UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Rafael Morais de Assis

Preparando bases de dados para uso em sistema preditivo que visa a reduzir a emissão de ordens de serviço em empresa de telecom

Uberlândia, Brasil

2022

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Rafael Morais de Assis

Preparando bases de dados para uso em sistema preditivo que visa a reduzir a emissão de ordens de serviço em empresa de telecom

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Profa. Dra. Rita Maria da Silva Julia

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Ciência da Computação

Uberlândia, Brasil

2022

Rafael Morais de Assis

Preparando bases de dados para uso em sistema preditivo que visa a reduzir a emissão de ordens de serviço em empresa de telecom

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Ciência da Computação.

Trabalho aprovado. Uberlândia, Brasil, 01 de abril de 2022:

Profa. Dra. Rita Maria da Silva Julia
Orientadora

Fabíola Souza Fernandes Pereira

Maria Camila Nardini Barioni

Uberlândia, Brasil
2022

Agradecimentos

Primeiramente, agradeço à minha orientadora, Profa. Dra. Rita Maria da Silva Julia, por ser paciente e pelo grande auxílio na elaboração deste trabalho.

Um agradecimento enorme à minha família, que sempre me incentivou a estudar e confiou em minha decisão por estudar computação.

Agradecimentos especiais à Fabíola Fernandez, por ter participado intensamente no projeto. Também a Algar e a seus funcionários: Umberto, Jhony e à Marcela Janaina, pela ajuda para lidar com os dados da empresa. Agradeço também aos participantes do projeto de pesquisa e aos colaboradores da empresa que possibilitaram a realização desta pesquisa.

Por fim, agradeço à Orlandina, por ter me acolhido em uma nova cidade, e aos meus familiares, por terem ajudado nessa longa caminhada.

"The readiness is all" -
Shakespeare

Resumo

A grande quantidade de dados gerados no dia-a-dia da sociedade moderna, quando usada adequadamente, pode representar uma rica fonte de informações. Dentre os complexos problemas presentes em empresas que lidam com grande volume de dados e que prestam serviços diretos à sociedade, destaca-se a inevitável dificuldade de manter suas bases de dados atualizadas e adequadas para serem usadas como fonte confiável de informação em sistemas que visam a melhorar a experiência do cliente com a empresa. Nesse cenário, empresas de Telecom representam um relevante estudo de caso por, naturalmente, implementarem tecnologias de comunicação que servem como vetores de disseminação da informação digital. Assim sendo, o presente trabalho se encaixa no escopo de um projeto de pesquisa maior efetuado no seio da empresa Algar Telecom. Um dos grandes desafios da referida empresa é usar as informações presentes nos dados relativos a seus clientes de forma a melhorar a satisfação deles com os serviços prestados. Neste contexto, o objetivo geral do referido projeto global é o de efetuar a predição de aberturas de Ordens de Serviço na Algar Telecom, ou seja, buscar prever com antecedência se um dado cliente, em função de sua experiência com os serviços prestados pela empresa, está propenso a acionar um processo de reclamação que tenda a desencadear a abertura de uma Ordem de Serviço o que representa uma dinâmica que, além de onerosa para a empresa, é desagradável para o cliente. Dessa forma, a execução do citado projeto global visa a melhorar a experiência do cliente com a empresa. Contudo, um dos grandes desafios encontrados em sua implementação é o fato de as bases de dados envolvidas no processo apresentarem sérias fragilidades que comprometem sua utilização. Assim sendo, o objetivo central desta proposta de TCC consiste em preparar e interligar essas bases de dados de tal forma que elas possam ser usadas pelo sistema preditivo global, permitindo-lhe produzir resultados confiáveis. Para tanto, tais bases inicialmente foram submetidas a um cuidadoso processo de análise e limpeza de dados e, posteriormente, agrupadas. Os resultados satisfatórios do tratamento dessas bases de dados efetuado na presente proposta puderam ser validados por meio das boas acurácias obtidas pelo sistema preditivo.

Palavras-chave: pré-processamento, dados de telecom, ordem de serviço, machine learning, integração de dados.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AM	Aprendizado de Máquina
API	*Application Programming Interface*
CGC	Cadastro Geral de Contribuintes
CNPJ	Cadastro Nacional de Pessoa Jurídica
CPF	Cadastro de Pessoa Física
CRM	*Customer Relationship Management*
CTBC	Companhia de Telecomunicações do Brasil Central
DW	*Data Warehouse*
dB	Decibéis
GPS	*Global Positioning System*
id	Identificador
Kbps	*Kilobits* por segundo
KDD	*Knowledge Discovery in Databases*
PCA	*Principal Component Analysis*
RPA	*Robotic Process Automation*
SI	*Script* Integrado
SQL	*Structured Query Language*
OS	Ordem de Serviço
TELECOM	Telecomunicações

Sumário

1 Introdução

A área de exploração de dados tem se destacado devido à capacidade de geração de dados que aumenta, exponencialmente, desde 2009 (GANTZ; REINSEL, 2012). Em virtude do crescimento massivo da web no mundo e de mudanças comportamentais como uso de e-commerce, aplicativos de celular e mídias sociais, um enorme volume de dados é criado todos os dias. Para lidar com isso, a tecnologia também se desenvolveu nos últimos anos, para gerar, processar e armazenar diversos de dados, o que configura a área como *Big Data* (MAURO; GRECO; GRIMALDI, 2014).

Diante disso, empresas dos mais diversos segmentos podem se beneficiar com a exploração do universo dos dados. A pesquisa "*Global data management benchmark report 2018*" aponta que 86% das empresas os veem como um fator preponderante para definir estratégias de negócios e que 54% consideram o uso como uma vantagem competitiva (EXPERIAN, 2018), logo atesta-se que eles são benéficos para o cliente. Ademais, 60% das empresas pretendem melhorar a qualidade de experiência do consumidor por meio dos dados, enquanto outro estudo revela que 48% dos casos de uso de Big Data, em empresas, se direcionam aos clientes (DATAMEER, 2013).

No que diz respeito a uso de dados, as empresas de telecomunicações (Telecom) possuem uma vantagem em relação a outras – elas podem, naturalmente, coletá-los no momento em que o cliente usa a Internet, por serem detentoras desse serviço. Com o crescimento da utilização de *smartphones* e da internet móvel, as telecoms poderão coletar cada vez mais dados, (TIINSIDER, 2021), o que representa uma oportunidade para trabalhar com eles e os capitalizar nesse contexto.

De fato, a utilização de dados possibilita melhorar as relações direta entre cliente e organização, que ocorre por meio da compra e do uso de serviços, bem como as indiretas por meio de publicidades e recomendações de produtos (MEYER; SCHWAGER, 2007). Em uma pesquisa destinada a identificar a experiência ideal para os clientes, notou-se que o principal critério deveria ser a resolução rápida de problemas pela empresa (KANG, 2013). Consequentemente, uma boa experiência consegue evitar a saída do produto ou serviço, evento conhecido como *churn* (KLEMZ, 2019).

Nesse entremeio, a concorrência entre as telecoms no serviço de banda larga no Brasil é acirrada em razão do crescimento de menores organizações provedoras de Internet (ESTADAO, 2019). Ainda que a banda larga seja o serviço mais buscado (ESTADAO, 2019), é também considerado o pior desse tipo de empresa (CANALTECH, 2019). Logo, a fim de se destacar, uma concorrente deve resolver os problemas com rapidez, com vistas a manter uma certa qualidade na manutenção para reter o cliente.

Dentre as grandes organizações de telecom brasileiras se destaca a Algar, fundada com o nome Companhia de Telecomunicações do Brasil Central (CTBC) em 1954 e com sede em Uberlândia, Minas Gerais. Em virtude dos esforços para melhorar a qualidade da experiência do cliente, tal empresa será o foco deste trabalho, voltado à exploração de dados no serviço de banda larga e ao relacionamento com o cliente.

Na Algar, a reclamação de clientes sobre problemas na banda larga perpassa um fluxograma de ações e eventos, cujo processo se inicia por meio do contato do consumidor com o atendente da empresa. Este último realiza uma série de ações e procedimentos básicos de forma sistematizada para resolver questões técnicas junto ao cliente – essas iniciativas serão denominadas como "Script Integrado (SI)". Contudo, quando a dificuldade não é resolvida, abre-se uma Ordem de Serviço (OS), o que implica no encaminhamento de um técnico de campo para o local.

Vale ressaltar que a OS não é uma particularidade da Algar, por fazer parte de todas as telecoms no Brasil. A Agência Nacional de Telecomunicações (ANATEL), com fins de regulação, estabelece, no Art. 25, §2 da Resolução n. 574/2011 (ANATEL, 2017), que a prestadora de serviço de banda larga precisa manter um histórico de OS com as ações adotadas no atendimento e o tempo necessário para realizar o reparo.

A geração de OS, além de implicar em prejuízos no tocante à imagem da organização perante a Anatel, representa um gasto com o técnico de campo que deverá ir ao local e nem sempre haverá alguém para atendê-lo. Além disso, ele pode encontrar um cenário diferente do descrito pelo atendente, no qual pode ou não conseguir resolver a situação, o que gera custos de deslocamento desnecessários. O fato de o problema não ter sido resolvido pelo atendente via SI também ocasiona espera até a chegada do técnico, algo prejudicial à experiência do cliente com a organização. Assim, a geração de OS prejudica ambos os lados, visto que a dificuldade poderia ter sido resolvida de forma rápida pelo atendente via SI.

Isso posto, o presente trabalho se enquadra em um âmbito maior do objetivo geral: usar técnicas de ciência de dados para analisar os bancos de dados da Algar, a fim de melhorar a experiência do cliente. Tal proposta é um projeto de pesquisa e inovação que envolve a Universidade Federal de Uberlândia (UFU), a Universidade de São Paulo – Campus São Carlos (USP-ICMC) e a Algar, por meio do financiamento do BRAIN, programa da Algar que pretende incentivar a inovação e conta com a participação de alunos de pós-doutorado e Iniciação Científica (IC).

Nesse caso, foram propostos dois protótipos: 1) análise do SI para atenuar ou eliminar possíveis inconsistências das etapas para propor a otimização, o que possibilita a resolução dos problemas do cliente na ligação telefônica; 2) descobrir o padrão de comportamento da rede do cliente antes de ele solicitar uma OS e prever a abertura.

No que tange às dificuldades encontradas, a principal consiste na relação da empresa com a burocracia de alguns processos para disponibilizar informações. Nesse caso, encontraram-se problemas como a falta de informação, a redundância e a dificuldade no cruzamento de banco de dados, pois a descrição de cada campo não era autoevidente, o que requereu assistência interna para se realizar de fato. Para resolver tais situações, foi necessário diálogo entre os pesquisadores e os funcionários da Algar para esclarecer as dúvidas.

