
**PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ**

**BIG DATA E INTELIGÊNCIA ANALÍTICA**

**DENISE TORQUATO**

**DOUGLAS BRANCO PESSANHA LOPES**

**RAFAEL FRANCO SILVA PINTO**

# OFICINA MAKER

## ATIVIDADE SOMATIVA 2

**CURITIBA**

**2022**

**DENISE TORQUATO**

**DOUGLAS BRANCO PESSANHA LOPES**

**RAFAEL FRANCO SILVA PINTO**

# OFICINA MAKER

## ATIVIDADE SOMATIVA 2

Trabalho de atividade prática da disciplina Oficina Maker à Pontifícia Católica Universidade do Paraná, como requisito para conclusão da matéria.

Orientador: Prof. Galbas Milleo Filho.

**CURITIBA**
**2022**

## RESUMO

A função de um cientista de dados engloba tarefas deste a análise de dados, até o processamento e criação de um modelo preditivo com objetivo de encontrar uma solução para um problema específico. Neste trabalho foram exploradas cada uma destas tarefas e apresenta uma abordagem para a solução de um problema social. Utilizado como base do projeto o dataset Sistema E-Saúde Médicos, abordando dados compostos pela rede de Unidades Municipais de Saúde de Curitiba, teve o objetivo de classificar os principais tipos de doenças afetados pela população e a sua relação com o saneamento básico e qualidade de moradia dos entrevistados. Para este objetivo foi utilizado técnicas de análise de dados, como Pandas e SQL, juntamente com exploração de Big Data utilizando Pyspark e técnicas de Machine Learning.


**Palavras-chave**: Big Data. Ciência de dados. Machine Learning. Spark. Pandas.

## LISTA DE FIGURAS

**SUMÁRIO**

# 1 INTRODUÇÃO

Uma das funções mais requisitadas para um cientista de dados atualmente é a habilidade de explorar grandes volumes de dados, também chamado de Big Data. O Big Data e as análises compostas a partir deste, como mencionado por (SAGIROGLU; SINANC, 2013), estão no centro dos negócios e da ciência moderna, com a expansão constante da tecnologia, já vivemos em um mundo onde é praticamente impossível viver sem criar mais dados. O Big Data trabalha com uma enorme, variada e complexa estrutura de dados, que apresenta desafios particulares para o armazenamento, análise e visualização até a chegada da conclusão de um resultado. Estes dados podem ser gerados via transações online, e-mails, pesquisas, vídeos, áudios, sensores e telefones e suas aplicações. Com esta base de dados sempre crescente fica inviável a utilização de técnicas convencionais para a exploração dos dados, o que torna o Big Data essencial.

Com a popularização da linguagem de programação Python, especialmente no campo da ciência de dados e Big Data, também há um crescimento de novas e gratuitas bibliotecas disponíveis para a utilização, conforme (STANČIN; JOVIĆ, 2019). Neste trabalho foram exploradas técnicas de análise de dados como, SQL e Pandas (utilizando o framework Pypandas), SparkML e biblioteca MLlib para modelo preditivo e PySpark para análise de Big Data.

## 2 OBJETIVOS

Propõe-se a implementação de ferramentas e bibliotecas encontradas para a utilização junto com a linguagem de programação Python para fazer uma análise dos dados referentes à atendimentos prestados pela Secretaria Municipal de Saúde de Curitiba, onde engloba as Unidades Básicas de Saúde, Unidades de Pronto Atendimento, Centros de Especialidades Médicas e Odontológicas, entre outros, coletados em um período de três meses. O dataset é chamado de "2022-04-06 Sistema E-Saude Medicos-Base de Dados"e foi coletado no site da Prefeitura de Curitiba.

Acreditamos que é possível melhorar a qualidade de vida das pessoas utilizando as ferramentas de big data. Sendo assim, nossa motivação ao escolher este dataset veio da oportunidade de podermos analisar os dados referentes a este grupo e assim, além de classificar as doenças que mais afetam as pessoas que moram nesta região de Curitiba, e com isso ajudar na diminuição de enfermos. Com a análise destes dados podemos adicionar mais um nível de conscientização sobre estas doenças e também deixar claro para os governantes a prioridade das ações que podem ser tomadas para o combate dessas doenças.

