitibana',
              'Destino Lixo',
              'Cômodos'
             ], axis=1)
```

```
#exibir informações sobre as colunas do dataset
df.info()

#converter as colunas de datas de string para datetime
df['Data do Atendimento'] = pd.to_datetime(df['Data do Atendimento'])
df['Data de Nascimento'] = pd.to_datetime(df['Data de Nascimento'])

#adicionar coluna idade
df['idade'] = df['Data do Atendimento'].dt.year - df['Data de Nascimento'].dt.year

#correção de nome da coluna
df = df.rename(columns={'Municício': 'Município'})

#gerar gráfico de percentual de dados não ausentes por variável
fig, ax = plt.subplots(figsize=(6, 4))
(1 - df.isnull().mean()).abs().plot.bar(ax=ax)

#preenchimento de dados ausentes
df_descarte['Código do CID'] = df_descarte['Código do CID'].fillna("Desconhecido")
df_descarte['Descrição do CID'] = df_descarte['Descrição do CID'].fillna("Desconhecido")
df_descarte['Tipo de Habitação'] = df_descarte['Tipo de Habitação'].fillna("NAO INFORMADO")
df_descarte['Meio de Transporte'] = df_descarte['Meio de Transporte'].fillna("NAO INFORMADO")

#exibir linhas com dados ausentes
df_descarte[df_descarte.isna().any(axis=1)]

#limpeza de linhas com dados ausentes
df_descarte = df_descarte.dropna(how='any')

#gerar dataset com linhas preenchidas pelo CID
df_cid = df.groupby(['Descrição do CID']).size().reset_index(name='QTD')
df_cid2 = df_cid[df_cid['QTD'] > 2000]
df_cid3 = df_cid2.drop(df_cid2[df_cid2['Descrição do CID'] == "Desconhecido"].index)
lista = list(df_cid3['Descrição do CID'])
cid = df['Descrição do CID'].isin(lista)
dffinal_cid = df[cid]

#adiciona coluna de faixa etária
faixa_etaria = pd.cut(df_descarte.idade,bins=[0,10,20,30,40,50,60,70,80,90,120],
```

```
                labels=['0-10','11-20','21-30','31-40','41-50','51-60','61-70',
                    '71-80','81-90','90-120'])
df_descarte.insert(24,'Faixa_etaria',faixa_etaria)

#gera gráfico de pizza de faixa etária
df_idade = df_descarte.groupby('Faixa_etaria').size().rename('qtd')
df_idade.plot.pie(figsize=(8, 8))

#exibe a média, mínimo e máximo entre as idades
print('Média', df_descarte['idade'].mean(), '\nMínimo', df_descarte['idade'].min(), '\nMáximo', df_descarte['idade'].max())

#gera gráfico de caixa de idades e sexo
sns.boxplot(x="Sexo", y="idade", data=df_descarte)

#agrupa os CID e exibe a contagem
dffinal_cid_qtd = dffinal_cid.groupby(['Descrição do CID']).size().reset_index(name='QTD')

#gera gráfico de linhas horizontal por CID
dffinal_cid_qtd.plot.barh(x='Descrição do CID', y='QTD')

#gera relacionamento entre CID e UPAs
df_similar = pd.crosstab(dffinal_cid['Descrição do CID'], dffinal_cid['Descrição da Unidade'])

#gera correlação entre as UPAs
df_corr = df_corr.style.background_gradient(cmap='RdBu')

#gera gráfico de linhas vertical de atendimentos por sexo
sns.factorplot('Sexo', data=df_descarte, kind='count')

#substitui o valor da coluna sexo por 0 e 1
df_descarte['Sexo'] = df_descarte['Sexo'].str.replace('F','0')
df_descarte['Sexo'] = df_descarte['Sexo'].str.replace('M','1')

#converte a coluna sexo para numérico
df_descarte["Sexo"] = pd.to_numeric(df_descarte["Sexo"])

#compara 2 categorias, no caso feminino e masculino, quanto as classes de idade geradas
fig, ax = plt.subplots(figsize=(8, 6))
```

```
(df_descarte.assign(idade_bin=pd.qcut(df_descarte.idade, q=10, labels=False ),
sexo_bin=pd.cut(df_descarte.Sexo, bins=2, labels=False),).pipe(lambda df:
pd.crosstab( df.idade_bin, df.sexo_bin) ).pipe(lambda df: df.div(df.sum(1), axis=0))
.plot.bar(stacked=True, width=1, ax=ax, cmap="viridis", )
.legend(bbox_to_anchor=(1, 1)))
```