Em virtude disso, o objetivo específico que motiva a proposta do presente trabalho se refere a analisar os problemas citados nas bases de dados da Algar e prepará-las para fundamentar o protótipo 2, com a predição de abertura de OS em decorrência da rede do cliente, o que pode otimizar a experiência com a empresa. Além disso, foram feitas diversas análises de dados para obter insights sobre eles, compreendê-los e identificar a viabilidade ou não da proposta. Como resultado, elaborou-se uma base de dados para conectar o cliente ao estado da rede antes de solicitar uma OS, bem como se criou uma base confiável para prever a abertura de OS.

1.1 Organização do Trabalho

Os próximos capítulos deste trabalho estão organizados da seguinte forma:

- **Capítulo 2 - Fundamentação Teórica**: apresenta referencial teórico para entendimento das atividades aqui conduzidas incluindo uma visão geral sobre banco de dados e pré-processamento de dados.
- **Capítulo 3 - Trabalhos Relacionados**: apresenta estudos relacionados a temática deste trabalho.
- **Capítulo 4 - Desenvolvimento**: expõe as etapas de pré-processamento das bases, sua integração e utilização para o modelo de Aprendizado de Máquina (AM) (do inglês *Machine Learning*).
- **Capítulo 5 - Conclusão**: contém as principais conclusões deste trabalho e sugestões para trabalhos futuros.

2 Fundamentação Teórica

Nesta seção são elencados os conceitos básicos necessários para a compreensão do presente Trabalho de Conclusão de Curso (TCC). A seção 2.1 apresenta conceitos de banco de dados, incluindo *Data Warehouse* (DW). Enquanto isso, a seção 2.2 cita os conceitos de pré-processamento de dados.

2.1 Banco de Dados

De acordo com Elmasri e Navathe (2015), o banco de dados diz respeito a uma coleção de dados que se relacionam entre si. O dado se refere a determinado fato, informação ou figura registrada que possui significado implícito, como o número "42" ou a palavra "xrl8".

2.1.1 Modelo relacional

Há várias formas para organizar conceitualmente um banco de dados. O modelo relacional, por exemplo, é comumente utilizado em aplicações de processamento de dados no âmbito comercial (RAMAKRISHNAN; GEHRKE, 2000). Esse modelo é baseado em esquema que define nome e aspectos de entidades dentro do banco de dados. Nesse caso, uma coleção de dados relacionados entre si forma um registro que representa os dados com sentido específico (RAMAKRISHNAN; GEHRKE, 2000).

Na sequência, apresenta-se um exemplo de registro:

Tabela 1 – Exemplo de registro

Número	Usuário	Idade
1	xrl8	42

Diante disso, o conjunto de registros é organizado para formar uma tabela, também chamada de relação. As colunas são chamadas de "campos" ou "atributos", enquanto as linhas são denominadas como "tuplas" ou simplesmente "registros". Os valores dos atributos devem ser simples e atômicos e se ao atribuir um valor ele não existir ou ser desconhecido, é atribuído o valor nulo.

Um exemplo rudimentar de relação se refere à lista telefônica que armazena nome, número de telefone e endereço. Ela pode ser interpretada no modelo relacional indicado na Tabela 2, em que os registros correspondem às informações e possuem os atributos de "nome", "telefone" e "endereço":

Tabela 2 – Exemplo de relação para uma lista telefônica

Nome	Telefone	Endereço
Ricardo Moreira	3499743-5634	Rua Suécia
Silvana Damasco	7397456-5263	Rua Davi Fadini
Ramon Santana	3295634-5723	Rua Antares

Nas tabelas, para distinguir uma tupla das outras, são escolhidos um ou mais atributos, cujos valores sejam não-vazios e únicos para cada tupla, sendo seu identificador. Os atributos capazes de fazer essa distinção e que possuem tais características são denominados como atributos primários. Cada conjunto de atributos com essas características são chamados de chaves candidatas e a partir delas é escolhido a chave primária. (ELMASRI; NAVATHE, 2015).

Chaves primárias precisam ser escolhidas de maneira cuidadosa, pois devem corresponder a um atributo ou mais atributos cujo valores não devem se repetir de fato. Na tabela da lista telefônica, "Nome" não poderia ser chave primária, pois duas pessoas podem ter o mesmo nome. Para esse propósito, pode-se adicionar o atributo "CPF", por não haver dois registros com o mesmo CPF; logo, é um identificador único, como demonstra na Tabela 3, com a inserção desse atributo:

Tabela 3 – Exemplo de Tabela com chave primária

CPF	Nome	Telefone	Endereço
043.365.875-00	Ricardo Moreira	3499743-5634	Rua Suécia
045.744.772-23	Silvana Damasco	7397456-5263	Rua Davi Fadini
046.244.563-78	Ramon Santana	3295634-5723	Rua Antares
047.822.453-78	Ricardo Moreira	3298694-5929	Rua Dom Pedro II

Uma tabela t1 pode se relacionar com t2 por meio de seus atributos – nesse caso, existe um atributo de mesmo domínio da chave primária de outra tabela, com a possibilidade de associar os registros de t1 aos de t2. Um campo que referencia a chave primária de outra tabela é chamado chave estrangeira. Juntamente à Tabela 3 pode haver no mesmo banco de dados, a Tabela 4, com uma chave estrangeira *CPF_Pessoa*, sempre com valores já presentes em *CPF* da Tabela 3. A presença de chave estrangeira desenvolve uma relação de referência; por conseguinte, *CPF_Pessoa* deve ter um valor que ocorra em algum registro da chave primária que relaciona ou terá valor nulo.

Tabela 4 – Exemplo de tabela com Chave Estrangeira

Usuario	Senha	**CPF_Pessoa**
xlr8	1234	**043.365.875-00**
4free	09121996	**045.744-772-23**
rafanthx13	alfa4thunder	**046.244..563-78**

2.1.2 Operação de junção de tabelas

Em um banco de dados relacional, há a operação de unir tabelas, usada para combinar informações de duas relações em consulta (RAMAKRISHNAN; GEHRKE, 2000); com isso, obtém-se uma nova relação cujo registro apresenta os dados de ambas as tabelas. Geralmente, tal ação ocorre desde que as duas relações tenham um mesmo atributo com o mesmo domínio, o que é possível a partir da relação das chaves primária e estrangeira.

Na linguagem padrão SQL (*Structured Query Language*), para as consultas e gerenciamento dos sistemas de banco de dados que utilizam o modelo relacional, tal operação ocorre nas consultas pela cláusula JOIN, que permite conectar duas tabelas por um ou mais atributos em relação a outros do mesmo domínio. Isso acontece, em geral, a partir da chave estrangeira de uma tabela com a chave primária de outra, à qual se referencia nesse contexto. Para realizar tal ligação, há as seguintes formas ilustradas na Figura 1, em que a tabela da esquerda é indicada por *t1*, e a da direita, como *t2*:

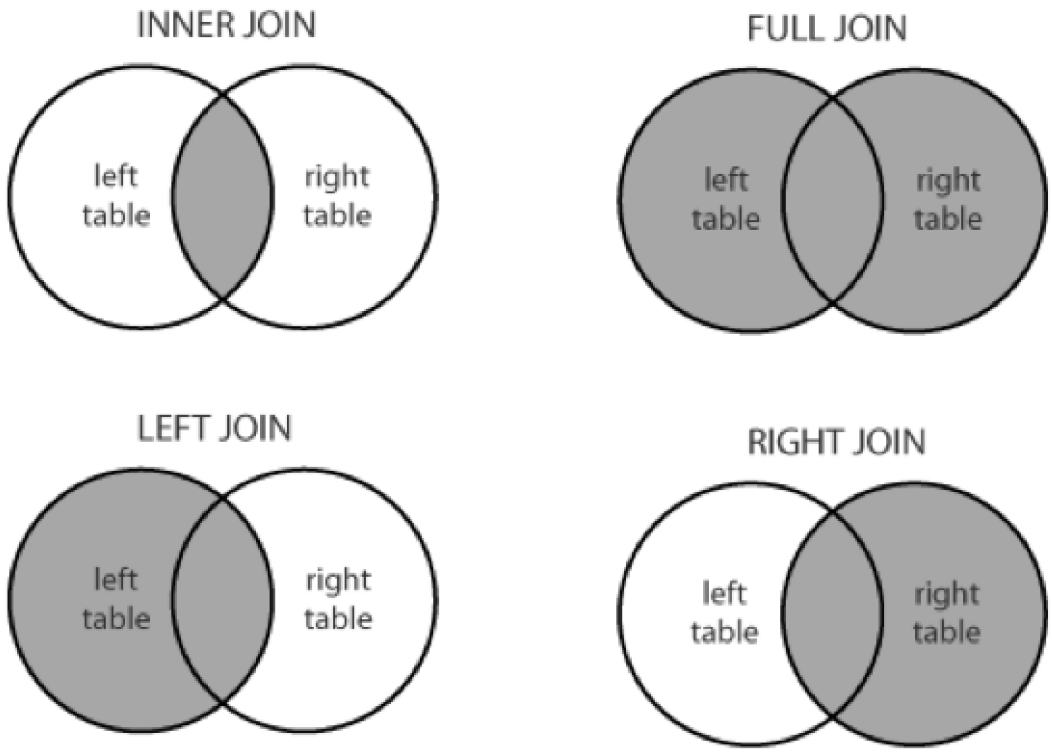

Figura 1 – Diferentes formas para obter Join em SQL (DELVA, 2020)

- **INNER JOIN**: retorna a relação dos registros combinados de t1 e t2 correspondentes.
- **LEFT JOIN**: retorna a relação de todos os registros de t1 correspondentes a t2 e os registros de t1 que não a corresponderam.

- **RIGHT JOIN**: retorna a relação de todos os registros de t1 correspondentes a t2 e os registros de t2 que não a corresponderam.
- **FULL JOIN**: retorna a relação de todos os registros de t1 com os de t2, independentemente se corresponderam ou não, o que equivale a realizar *LEFT JOIN* com *RIGHT JOIN*.

2.1.3 Conceitos de *Data Warehouse*

Turban et al. (2009) definem *Data Warehouse* (DW) como um repositório de dados organizado para obter fácil acesso e manipulação. Enquanto isso, Kimball e Ross (2002) citam que o principal objetivo do DW é prover informações de fácil acesso, em que as deve apresentar de forma consistente. O modelo de DW deve ser adaptável a mudanças e visa dar suporte à tomada de decisões.

Pelo fato de fundamentar análises mais profundas e tomadas de decisão, o DW possui as seguintes características:

- **Integral**: unifica dados de variadas fontes, como banco de dados, transações ou arquivos, etc. (HAN; KAMBER; PEI, 2011).
- **Orientado a tempo**: é capaz de mostrar tendências e variações ao longo de determinado período. Aqui, os dados provém da perspectiva histórica.
- **Informação não voláteis**: não se excluem os dados após serem inseridos, exceto quando estão extremamente obsoletos (nesse caso, podem ser inúteis) ou incorretos (não são relevantes para análises sérias).