Com o decorrer das semanas houve uma alteração no objetivo da atividade. Passou-se a buscar saber se um paciente será internado ou não com base em algumas informações fornecidas sobre ele. A princípio procuraríamos descobrir se havia alguma correlação entre determinados atributos do dataset e as doenças mais encontradas. Entretanto, devido aos problemas na máquina virtual do curso, optamos por executar a atividade na máquina local dos estudantes envolvidos na disciplina, realizando assim a execução de um código que utiliza de menos recursos computacionais.

# 3 REVISÃO DE LITERATURA

## 3.1 HADOOP E SPARK

Hadoop é um software *open-source* que permite o acompanhamento e manipulação de dados de Big Data através de uma rede de computadores, chamados de nós, para solucionar intrigantes problemas com dados. Segundo (WHITE, 2012), Hadoop é um software altamente escalável, com um custo reduzido que processa dados não-estruturados e semiestruturados, como por exemplo, dados de e-mails, sensores de IoT e dados de redes sociais. Seus maiores benefícios são: proteção de dados à uma eventual falha de hardware, vasta escalabilidade e possível análise dos dados em tempo real para uma melhor tomada de decisão.

O Apache Spark, que também é uma ferramenta *open-source*, é utilizado para processar os dados de um dataset de Big Data. Assim como visto pelo Hadoop, o Spark também divide as tarefas por vários nós, entretanto, tende-se a ter uma melhor performance em relação ao Hadoop. Esta vantagem tem-se pois o Apache Spark possui a capacidade de processar os dados previamente armazenados, chegando a ser até 100 vezes mais rápido quando executando em memória e 10 vezes mais rápido quando executando em disco.

## 3.2 PANDAS

Pandas é uma biblioteca para a linguagem de programação Python. Assim como as outras bibliotecas utilizadas neste projeto Pandas é *open-source*, sendo uma biblioteca rica em ferramentas para utilização em tarefas envolvendo ciência/análise de dados e aprendizagem de máquina. Assim como dito em (MCKINNEY et al., 2011), Pandas foca em ser uma camada fundamental para o futuro da computação estatística em Python.

## 3.3 SPARK ML

Segundo (MENG et al., 2016), Spark ML é uma popular plataforma *open-source* para o processamento em larga escala utilizando interatividade com aprendizagem de máquina.

Uma biblioteca essencial para esta atividade, derivada do Spark ML, é a MLlib. Segundo o mesmo autor, esta biblioteca de aprendizagem de máquina fornece funcionabilidades muito eficientes para tarefas envolvendo procedimentos de estatística, otimização e álgebra linear, e fazendo parte do ecossistema do Spark ML faz ser prática a sua utilização.

## 3.4 COMPARAÇÃO DAS TÉCNICAS

Como podemos perceber, cada uma destas três das principais técnicas utilizadas neste projeto, e adquiridas ao longo do curso, sobressai em uma determinada área do campo de Big Data. Sendo assim, essas técnicas são utilizadas em diferentes etapas de um projeto envolvendo dados, porém sempre trabalhando cooperativamente.

A princípio, a utilização do Hadoop e Spark faz possível a mineração de grandes volumes da dados, após, técnicas utilizando a biblioteca Pandas ajuda na análise e permite que se crie tabelas e gráficos para facilitar a visualização dos *insights* , tanto da equipe que está trabalhando diretamente no projeto, quanto à terceiros. Na etapa final são utilizadas as bibliotecas de *machine learning*, neste caso a SparkML. Deste modo podemos além de saber o que está ocorrendo, pela análise dos dados, também criamos predições do que pode vir a acontecer.

# 4 MATERIAIS E MÉTODOS

## 4.1 PLANEJAMENTO

Este projeto teve início na data de 25 de abril de 2022, sendo utilizado a primeira semana para a formação do grupo e entendimento do planejamento das atividades e seleção do dataset a ser utilizado. Também, em paralelo, foi feita a aprovação do dataset com o professor.