Afirma-se, pois, que o DW é otimizado para manter o histórico dos dados e se difere da arquitetura de um banco de dados comum, cujo foco principal é processar transações. Ademais, o DW visa ser um armazém de informação de forma rápida e otimizada.

2.1.4 Modelo Dimensional

O modelo dimensional se estrutura para a recuperação e consulta de um alto volume de dados. Ele é necessário ao DW porque, na modelagem relacional do banco de dados, é mais difícil fazer consultas complexas do que na dimensional. O paradigma mais adotado a essa modelagem é o esquema estrela, com um projeto de várias tabelas menores (tabelas dimensões) orientadas a apenas um assunto e a uma grande tabela central (tabela fato) (HAN; KAMBER; PEI, 2011).

Tabela fato: possui vários registros que correspondem a fatos, observações, medidas e métricas de acordo com seu propósito. Inclui diversas chaves para as tabelas

dimensões, levando a mais detalhes sobre um determinado fato. Sobre ela, elaborada em conjunto às tabelas dimensões, realiza-se a operação de consulta e análise de dados para a tomada de decisão.

Tabela dimensão: contém diversos detalhes descritivos sobre determinado assunto referenciado na tabela fato.

Um exemplo do esquema estrela é apresentado na Figura 2. Nesse contexto, Venda (*sales*) é a tabela fato que aglutina diversas chaves de assuntos específicos sobre a venda e se origina das tabelas dimensões.

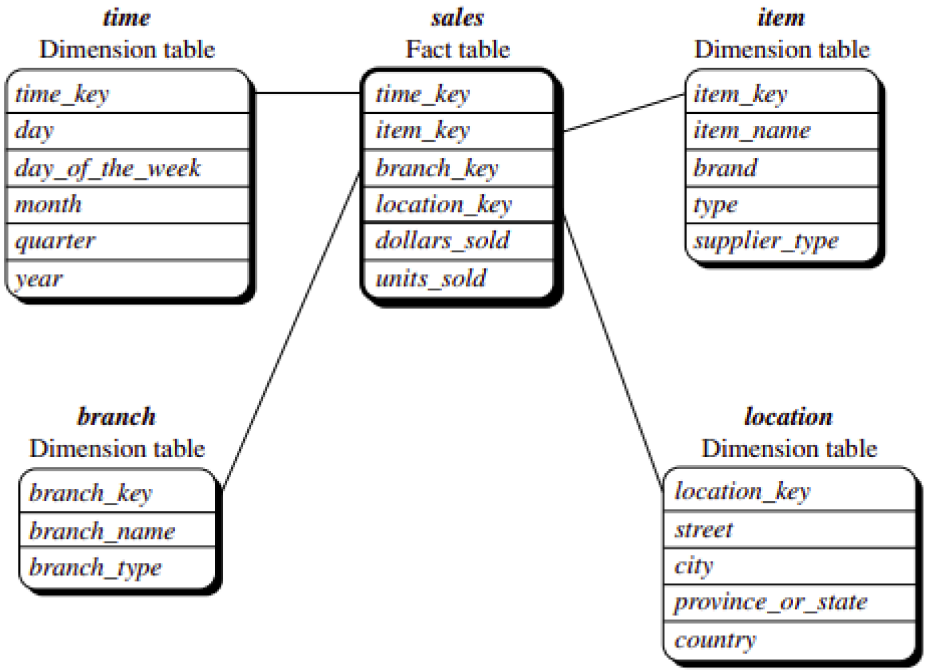

Figura 2 – Exemplo de esquema estrela para o assunto “venda” (HAN; KAMBER; PEI, 2011)

2.2 Pré-processamento de dados

Devido ao tamanho elevado e por terem diversas fontes de origem, as informações de bancos de dados podem apresentar problemas como dados inconscientes ou ausentes (HAN; KAMBER; PEI, 2011). Ocasionados por erro humano ou problemas na coleta, é necessário submetê-los a procedimentos para garantir a qualidade e minimizar quaisquer problemas que possam ser encontrados. Tanto para AM quanto para Descoberta de Conhecimento em Base de Dados – *Knowledge Discovery in Databases* (KDD), recomenda-se submeter os dados a tais processos (FACELI, 2011) exemplificados na Figura 3.

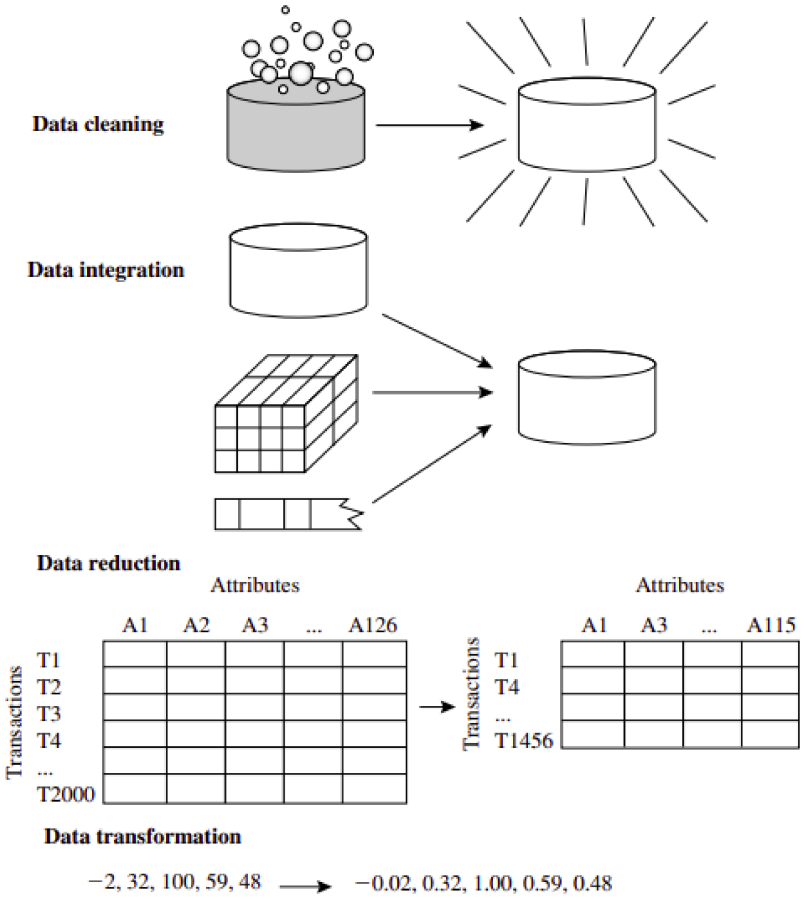

Figura 3 – Etapas de pré-processamento de dados (HAN; KAMBER; PEI, 2011)

2.2.1 Limpeza de dados

Dados do mundo real podem apresentar ruído, ser inconsistentes ou ter outros problemas (HAN; KAMBER; PEI, 2011). As subseções a seguir descrevem alguns cenários que podem ser encontrados na análise dos dados e sugerem procedimentos para a sua limpeza.

2.2.1.1 Dados com Ruído

Dados com ruído contêm valores que parecem não pertencer à distribuição do seu domínio. Nesse sentido, o dado apresenta um valor discrepante do restante da coluna, distante do desvio padrão e pode ter sido originado de um erro aleatório na aquisição (FACELI, 2011).

Valores diferentes de outros em um mesmo atributo são chamados de *outliers*, mas nem sempre é possível determinar se são apenas valores extremos ou se foram gerados por ruído (FACELI, 2011). Para verificar esse aspecto, é necessário analisar o seu domínio: por exemplo, a média de gols em uma partida de futebol do Campeonato Brasileiro em 2021 foi de três gols, aproximadamente (BRAZILIENSE, 2021), mas, se houvesse uma partida com 12 gols, ela iria corresponder a um *outlier*, mas não a um dado com ruído. Todavia, se ao medir a altura da pessoa em uma sala e encontrar alguém com mais de três metros, acredita-se que esse indivíduo se refere a um *outlier* realmente inválido.

Há diversas técnicas estatísticas para lidar com dados ruidosos, mas a principal questão diz respeito a analisar o impacto da remoção deles. Os *outliers*, apesar de serem reais, ainda podem influenciar o resultado de processos subsequentes devido ao valor ser extremo, como nos processos de Aprendizagem de Máquina (AM).

2.2.1.2 Dados Faltantes

Determinado registro pode apresentar diversos valores ausentes, como desconhecimento do valor no momento de preenchimento; distração, desleixo ou falta de obrigatoriedade na aquisição de dados; inexistência de valor para o atributo em um registro específico; e problemas na coleta, transmissão e armazenamento (FACELI, 2011). Independentemente dos motivos, é necessário lidar e analisar os dados, observar a quantidade de ocorrências e tomar as seguintes ações:

- **Ignorar**: conforme o domínio do atributo e do objetivo da base de dados, a falta de valores pode não apresentar nenhum impacto. Dessa maneira, não se deve realizar nenhuma ação para lidar com ele (TAN et al., 2018).
- **Eliminar registro**: de acordo com a importância do campo faltante ou pela ausência de dados em diversos atributos, pode-se excluir a linha da base de dados (HAN; KAMBER; PEI, 2011).
- **Inserir automaticamente valores**: segundo o domínio do atributo, é possível substituir os valores faltantes por outro, a exemplo dos valores numéricos compensados pela média ou moda do atributo; e dos valores textuais, com a substituição por um valor indicado como “desconhecido” ou outra constante (FACELI, 2011).

2.2.1.3 Dados Inconsistentes

Dados inconsistentes possuem valores que não se adéquam ao domínio, ou seja, em relação a valores de outros atributos. É possível detectá-los quando a relação lógica e contextual entre valores e atributos é violada pelos valores (FACELI, 2011).

Para um valor ser inconsistente com a coluna, pode-se ter, por exemplo, o valor "Ronaldo" para a coluna "profissão". Tal elemento não é consistente, pois representa um nome pessoal, e não uma profissão. Isso também pode acontecer entre valores de atributos: para uma tabela que registra dados médicos, é considerada inconsistência uma pessoa com idade de 10 anos ter o peso de 99 quilos (FACELI, 2011).

Com o intuito de lidar com tais aspectos, pode-se usar os mesmos procedimentos dos dados faltantes ou removê-los de maneira manual.

2.2.1.4 Representação inconsistente dos dados

Inconsistências na representação dos dados podem ocorrer quando um mesmo dado é representado por valores diferentes (BATISTA, 2003). Por exemplo, para a coluna "nome da faculdade", valores como "UFU", "Ufu" e "Universidade Federal de Uberlândia" são diferentes, mas representam a mesma instituição (HAN; KAMBER; PEI, 2011). Com o intuito de resolver esse tipo de problema, detectam-se e se substituem os valores para uma mesma forma, como a troca de tais ocorrências por somente "UFU".

2.2.1.5 Dados Redundantes e Duplicados

Um conjunto de dados pode apresentar redundância e duplicação nos valores e entre atributos. Esses problemas geralmente ocorrem em etapas de coleta, integração e transmissão de dados, como descrito a seguir (FACELI, 2011):

- **Registros duplicados**: acontecem quando um ou mais registros possuem o mesmo valor nas mesmas colunas. A depender do tema da tabela e do propósito dos dados, torna-se imprescindível eliminar essa duplicação.
- **Atributos duplicados**: se referem ao momento em que duas ou mais colunas possuem os mesmos valores para cada registro.
- **Atributos Redundantes**: dizem respeito a dois atributos que, apesar de serem diferentes, possuem uma correlação e o mesmo perfil de variação. Exemplos da redundância seriam os atributos "data de nascimento" e "idade", em que a última pode ser definida pela data de nascimento.

A presença de duplicação e redundância contribui para que esses dados sobressaiam em relação aos outros, por terem mais ocorrências quando, na verdade, não deveriam

existir nesse contexto. Por isso, geralmente é desejável identificá-los e eliminá-los, o que usualmente ocorre ao final da limpeza de dados (FACELI, 2011).

2.2.2 Transformação de dados

Após o processo de limpeza, os dados podem ser transformados, a fim de adequar a representação a algoritmos como os usados em AM ou KDD. A decisão ou não por esse procedimento depende do algoritmo a ser utilizado, a exemplo das ações citadas a seguir (BATISTA, 2003):

Normalização em atributos numéricos

Consiste em transformar valores numéricos de qualquer intervalo para intervalos menores (entre -1 e 1 ou entre 0 e 1, por exemplo). Algoritmos como redes neurais se beneficiam dessa transformação, enquanto outros não possuem nenhum impacto, como as árvores de decisão (BATISTA, 2003).

Conversão de simbólico para numérico

Se o atributo possui dois valores textuais, é possível substituí-lo por *0* e *1*, a fim de indicar a ausência ou não de um valor. Exemplo: o atributo "sexo" aceita "F" ou "M"; tendo como base o valor "M", pode-se converter para *0* ou *1*, em que 0 indica a ausência (no caso, "F"), e um, a presença ("M") (FACELI, 2011).

Quando há mais valores, pode-se mapeá-los para valores e números arbitrários. Se não houver nenhuma ordenação nos valores simbólicos, isso pode ocorrer de forma totalmente arbitrária; porém, caso se apresentem como "criança", "jovem" e "adulto", o mapeamento deverá seguir a mesma hierarquia de idade (criança = 2, jovem = 4 adulto = 6, por exemplo) (BATISTA, 2003).

Converter de numérico para simbólico

Um exemplo é a técnica de discretização, que converte valores numéricos em textuais. Para isso, torna-se necessário definir um número "n" de intervalos e dividir o intervalo dos dados numéricos em "n" partes. Vale ressaltar que, a cada valor que pertence a um intervalo, se atribui um valor (FACELI, 2011).

2.2.2.1 Criação de Atributos

Por meio do conjunto original de dados, é possível obter novos atributos ao capturar as informações mais importantes desses dados e torná-los mais eficientes (TAN et al., 2018). Há três formas para a criação dos atributos:

- **Extração de atributos**: desenvolve um conjunto de novas colunas diretamente dos dados originais. É uma técnica bastante utilizada para lidar com o processamento

de imagens.

- **Mapeamento para um novo domínio**: criam-se atributos a partir da aplicação de uma função a um ou mais atributos.
- **Construção de atributo**: utilizada quando um atributo tem a informação completa, mas não está em um formato suficientemente explícito. Nesse caso, constroem-se novas colunas a partir da antiga que, por evidenciar a informação, torna irrelevante o atributo original.

2.2.3 Redução de dimensionalidade

Tornar a base de dados enxuta é essencial. Ter vários atributos pode prejudicar o processo de AM e KDD, pois tornam o conjunto de dados mais complexo, e a aplicação de algoritmos, mais custosa. As técnicas para realizar a redução de dimensionalidade podem ser divididas em seleção e agregação de atributos (FACELI, 2011).

2.2.3.1 Seleção de atributos

Após a limpeza dos dados, alguns atributos podem ser irrelevantes, seja por haver muito ruído ou por não ter importância em relação ao problema a ser trabalhado. Há três abordagens usadas para escolher os atributos (KOHAVI; JOHN, 1997):

- ***Embedded***: atributos são selecionados durante o processo de algoritmo de AM, como nos algoritmos de árvores de decisão.
- **Filtro**: é uma abordagem que pode ser utilizada na etapa de pré-processamento e independe de qualquer algoritmo, a qual consiste em criar critérios numéricos e lógicos para filtrar os atributos, com base na tarefa a ser realizada. A natureza do conjunto de dados pode indicar esses critérios, ao definir se um atributo pode ou não ser relevante (BATISTA, 2003).
- ***Wrapper***: se refere à realização de testes para diversos subconjuntos de atributos, a fim de testá-los e avaliá-los por meio de algoritmos de AM. Escolhe-se, assim, o subconjunto que obtiver a melhor avaliação (FACELI, 2011).

2.2.3.2 Agregação

As técnicas de agregação combinam atributos por meio de técnicas estatísticas, o que reduz o número de atributos e busca manter a integridade dos dados. Uma das técnicas mais utilizadas é a Análise de Componente Principal (PCA, do inglês *Principal Component Analysis*), que reduz a dimensionalidade, além de identificar e extrair a redundância das colunas (FACELI, 2011).

2.2.4 Integração de Dados

Os dados de diversas fontes são unificados com frequência. Por isso, é necessário se atentar a esse aspecto, pois a integração pode acarretar redundâncias e inconsistências no conjunto resultante de dados (HAN; KAMBER; PEI, 2011).

Problema de Identificação de entidades

Ao unificar os dados de diferente fontes, a primeira etapa a ser considerada se refere à combinação entre eles, algo visto como um problema de identificação de entidades (HAN; KAMBER; PEI, 2011). Para tanto, é preciso escolher um atributo de cada tabela e se certificar de que eles se relacionam ao mesmo elemento, por meio da análise dos metadados, da observação de nome, significado, tipo, intervalo dos valores permitidos etc. nesses atributos, o que acontece no momento em que uma chave estrangeira referencia a chave primária de outra tabela.

Análise de Correlação e Redundância

Com a junção de vários atributos em apenas uma base de dados, pode ocorrer uma redundância que pode passar despercebida por haver origens diferentes. Além de observar os metadados, outra forma de detectar a derivação de uma coluna em relação a outra é a partir da análise de correlação, que consiste em verificar os impactos de um atributo em outros, com base em seus dados. Para os dados textuais, existe o teste de qui-quadrado, e para os numéricos, o cálculo de coeficiente de correlação e a covariância (HAN; KAMBER; PEI, 2011).

Cumpre afirmar que a mesma informação pode ser repetida por mais de um atributo, seja codificação, escala ou diferentes níveis de abstração. Por exemplo: uma tabela que calcula a venda de produtos tem a possibilidade de apresentar a coluna “valor total”, ao passo que, em outra fonte, pode haver duas colunas como “valor unitário” e “quantidade comprada”. Após a análise, não há a necessidade da repetição da mesma informação; por isso, ela deve ser descartada.

3 Trabalhos Relacionados

Nesta seção serão apresentados alguns trabalhos relacionados à etapa de preparação das bases de dados. Diversas pesquisas possuem a fase de pré-processamento de dados, mas nenhuma com o escopo específico deste estudo, isto é, tratar bases de redes de telecom para predizer a abertura ou não de OS.

Assim, encontrou-se a investigação de Francischelli (2013), que trata do impacto na tomada de decisão da empresa H&TEC Soluções, por utilizar o processo de KDD sobre dados da *web*. Foram extraídos documentos da base de dados *Brown Corpus* da coleção de documentos públicos da Internet *Archive Text*, relativa a uma biblioteca digital livre. Depois disso, realizou-se o pré-processamento dos textos escritos em linguagem natural para convertê-los em uma representação mais estruturada e manipulável por algoritmos de KDD por meio da ferramenta Weka, para a descoberta de conhecimento em consonância às técnicas de associação e clusterização. Por fim, verificou-se que os dados públicos podem ser usados para haver uma seleção mais adequada de dados para processos de mineração de texto, o que representa um diferencial competitivo para a empresa.

Outro trabalho relativo à coleta e limpeza foi elaborado por Ferreira (2017) , que aborda a preparação de dados provenientes do Sistema de Posicionamento Global – *Global Positioning System* (GPS). Devido a diversos fatores que podem ocorrer na coleta de dados de GPS, como ruídos e interferências, torna-se importante realizar o pré-processamento dos dados não apenas para eliminá-los, como também para padronizar ao domínio das trajetórias. Após isso, são listados os principais métodos de tratamento desse tipo de dado na literatura para os aplicar a dois estudos de casos das bases públicas *Geolife* e *Taxi San Francisco*. Nesse caso, desenvolveu-se um programa em Java para extrair e padronizar qualquer formato com dados geoespaciais. Por fim, foi verificado que, a partir de métodos de análise de similaridades, os dados de GPS ficaram mais discerníveis após o pré-processamento em relação ao período anterior, o que facilita o trabalho de KDD ou AM e demonstra a relevância da preparação de uma base de dados.

Silva (2018) prepara uma base de dados dos benefícios previdenciários do Ministério Público para a extração de conhecimento. Além do pré-processamento, criaram-se bases para obter diversas versões e testar algoritmos de agrupamento para diversas visões. Para a validação, foi utilizado o índice de silhueta simplificada, no qual os agrupamentos não conseguiram um desempenho adequado.

Em Umezawa (2021) é elaborado um projeto para monitoramento e consulta de processos robóticos automatizados (Automação Robótica de Processos – RPA) pelo usuário. A interação com o programa é feita a partir de um *chatbot* criado na linguagem

python, desenvolvido com *layout* simples e intuitivo para respostas rápidas. A meta desse programa é agilizar o acesso às informações dos bancos de dados nos processos de RPA por meio de ações estáveis e cruzamento de tabelas. Isso, de acordo com os resultados obtidos, se mostrou eficiente ao proporcionar economia no tempo, em se tratando do acesso aos dados, em detrimento ao método convencional de consulta direta.

4 Desenvolvimento

Neste capítulo será descrito o desenvolvimento deste estudo. Na Seção 4.1 é descrito em que medida esse trabalho se encontra no escopo geral do projeto de pesquisa. A Seção 4.2 lista os procedimentos de pré-processamento para as quatro bases. A Seção 4.3 descreve como foi feito o agrupamento dos dados. A Seção 4.4 descreve de forma geral como a base foi usada para o processo de AM. A Seção 4.5 é desenvolvido um programa usando o modelo criado anteriormente para predizer dados novos e antigos, obtendo também estatísticas sobre seu desempenho.