A partir da segunda semana, o grupo já tinha se encontrado em reunião *on-line* e definido as tarefas a serem completadas da semana. Estas sendo a análise do dataset escolhido, e a definição do objetivo do projeto, juntamente com as ferramentas a serem utilizadas.

Na terceira semana, o grupo se dividiu para fazer a análise exploratória dos dados. O maior desafio desta etapa foi quanto à utilização ou não de ferramentas de SQL juntamente com o Spark para a exploração de big data. Enfim foi escolhido a utilização da biblioteca Pandas, já que com ela é possível chegar à resultados similares do que com SQL e ainda tem a vantagem da facilidade devido à compatibilidade de *softwares*.

Durante a quarta semana foram feitas as análises dos dados restantes e colhidos os insights das atividades das semanas anteriores. Em paralelo foi planejado e escrito o relatório parcial, para ser entregue como a atividade somativa 1, feito pelo Rafael Franco, os insights das medidas de métricas de estatísticas descritiva obtidos pela Denise Torquato e tanto a preparação inicial do dataset, quanto a remoção dos outliers foram tarefas realizadas pelo Douglas Branco.

Devido à problemas relacionados à máquina virtual da PUCPR, a data para a entrega do relatório da atividade somativa 1 foi alterada para a semana seguinte. Seguindo este prazo, o relatório parcial, referente à atividade somativa 1, foi entregue durante a quinta semana da matéria Oficina Maker.

Na semana 5, ao verificar a dificuldade de se executar a atividade inicialmente proposta, o grupo decidiu alterar o objetivo da atividade. Para isso, foram feitas algumas mudanças no código implementado e consequentemente nos dados gerados.

Com a alteração no cronograma, as fases de modelagem e treinamento passaram para as semanas 6 e 7. Foi feita a modelagem necessária para correta utilização

dos dados, além do treinamento dos modelos.

Após os resultados obtidos na modelagem e treinamento dos dados, foram feitas algumas otimizações no modelo para que o código pudesse funcionar perfeitamente. Com os resultados obtidos, foi realizada a integração dos dados e preparação deste relatório final da atividade.

## 4.2 RESCURSOS HUMANOS

A equipe para a realização do projeto é composta por três integrantes e o dataset E-Saúde Médicos já foi validado pelo professor através do fórum.

## 4.3 DESCRIÇÃO MACRO DA SOLUÇÃO

Este projeto tem o objetivo maior de ajudar a reduzir a quantidade de pessoas acometidas por doenças na cidade de Curitiba. Para isso é feita a análise dos dados retirados do dataset fornecido através do site da prefeitura e realizado um estudo para identificar as principais doenças identificadas nos últimos 3 meses e quais suas relações com o tipo e qualidade de moradia e saneamento básico dos moradores. Acreditamos que com isso será possível criar uma maior conscientização da população e também dos políticos relacionados à cidade para assim ter um objetivo claro dos principais pontos a serem melhorados para beneficiar a saúde da população.

As tecnologias a serem utilizadas neste projeto são: Hadoop, Spark, PySpark, Jupyter Notebook e Python; junto com bibliotecas: Pandas, Spark SQL, Spark ML e Pandas.

Os prazos de entregas então alinhados com a tabela da Figura 1, assim sendo, até o final da semana 4 foi realizado o planejamento, coletado dos dados e preparação e análise dos dados.

Para a segunda entrega, planejada para o final da semana 8, será feita as demais etapas, incluindo: modelagem e treinamento, otimização do modelo e a integração.

Figura 1 – Tabela apresentando o cronograma do projeto

| | Etapas | Semana 1 | Semana 2 | Semana 3 | Semana 4 | Semana 5 | Semana 6 | Semana 7 | Semana 8 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Planejamento | | | | | | | | |
| 2 | Coleta dos dados | | | | | | | | |
| 3 | Preparação e Análise dos Dados | | | | | | | | |
| 4 | Modelagem e Treinamento | | | | | | | | |
| 5 | Otimização do Modelo | | | | | | | | |
| 6 | Integração | | | | | | | | |

Fonte: Os autores

## 4.4 DESCIÇÃO DO FORMATO DOS DADOS E RISCOS DO PROJETO

O *dataset* selecionado é um *dataset* semiestruturado, formatado como um CSV (*Comma-separated values*).

Quanto aos riscos do projeto acreditamos que, como nossa equipe é composta por três pessoas, não esperamos ter dificuldades neste quesito. Apesar disso, a exploração e utilização de técnicas de big data é um do desafio que estamos superando ao realizar esta atividade, pois nossa experiência neste sentido não é tão grande. Para mitigar este desafio, há um estudo constante destas técnicas, inclusive vindo da matéria sendo cursada em paralelo com esta, Big Data Stream, que também abrange técnicas de Big Data que podem ser utilizadas neste projeto.

Fizemos uma seleção de um *dataset* da área de saúde, na qual nenhum dos integrantes do grupo possui experiência. A seleção dos atributos mais importantes foi pautada apenas no bom senso dos participantes, podendo ser limitada nesse sentido. Também existe um risco da aplicação de métodos estatísticos e de machine learning não produzirem resultados significativos.

O prazo de oito semanas, comumente é um desafio na realização das atividades somativas ao longo do curso. Sabendo disso, planejamos reuniões frequentes com todos os membros para sempre manter a organização das atividades e assim completar o projeto no tempo requisitado.

Os recursos necessários para a realização com sucesso do projeto incluem-se computadores capazes de rodar o Jupyter Notebook e as bibliotecas necessárias para cada etapa. Incluindo também softwares como Java e Spark.

## 4.5 COLETA DOS DADOS

Para a realização desta etapa de coleta de dados foi utilizado o *dataset* completo, pois este não conta com a versão de amostra. Porém, após a realização de uma série de testes com este *dataset*, foi identificado que seria possível a realização das atividades sem maiores dificuldades mesmo manipulando desde o começo o *dataset* completo.

Nesta etapa, também foram inicializadas a manipulação do *dataset* com as ferramentas escolhidas anteriormente. O maior contratempo desta etapa foi que alguns membros do grupo tiveram dificuldades na importação de todas as bibliotecas utilizadas. Mas a migração do Jupyter Notebook para o Google Colab, junto com orientações dos próprios membros do grupo foram suficientes para resolver estes problemas e prosseguir como o planejado.

As principais variáveis para encontrarmos a solução do problema em questão (classificar as principais doenças afetadas pela população e relacionar com o ambiente de moradia dos moradores) foram definidas como: sexo do paciente, desencadeou atendimento, código e descrição do CID, energia elétrica, tipo de habitação, destino lixo, fezes/urina e meio de transporte, e estes são escritos e analisados na etapa a seguir.

Durante a exploração dos dados foram encontrados alguns casos de outliers, e estes foram removidos. Casos como atributos nulos, 1 caso em “Sexo”, 1 caso em “Desencadeou Internamento”, 3112 em “Abastecimento”, 3112 em “Tipo de habitação”, 1 em “Energia elétrica”, 3112 em “Destino de lixo”, 3112 em “Fezes/Urina”, 985 em “Cômodos”, 3116 em “Meio de transporte” e 8 em “Descrição do CID”.

Nesta etapa é pedido a análise métrica das varáveis, porém este *dataset* conta somente com a variável “Cômodos” sendo numérica, e entendemos que esta variável, do jeito que é apresentada no *dataset*, não tem geração de valor para o resultado, ainda mais por conter valores extremamente fora da realidade, então ela não está sendo utilizada. Sendo assim não resta uma variável numérica para a realização da análise métrica, mas é encontrado a seguir a análise estatística de todas as variáveis úteis para o projeto. Houve uma tentativa de normalizar os dados da variável “cômodos”, porém do jeito que foi coletado estes dados não foi possível fazer uma normalização dos dados adequada.