4.1 Objetivo do protótipo

Este trabalho corresponde a uma etapa de um projeto maior de pesquisa que visa realizar o processo de AM para gerar um modelo preditivo que possa prever abertura de OS. Para tal atividade, foi disponibilizado pela empresa de telecomunicações Algar, bases de dados e suporte para tal trabalho de pesquisa.

O objetivo é preparar os dados fornecidos pela empresa para servir como base de dados para geração de tal modelo. Foi fornecido por ela 4 bases de dados: base de cliente, de OS, do Rubi e dos equipamentos. Tais bases de dados foram obtidas de forma direta, sem qualquer tratamento, inclusive sem realizar a anonimização de dados pessoais. Por conta disso, foram encontrados vários problemas como inconsistências, dados faltantes e até mesmo informações repetidas. O presente trabalho tem como meta resolver tais problemas para cada base, interligá-las e assim obter uma base de dados única, otimizada e integral.

4.2 Bases de dados

Nas seções a seguir, serão descritas as bases utilizadas, os problemas encontrados nesse contexto e a maneira como foram resolvidos por meio das atividades conduzidas neste estudo.

4.2.1 BD-1: Base de dados Cliente

Descrição da base: possui 1.239.483 registros com 44 colunas de informações dos clientes cadastrados. Os atributos principais são:

- *CUSTOMER_KEY*: identificador do cliente no DW.

- *NAME*: nome do cliente.
- *CPF_CGC*: número do documento de identificação do cliente. Cadastro de Pessoa Física (CPF) para um individual e Cadastro Geral de Contribuintes (CGC), equivalente ao Cadastro Nacional de Pessoa Jurídica (CNPJ) para corporativo.
- *DESCRIPTION*: nome do produto
- *CONTRACT_KEY*: chave identificadora do contrato.

A base possui uma quantidade enorme de registros, mas, não quer dizer que cada uma das linhas representa clientes distintos. Esta base vem do sistema de gestão de relacionamento com o cliente (em inglês, CRM) da Algar: um sistema corporativo para gerenciar o contato entre o cliente e a empresa e funciona como uma Tabela Fato, uma estrutura comum em *Data Warehouses*. Assim sendo, pode haver mais de um registro que se refere a um mesmo cliente, representando alterações ocorridas nos valores anteriores de seus atributos descritos nos registros anteriores. Portanto, a quantidade de clientes envolvidos em tal base não coincide, necessariamente, com a quantidade de registros dela. Logo, a sequência de registros relativos a um mesmo cliente permite que se analise o histórico das alterações referentes a ele, desde a contratação de um serviço da Algar até a sua desativação, pois cada alteração de um atributo leva a criação de um novo registro.

4.2.1.1 Problemas levantados na base e soluções propostas

O histórico acima citado contém impressões sobre alterações dos dados pessoais que não tem qualquer impacto sobre a problemática de OS. Assim sendo, foi feita sua remoção usando os seguintes atributos: *CUSTOMER_KEY*, *CONTRACT_KEY* e *NAME* mantendo o último registro inserido na base, indicado pela coluna *DATA_ATIVACAO*. Portanto, cada linha representaria apenas um cliente com o seu respectivo produto.

Na base de dados Cliente, vários atributos não apresentam dados condizentes com as respectivas descrições. Em uma análise breve, nota-se que os campos estão preenchidos com valores incompatíveis aos esperados naquela coluna. Por exemplo, a coluna *EDUCATION_LEVEL*, nível escolar, está totalmente preenchida, mas grande parte dos dados se refere a valores numéricos como -1, -3 e -5, que não indicam nenhum grau de escolaridade. Para resolver essa situação são listados todos os valores possíveis dessa coluna, conforme o Apêndice A; em seguida, verifica-se se cada um deles tem sentido ou não – se não for observada essa característica, substitui-se por um valor nulo, como o caractere vazio (“”). Após realizar esse procedimento a coluna passa a ter somente valores coerentes como demonstrado comparando o Apêndice A com Apêndice B.

Constatou-se, ainda, a redundância de dados: para a coluna *NUMBER_OF_HOUSEMATES* (número de pessoas na casa), por exemplo, foram encontrados valores como

'ACIMA de 5' e *'Acima de 5'*. Semanticamente significam a mesma coisa, mas, pelo fato de a palavra 'Acima' não estar em caixa alta, como no primeiro caso, elas seriam interpretadas de formas diferentes em um processo de AM. Isso foi corrigido ao listar todos os valores possíveis do atributo, observados um a um, e se fazendo, sempre que necessário, substituições que padronizem referências semânticas idênticas. A listagem para esta coluna é ilustrada no Apêndice C e sua correção no Apêndice D.

Como a presença de dados informativos em uma coluna é fundamental para um bom processo de AM, será usado o seguinte critério para todas as bases: a coluna que tiver menos de 40% de dados válidos (dados condizentes a descrição de sua coluna, que sejam não-nulos), após a limpeza, será excluída, uma vez que, se há poucos dados para os registros, não é possível extrair muita informação.

Esses procedimentos de análises e de correções dos atributos foram empregados em todas as colunas de todas as bases.

4.2.1.2 Situação da base após limpeza

A seguir, os gráficos de barras indicam a proporção dos dados removidos após o tratamento dessa base (Figuras 4 e 5).

Cada barra representa um atributo, cujo nome está na abscissa inferior da respectiva coluna – aqui, o preenchimento indica que os dados são não nulos, ou seja, aparentemente válidos. A cor vermelha representa a proporção dos dados removidos após o processo de limpeza; a parte cinza demonstra a porção de dados válidos restantes; e a linha azul se refere ao ponto de corte, isto é, aos 40% de dados mínimos necessários para a coluna ser mantida.

O eixo das ordenadas à esquerda da Figura 4 representa a porcentagem dos dados relativos a cada atributo, variando de 0% (0,0) até 100% (1,0), por outro lado, o eixo das ordenadas à direita da figura representa o número total de registros correspondente a essas porcentagens. A abscissa superior representa a quantidade de dados referentes a cada atributo que permanece na base após executado o procedimento.

Pode-se citar, por exemplo, a coluna *CUSTOMER_KEY* na Figura 4 que, com a barra cinza totalmente preenchida, indica a inexistência de alterações, ao passo que a coluna *MARITAL_STATUS* (estado civil) foi reduzida para menos de 20% após a limpeza. Esta última será deletada por estar abaixo da linha azul, relativa ao ponto de corte.

Nesse prisma, a base adquire uma nova configuração após o processo conforme a Tabela 5, que ilustra os períodos anterior e posterior ao procedimento. A redução de registros se justifica, principalmente, após a ação de deletar o histórico e a diminuição de colunas por meio da remoção de atributos que não perpassaram o critério ora estabelecido.

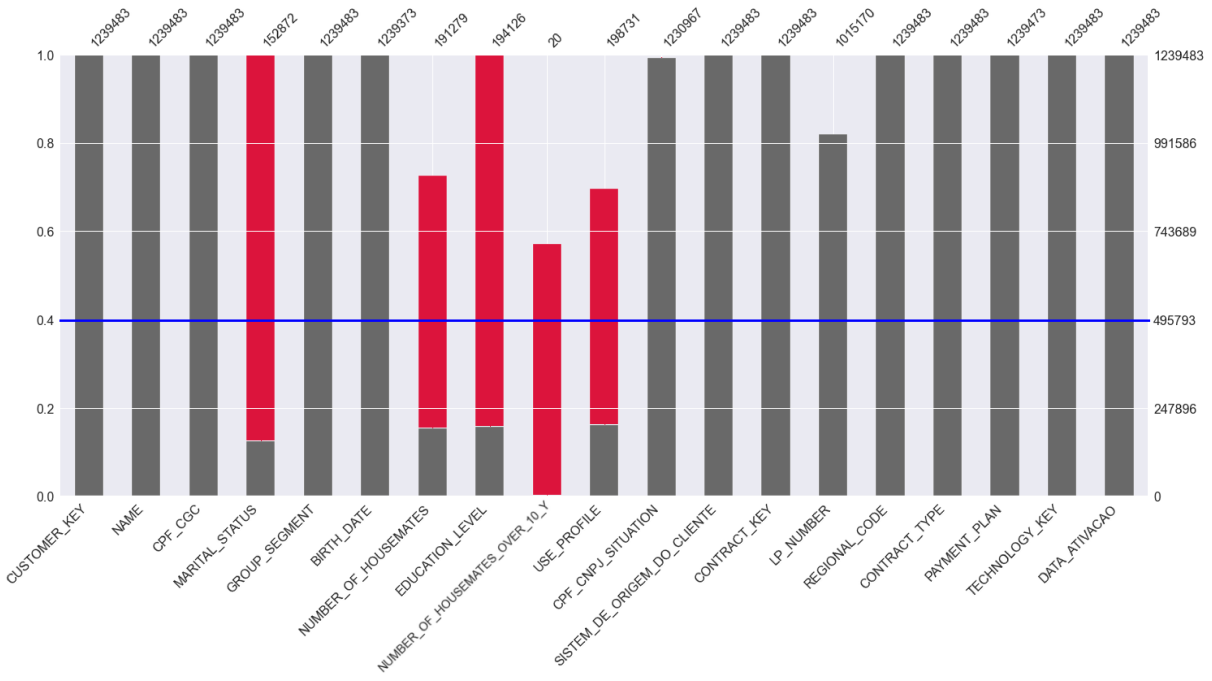

Figura 4 – Dados removidos da base de Cliente. Parte 1

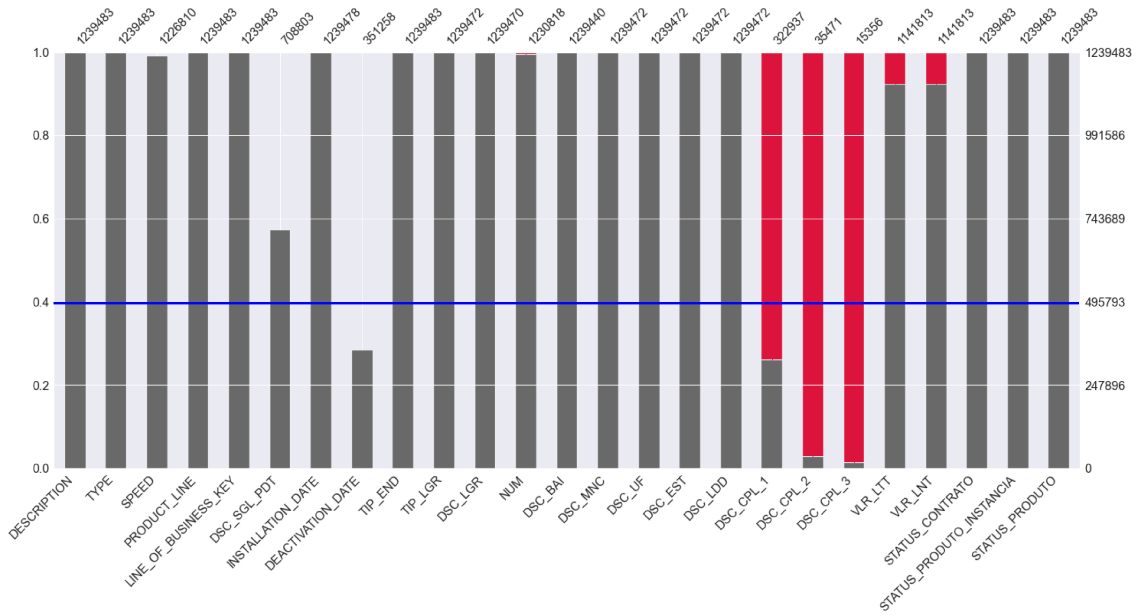

Figura 5 – Dados removidos da base de Cliente. Parte 2

Tabela 5 – Resumo da base Cliente antes e após a limpeza

Parâmetros da base	Antes	Depois	Redução (%)
Número de registros	1.239.483	583.886	53%
Número de atributos	44	35	20%