A “Descrição do CID” apresenta algumas informações que possuem pouco

significado para uma análise, tratando-se mais de descrições genéricas que informações sobre alguma doença. Um exemplo é a descrição com mais ocorrências “EXAME MÉDICO GERAL”, esta não expressa nenhuma informação sobre a enfermidade do paciente. Deliberadamente optamos por excluir as descrições excessivamente genéricas. Como os atributos escolhidos são representados por variáveis categóricas, a única medida estatística utilizada na análise exploratória de dados é a moda.

Após a alteração do objetivo do projeto, passamos a analisar as colunas apresentadas na figura 2:

Figura 2 – Colunas analisadas para o resultado final

```
 #   Column                                    Non-Null Count   Dtype
---  ------                                    --------------   -----
 0   Sexo                                      874894 non-null  object
 1   Descrição do CID                          874613 non-null  object
 2   Solicitação de Exames                     874894 non-null  object
 3   Encaminhamento para Atendimento Especialista  874894 non-null  object
 4   Desencadeou Internamento                  874894 non-null  object
 5   Abastecimento                             785548 non-null  object
 6   Energia Elétrica                          874894 non-null  object
 7   Tipo de Habitação                         785561 non-null  object
 8   Destino Lixo                              785563 non-null  object
 9   Fezes/Urina                               785561 non-null  object
dtypes: object(10)
```

Fonte: Os autores

A coluna “Desencadeou Internamento” passou a ser a informação a ser prevista a partir das outras. Alteramos o seu tipo de dados para numérico. Eliminamos todos os dados nulos das colunas selecionadas.

## 4.6 PREPARAÇÃO E ANÁLISE DOS DADOS

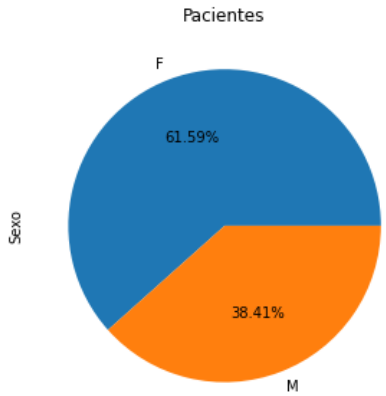
Utilizando as variáveis descritas anteriormente, foi possível tirar algumas conclusões. A maioria do atendidos foram do sexo feminino, aproximadamente 62%.

Dentre todos os pacientes atendidos, somente em 139 casos foi desencadeado o internamento logo após a consulta. Sendo que outros 20275 casos, o paciente pode voltar para casa após a consulta, mais de 99% dos casos.

Foi analisado que o abastecimento de água mais utilizado pelos moradores é o da rede pública, sendo equivalente à 98% dos casos, contando um total de 16698 moradores. No restante 2% dos casos, o que prevaleceu mais foi o “Outros”.

Figura 3 – Gráfico análise variável ’Sexo’

```
[41] plt.figure(figsize=(10,5))
     df3['Sexo'].value_counts(sort=True).plot.pie(autopct='%.2f%%')
     plt.title("Pacientes")

     Text(0.5, 1.0, 'Pacientes')
```




Fonte: Os autores

Figura 4 – Dados análise variável ’Desencadeou atendimento’

```
[18] df["Desencadeou Internamento"].value_counts()

     Nao    20275
     Sim      139
     Name: Desencadeou Internamento, dtype: int64

[19] df["Desencadeou Internamento"].isna().sum()

     1
```

Fonte: Os autores

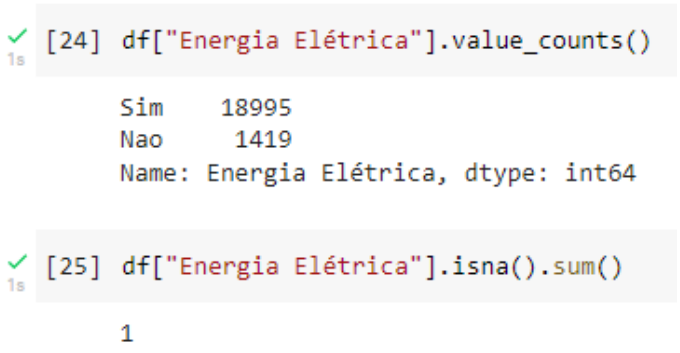
Figura 5 – Dados análise variável ’Abastecimento’

```
[20] df["Abastecimento"].value_counts()

     REDE PÚBLICA      16968
     OUTROS              286
     POÇO ARTESIANO       38
     CARRO PIPA            6
     CISTERNA              5
     Name: Abastecimento, dtype: int64

[21] df["Abastecimento"].isna().sum()

     3112
```

Fonte: Os autores

Foi observado que em um total de 18995 casos, foi relatado que dispõem de energia elétrica onde moram.

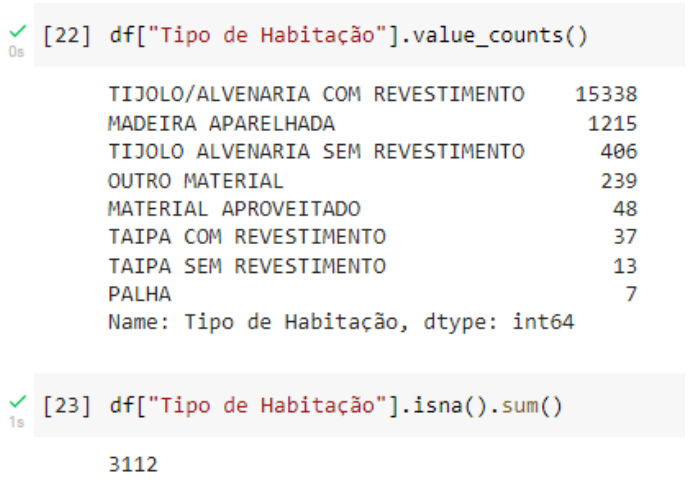
Figura 6 – Dados análise variável ’Energia Elétrica’




Fonte: Os autores

O maior valor obtido com os dados referente ao tipo de habitação foi em “Tijolo/Alvenaria com revestimento”, com um total de 15338. Seguindo de “Madeira Aparelhada” e “Tijolo alvenaria sem revestimento”.

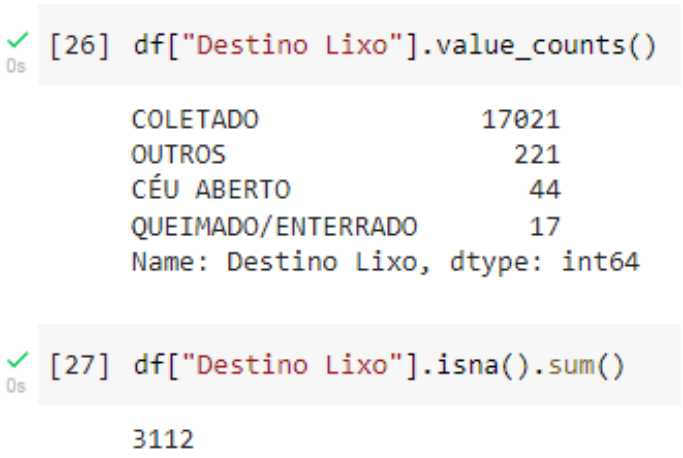
Figura 7 – Dados análise variável ’Tipos de Habitação’




Fonte: Os autores

Dentre os dados, 17021 casos foram relatados que o lixo é coletado em sua habitação, sendo o equivalente a pouco mais de 98% dos casos.

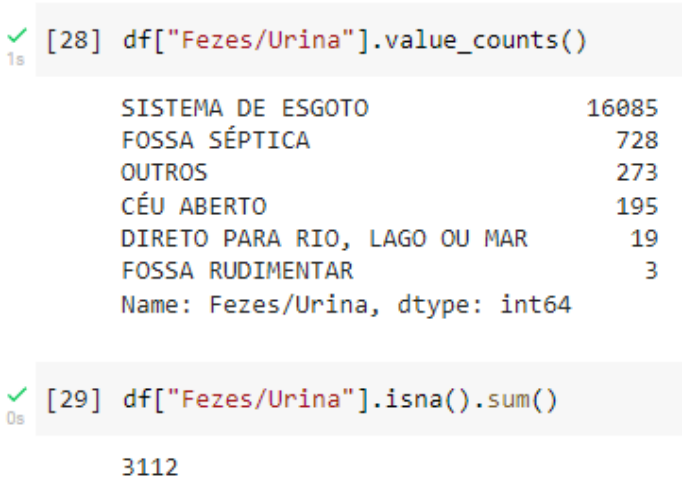
Figura 8 – Dados análise variável ’Destino Lixo’