4.2.2 BD-2: Base de dados OS

Descrição da base: apresenta 470.948 registros com 47 colunas sobre OSs geradas pelo sistema de janeiro a abril de 2018. Os principais atributos são: chave do cliente (*CHAVE_CLIENTE*), que se conecta com *CUSTOMER_KEY* da base Cliente; número de protocolo de OS (*NUM_PROCOLO*); chave identificadora do contrato (*CHAVE_CONTRATO*); além de datas, códigos de abertura e fechamento de OS; e informações e descrições do técnico e do atendente da OS.

A tabela de OS se comporta como uma tabela fato e mantém um histórico como a base Cliente. O histórico é referente aos comentários feitos pelo técnico de campo, atendente ou gerados automaticamente pelo sistema da empresa durante o atendimento da OS.

Para detectar esse histórico, foi desenvolvido um processo para listar a possibilidade de duplicatas em determinado atributo com um valor específico, além de mostrar o que se modifica entre os registros. Isso foi feito, por exemplo, para o atributo "número de protocolo" com valor "173274782", o qual apresentava mais casos (92 ocorrências), mas os únicos dados que variam entre os registros se referem aos detalhes de campo. O resultado do procedimento para esse caso de exemplo é demonstrado no Apêndice E.

4.2.2.1 Problemas levantados na base e soluções propostas

A base OS possui problemas similares aos da tabela de cliente, em que são tratados da mesma forma. Como exemplo disso, a coluna CRIADOR_COMENTARIO possui valores redundantes, como "SysMobile" e "SYSMobile", padronizados posteriormente para "SYSMobile"; e valores inválidos, como "-3", os quais foram substituídos pelo carácter vazio "".

Houve também casos de atributos com valor constante para todos os registros, como a coluna *TIPO_ENDERECO*, que possui somente "INSTALAÇÃO", e *STATUS_OS*, com o valor "FECHADO". Colunas com tais características foram detectadas e removidas, pois, em um processo de AM, um atributo sem variação nos dados não representa nenhuma informação. Além disso, deletaram-se colunas sem informações, como *NUM_PROTOCOLO_UNICO*, por terem menos de 40% de dados válidos.

O histórico dos comentários técnicos foram considerados irrelevantes para o presente trabalho, visto que necessitaria de um procedimento mais aprofundado para lidar com linguagem natural. Para removê-lo, realizou-se o mesmo procedimento da base Cliente, com a remoção de duplicatas a partir do atributo *NUM_PROCOLO*, identificador da OS. Assim, as 92 ocorrências do protocolo "173274782", citado anteriormente, são reduzidas a apenas uma.

4.2.2.2 Situação da base após limpeza

A remoção dos registros e das colunas consta nas Figuras 6 e 7. Enquanto isso, na Tabela 6, caracteriza-se a base OS após o procedimento.

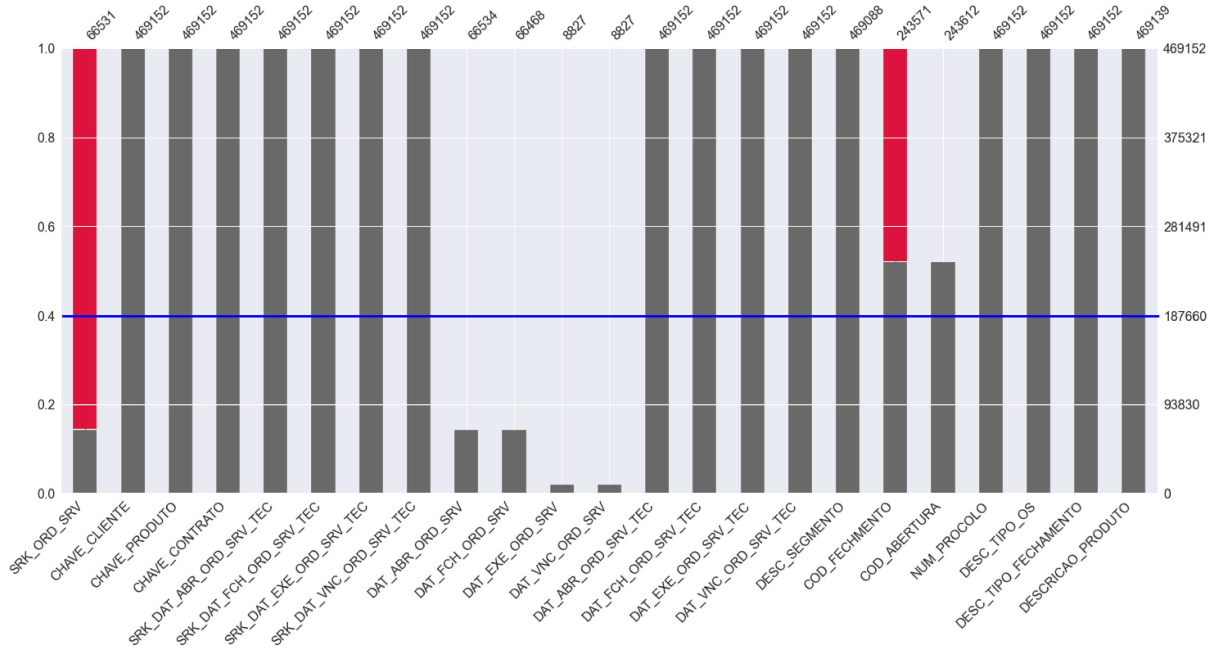

Figura 6 – Dados removidos da base de OS. Parte 1

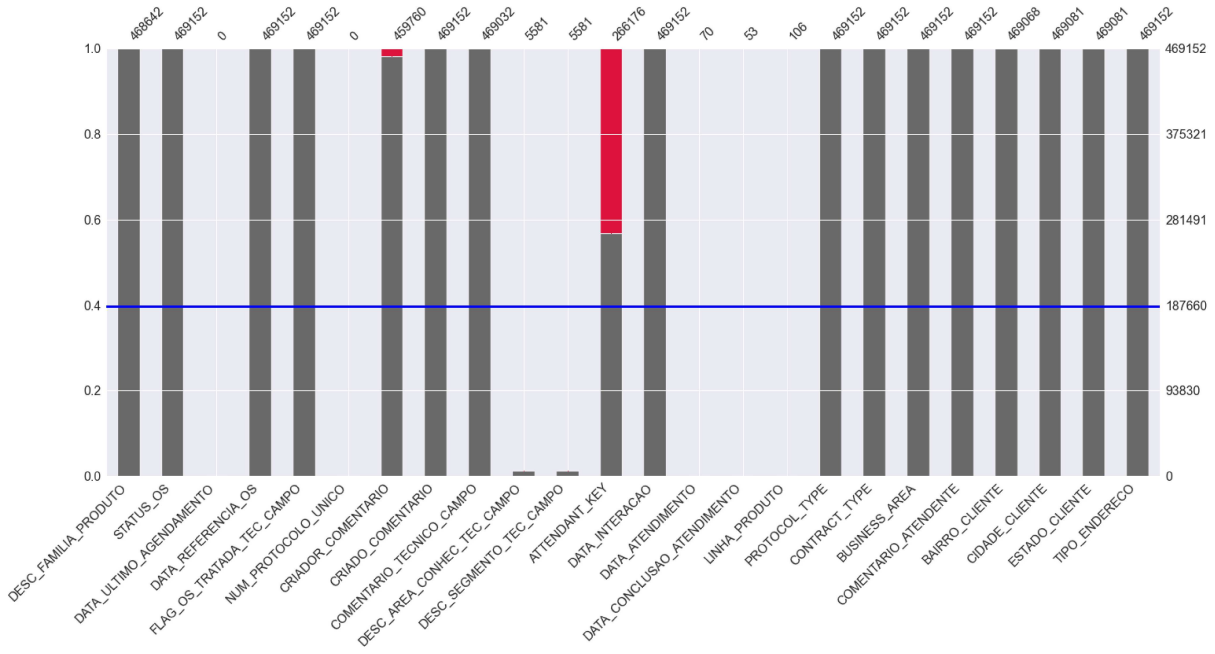

Figura 7 – Dados removidos da base de OS. Parte 2

Tabela 6 – Resumo da base OS antes e após a limpeza

Parâmetros da base	Antes	Depois	Redução (%)
Número de registros	682.233	107.535	84%
Número de atributos	47	35	25%

4.2.3 BD-3: Base de dados Rubi

Descrição da base: representa o inventário dos equipamentos de telecomunicações da empresa contendo seus nomes e localizações. Contém informações gerais que permite ligar um equipamento a determinados serviços, cliente e contratos. Possui 421.964 registros com 85 atributos sendo os principais:

- O par *banco_cliente* e *banco_produto*: se conectam com a base cliente.
- O par *inner_vlan* e *outer_vlan*: atributos numéricos de 4 dígitos que se unem à base Equipamentos.

4.2.3.1 Problemas levantados na base e soluções propostas

Dentre as tabelas elaboradas neste trabalho, a base Rubi apresenta a maior quantidade de valores incoerentes com o respectivo atributo. Em vários atributos foram feitas diversas substituições pelo carácter vazio, pois um dos valores mais recorrentes era "\n". Assim como na tabela OS, verificaram-se colunas com apenas um valor constante (atributos *tipo_estacao*, *velocidade_da_porta* e *proprietário*), as quais foram descartadas.