Fonte: Os autores

Também foi observado um valor majoritariamente acima dos demais, aproximadamente 93% dos moradores relataram que têm sistema de esgoto em sua moradia.

Figura 9 – Dados análise variável ’Fezes/Urina’




Fonte: Os autores

No caso do atributo “Meio de transporte” foi somado os valores correspondentes à “Ônibus” e “Ônibus e Carro”, obtendo um valor de 14927 casos, sendo o equivalente à 86% dos casos relatados.

Figura 10 – Dados análise variável ’Meio de transporte’

```
[93] df["Meio de Transporte"].value_counts()
```

| | |
|---|---|
| ONIBUS | 9419 |
| ONIBUS,CARRO | 5508 |
| OUTROS | 928 |
| CARRO | 756 |
| OUTROS,ONIBUS,CARRO | 323 |
| OUTROS,ONIBUS | 144 |
| ONIBUS,CAMINHAO | 74 |
| ONIBUS,METRO,CARRO | 36 |
| ONIBUS,CARRO,CAMINHAO | 26 |
| CAMINHAO | 16 |
| METRO | 10 |
| ONIBUS,CARROCA,CARRO,CAMINHAO | 9 |
| OUTROS,ONIBUS,CARROCA,CARRO,CAMINHAO | 9 |
| ONIBUS,CARROCA | 6 |
| OUTROS,CARRO | 6 |
| OUTROS,ONIBUS,CARRO,CAMINHAO | 5 |
| CARROCA | 4 |
| OUTROS,ONIBUS,METRO | 3 |
| OUTROS,ONIBUS,CARROCA,CAMINHAO | 3 |
| CARRO,CAMINHAO | 3 |
| CARROCA,CARRO | 2 |
| METRO,CARRO | 2 |
| ONIBUS,CARROCA,CARRO | 2 |
| CARROCA,CAMINHAO | 1 |
| OUTROS,CARROCA,CAMINHAO | 1 |
| OUTROS,METRO | 1 |
| ONIBUS,METRO | 1 |
| OUTROS,ONIBUS,METRO,CARROCA,CARRO,CAMINHAO | 1 |

Name: Meio de Transporte, dtype: int64

Fonte: Os autores

Por fim, quanto aos casos de “Descrição do CID” em que relata o tipo de doença afetado pelo paciente, tivemos dados um pouco mais homogêneos, sendo os três principais: “diagnostico de covid-19 confirmado por exames laboratoriais”, "hipertensao essencial (primaria)"e "diarreia e gastroenterite de origem infecciosa presumivel". Outras variáveis foram desconsideradas da análise por se tratarem de descrições genéricas, não representando uma doença específica.

Figura 11 – Dados análise variável ’Descrição do CID’

| | |
|---|---|
| EXAME MEDICO GERAL | 120281 |
| DIAGNOSTICO CLINICO OU EPIDEMIOLOGICO COVID-19, QUANDO A CONFIRMACAO LABORATORIAL E INCONCLUSIVA OU NAO ESTA DISPONIVEL | 68669 |
| EMISSAO DE PRESCRICAO DE REPETICAO | 58370 |
| DIAGNOSTICO DE COVID-19 CONFIRMADO POR EXAMES LABORATORIAIS | 25212 |
| EXAME DE ROTINA DE SAUDE DA CRIANCA | 24578 |
| HIPERTENSAO ESSENCIAL (PRIMARIA) | 21819 |
| PROCEDIMENTO NAO REALIZADO DEVIDO A DECISAO DO PACIENTE POR OUTRAS RAZOES E AS NAO ESPECIFICADAS | 15896 |
| EXAME NAO ESPECIFICADO COM FINALIDADES ADMINISTRATIVAS | 14625 |
| DIARREIA E GASTROENTERITE DE ORIGEM INFECCIOSA PRESUMIVEL | 14254 |

Fonte: Os autores

## 5 ANÁLISE PARCIAL

Com a análise dos dados até o momento deste dataset semiestruturado, foi possível concluir que, apesar de apresentar alguns dados nulos e principalmente na variável “Cômodos”, conter também outliers que impossibilitaram a utilização da variável neste projeto, podemos dizer que o dataset apresenta dados suficientes para chegarmos a uma conclusão parcial.

As principais doenças que afetaram a população foram COVID-19, infecção aguda das vias aéreas e hipertensão essencial. Sendo que observando os dados da variável “Meio de transporte” vimos que o fato de 86% da população utiliza o ônibus como meio de transporte, certamente impactou para a contaminação de COVID. Mais informações detalhadas vai ser possível após a conclusão do projeto. (Esta análise parcial foi apresentada no final da primeira etapa do projeto, sendo que a análise final do trabalho encontra-se no capítulo 8, "Resultados".)

# 6 ETAPA DO APRENDIZADO DE MÁQUINA

Todos os dados utilizados na etapa do aprendizado de máquina são categóricos. Deste modo, foi necessário transformá-los em dados numéricos. Inicialmente utilizados o API do Pandas para Spark para manipularmos o dataset, desta forma, transformamos a coluna “Desencadeou Internamento” em dados numéricos com essa biblioteca. “Sim” passou a ser lido com 1 e “Não” como 0.

A biblioteca Mlib do Spark não lê objetos Pandas, deste modo convertemos os dados para um aquivo .csv antes de utilizarmos a biblioteca. Para transformarmos os dados categóricos em numéricos utilizamos as classes StringIndex, VectorAssembler, OnehotEnconder e VectorIndex. Ao final, os dados adquiriram forma binária.

Utilizados os algoritmos LogisticRegression e DecisionTree para o aprendizado de máquina. Como forma de avaliar os algoritmos, utilizamos as classes Multiclass-ClassificationEvaluator e BinaryClassificationEvaluator.

## 7 ANÁLISE MACHINE LEARNING

Ao utilizar o LogisticRegression com o método AUC, obtivemos uma acurácia de 0.5. Isso significa que nosso algoritmo não é bom para prevermos se uma pessoa será internada ou não. Com Decision Tree obtivemos resultados bem melhores, a acurácia foi de 0,99, sendo excelente para prever o internamento do paciente.

# 8 RESULTADOS

Consideramos que os objetivos pretendidos na atividade foram alcançados. Construímos um modelo bastante eficiente de prever se um paciente que passa por uma consulta na prefeitura de Curitiba será internado ou não. Assim, a secretaria de saúde teria muito mais facilidade em organizar a sua estrutura e planejar os internamentos de forma muito mais rápida, ganhando em eficácia no serviço.

# REFERÊNCIAS

MCKINNEY, W. et al. pandas: a foundational python library for data analysis and statistics. **Python for high performance and scientific computing**, Seattle, v. 14, n. 9, p. 1–9, 2011.

MENG, X.; BRADLEY, J.; YAVUZ, B.; SPARKS, E.; VENKATARAMAN, S.; LIU, D.; FREEMAN, J.; TSAI, D.; AMDE, M.; OWEN, S. et al. Mllib: Machine learning in apache spark. **The Journal of Machine Learning Research**, JMLR. org, v. 17, n. 1, p. 1235–1241, 2016.

SAGIROGLU, S.; SINANC, D. Big data: A review. In: IEEE. **2013 international conference on collaboration technologies and systems (CTS)**. [S.l.], 2013. p. 42–47.

STANČIN, I.; JOVIĆ, A. An overview and comparison of free python libraries for data mining and big data analysis. In: IEEE. **2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)**. [S.l.], 2019. p. 977–982.

WHITE, T. **Hadoop: The definitive guide**. [S.l.]: "O'Reilly Media, Inc.", 2012.