4.2.3.2 Situação da base após limpeza

Antes da limpeza, havia 10,6% de dados inválidos e, depois disso, 51,3%. Pelo fato de o procedimento ter invalidado vários campos dos registros, 41 colunas foram deletadas ao final. Tanto a base Rubi como a base de equipamentos não são provenientes de um DW, assim elas não mantêm histórico e portanto não houve registros deletados por haver duplicação de informação .Na sequência, os resultados obtidos são ilustrados nas Figuras 8 e 9 e na Tabela 7:

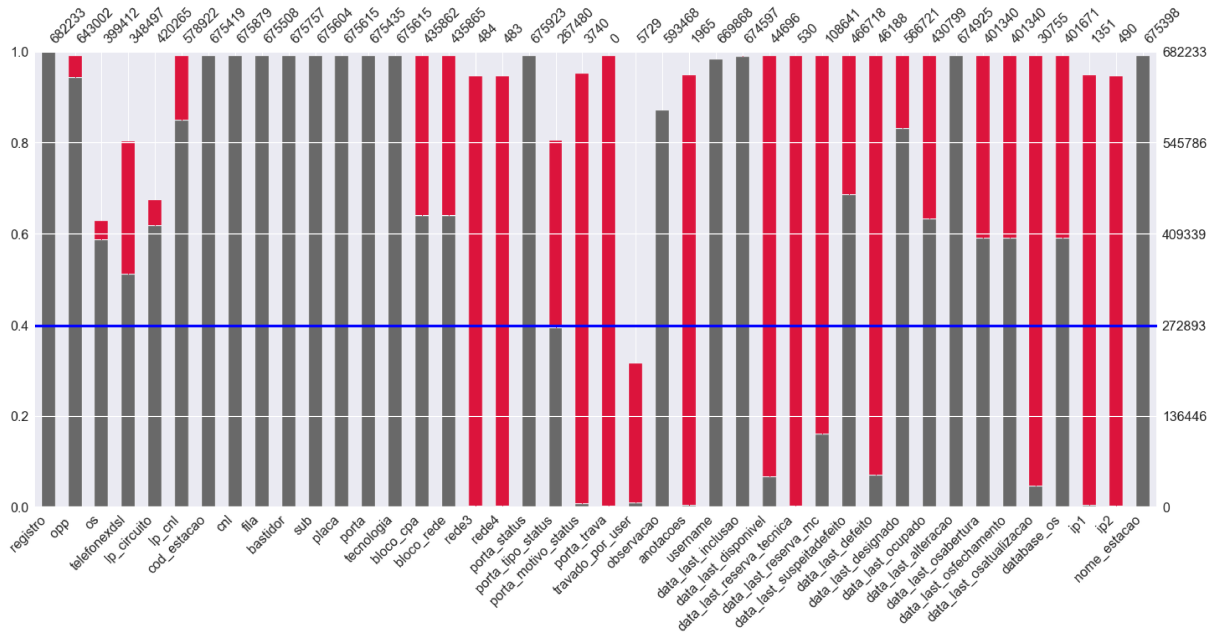

Figura 8 – Dados removidos da base Rubi. Parte 1

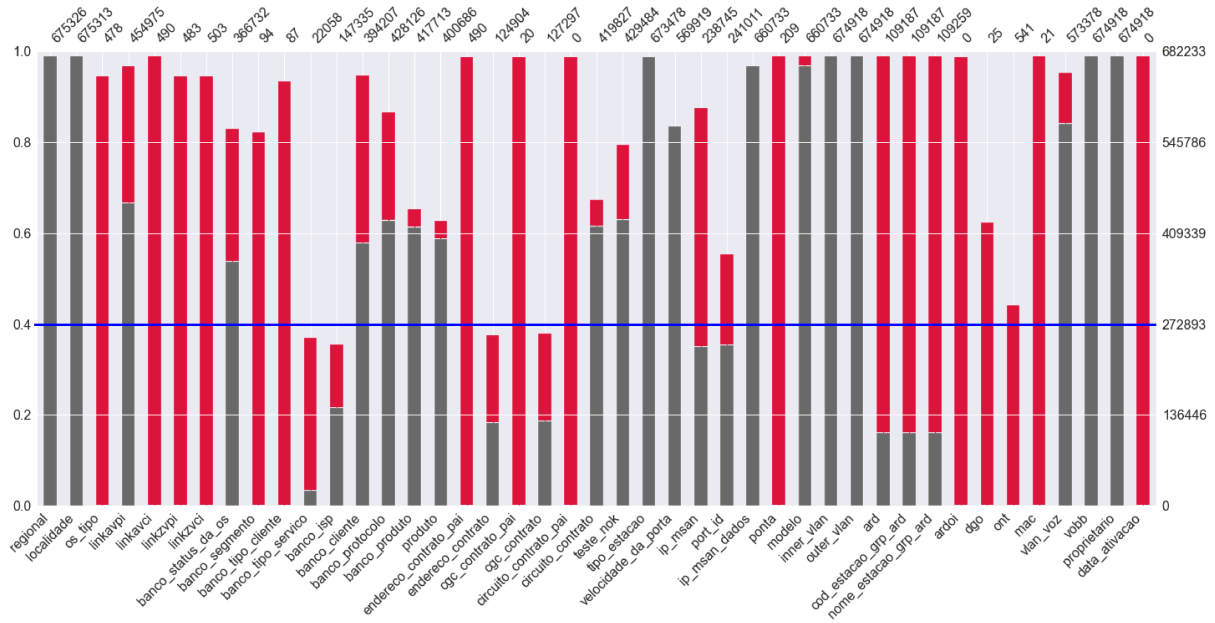

Figura 9 – Dados removidos da base Rubi. Parte 2

Tabela 7 – Resumo da base Rubi antes e após a limpeza

Parâmetros da base	Antes	Depois	Redução (%)
Número de registros	682.233	682.233	0%
Número de atributos	85	44	48%

4.2.4 BD-4: Base de dados Equipamentos

Descrição da base: composta por quatro tabelas separadas, contém informações técnicas sobre os equipamentos da empresa instalados pela cidade e suas configurações internas.

- **Tabela DEVICE**: cada registro de *DEVICE* corresponde a uma caixa de rede fixada nos postes da cidade. Possui informações como modelo, usuário e senha para acessá-la.
- **Tabela PORT**: apenas um DEVICE possui várias portas representadas na tabela PORT, as quais interligam o poste até as casas dos clientes. Apresenta os atributos *VLAN_IN* e *VLAN_OUT*, que referenciam *inner_vlan* e *outer_vlan* da base Rubi.
- **Tabela PROFILE**: descreve o perfil de rede de cada porta, ao indicar os índices máximos de upload (*MAX_RATE_UP*) e download (*MAX_RATE_DOWN*) configurados para o cliente.
- **Tabela SNAPSHOTS**: registra *snapshots* (fotografias de um momento) das portas para sinais de rede. Essa fotografia é feita de três em três dias e a tabela possui dados de dezembro de 2017 à novembro de 2018. São registrados os seguinte sinais de rede.
  - Taxa de sinal de *upload* e *download*: medida expressa em decibéis (dB) que indica a força da conexão do dispositivo do poste com o do cliente.
  - Taxa de atenuação de *upload* e *download*: medida também em dB, representa a perda do sinal de rede, ou seja, o ruído.
  - Taxa corrente de *upload* e *download*: taxa relativa ao momento de *upload* e *download*, medida em *Kilobits* por segundo (Kbps).
  - Taxa atingível de *upload* e *download*: variação da taxa de *upload* e *download*, com grande influência do plano do cliente e expressa em Kbps.

4.2.4.1 Problemas levantados na base e soluções propostas

Cada uma das quatro tabelas da base Equipamentos apresentaram menos problemas em relação às outras analisadas nesta pesquisa, sem dados redundantes, nem inválidos. Por isso, nelas não ocorreram nenhuma substituição. O único problema verificado demonstra que cinco atributos em *DEVICE* e dois em *PORT* possuíam menos dados do que o critério estabelecido para as colunas; assim, foram deletados, como ilustram as Figuras 10 e 11:

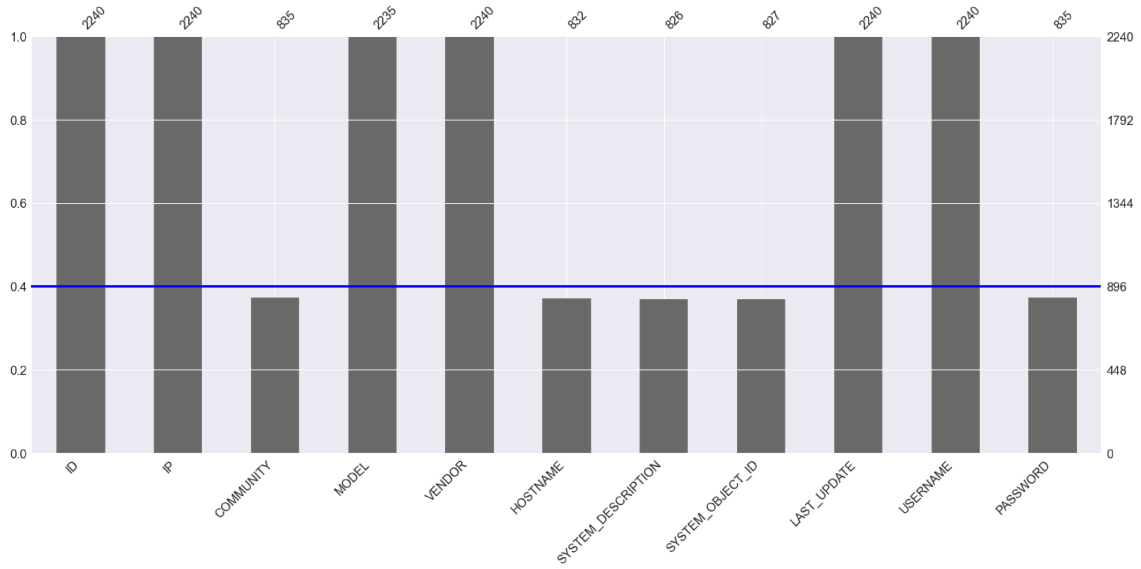

Figura 10 – Colunas removidas da base DEVICE

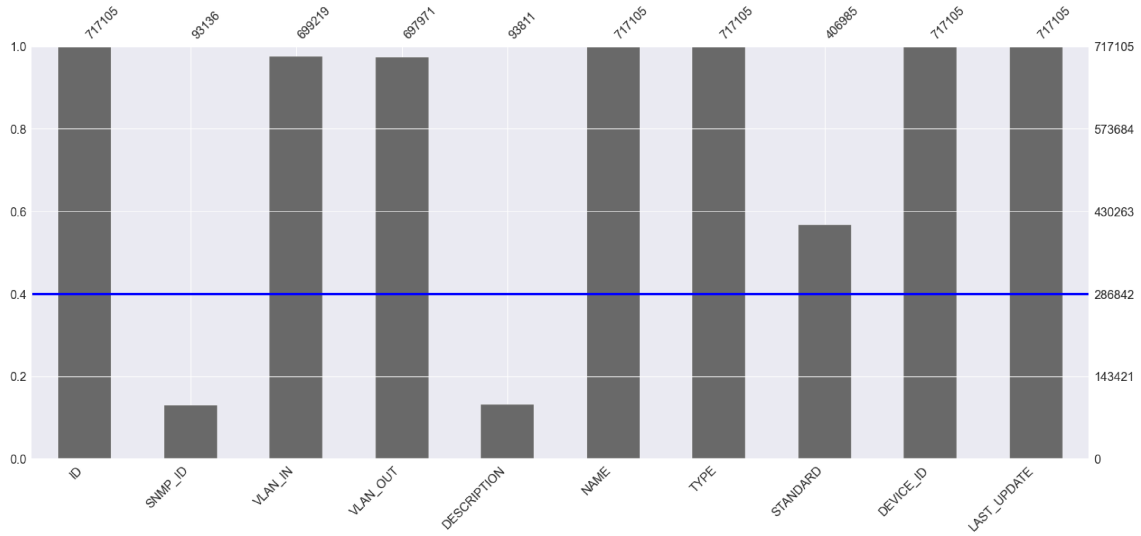

Figura 11 – Colunas removidas da base PORT

4.2.4.2 Junção das tabelas da base de Equipamentos

Após o processo de limpeza, realiza-se a junção das quatro tabelas para centralizar seus dados. Assim, cada registro da nova base Equipamento terá os dados agrupados de *snapshots*, perfil da rede, configuração do *DEVICE* e dados da porta que permitem se conectar à base Rubi. Isso é possível pelo fato de as tabelas possuírem referências (chave estrangeira) aos identificadores (chave primária) entre si, de forma a se unirem conforme a Figura 12. A cada um dos dos atributos foi acrescentado como préfixo o nome de sua respectiva tabela para não confundir colunas com mesmo nome. A junção é feita na seguinte ordem:

1. Junção pelo atributo *SNAP.PROFILE_ID* de SNAPSHOT com a coluna *PROFILE.ID* de PROFILE.
2. Junção pelo atributo *SNAP.PORT_ID* de SNAPSHOT com a coluna *PORT.ID* de PORT.
3. Junção pelo atributo *PORT.DEVICE_ID* de PORT com a coluna *DEVICE.ID* de DEVICE.

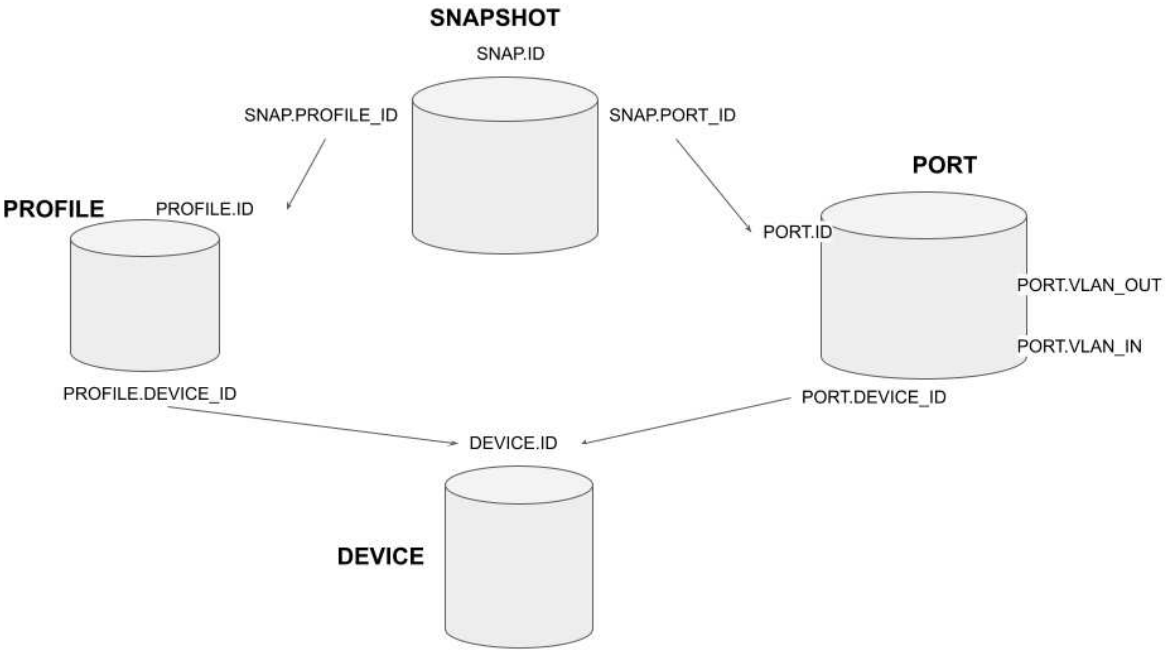

Figura 12 – Junção das tabelas de equipamentos

4.3 Base Join: junção das bases

A junção das bases de cliente, OS, Rubi e equipamentos será intitulada como "base Join". Ela sintetiza as informações mais relevantes sobre os clientes, como ter ou não OS e snapshots da própria rede. Ao desenvolvê-la foram contabilizados 31.485 clientes.

Nas próximas seções, são apresentados duas formas de criar a "base join". A primeira é a partir de um algoritmo, descrito pelo pseudocódigo 1 na subseção 4.3.1 que mostra o conjunto dos clientes pertencentes à base Join representados pelo vetor Cliente.

Na sequência, a seção 4.3.2 diz respeito ao seu desenvolvimento usando conceitos de SQL para realizar as junções da bases.

4.3.1 Pseudocódigo vetor do cliente

Algoritmo 1: Pseudocódigo vetor do cliente

```
1  início
2      Inicializar o vetor de clientes VetC.
3      para cada cliente ← base_Cliente faça
4          se cliente[nome] está em base_Rubi então
5              rubi_c ← registro do cliente na base Rubi
6              se rubi_c[vlan_in] está em base_Equipamentos e
               rubi_c[vlan_out] está em base_Equipamentos então
7                  equipamentos_c ← registro do cliente na Base Equipamentos
8                  se cliente[nome] esta em base_OS então
9                      os_c ← registro do cliente na base OS
10                 fim
11                 senão
12                     os_c ← registro vazio da base OS
13                 fim
14                 VetC armazena (cliente, rubi_c, equipamentos_c, os_c)
15             fim
16         fim
17     fim
18     retorna VetC
19 fim
```

O algoritmo 1 descreve o processamento para elaborar o vetor cliente. Na linha 2, cria-se o vetor que inicializa como vazio; em seguida, é feita uma sequência de verificações para cada cliente da base Cliente: se o nome dele estiver presente em algum registro da base Rubi (linha 4), ele é armazenado (linha 5) e, no registro da base Rubi que pertence ao cliente, caso o par *vlan_in* e *vlan_out* estiverem na base Equipamentos (linha 6)

armazena-se o registro de equipamento do cliente (linha 7).

A última verificação compreende o fato de ter ou não OS na base OS (linha 8): em caso positivo (linha 9), armazenam-se os dados e, se não houver, cria-se um registro vazio de OS (linha 10). Ao final do processo, os registros são armazenados de maneira conjunta e adicionados ao vetor cliente (linha 14). Por fim, obtêm-se 150 colunas: 1 a 35 pertencem à base Cliente; 36 a 79, Rubi; 80 a 115, Equipamentos; e 116 a 150, OS.

4.3.2 Junção de todas as bases

Como dito anteriormente, o objetivo deste trabalho é gerar uma base de dados central e confiável, proveniente das quatro bases de dados após a etapa de pré-processamento, para garantir a confiabilidade e a otimização. A partir da junção, forma-se a base em que cada registro contém os dados pessoais do cliente, atrelados aos dados de rede (modems internos e *snapshots*), além da possibilidade (ou não) de OS.

A seguir, é descrita a junção das bases em conformidade aos conceitos de banco de dados (linguagem SQL) e ao esquema da Figura 13.

1. Realiza-se a junção *INNER JOIN* das colunas (*CHAVE_CLIENTE*, *DESCRICAO*), da base Cliente, com os atributos (*banco_cliente*, *banco_produto*) da base Rubi.
2. Em seguida, há *INNER JOIN* dos atributos (*inner_vlan*, *outer_vlan*) da base Rubi com as colunas (*PORT.VLAN_IN*, *PORT.VLAN_OUT*) da base Equipamentos.
3. Por fim, faz-se a operação de junção *LEFT JOIN* dos atributos (*CHAVE_CLIENTE*, *CONTRACT_KEY*) da base resultante da etapa anterior com as colunas (*CUSTOMER_KEY*, *CHAVE_CONTRATO*) da base OS.

O *LEFT JOIN* garante que, mesmo se o cliente não tiver aberto nenhuma OS, os dados e as informações de rede ainda permanecerão na base de dados final. Aqui, os dados de OS serão nulos e indicarão a inexistência de OS. Ao final do processo, parte das colunas usadas para o cruzamento de dados são deletadas para evitar repetição.

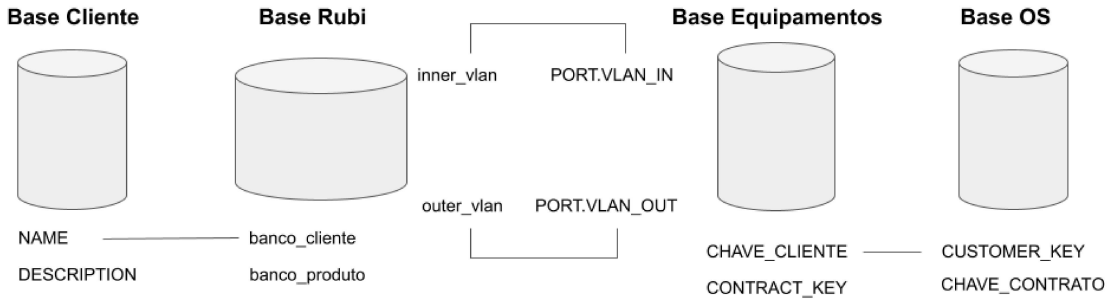

Figura 13 – Esquema da junção de todas as bases

4.4 Como a base é usada para o protótipo

Nesta seção será conduzido um breve exemplo de como a base Join produzida foi utilizada no contexto global do projeto desenvolvido: como base para construção de um modelo de AM.

A pergunta a ser investigada é a predição de um dado cliente abrir ou não OS nos 15 dias subsequentes ao dia 01/04/2018.

4.4.1 Formatação dos dados

Idealizaram-se duas abordagens para investigar os dados. A longo prazo, busca-se responder à pergunta de acordo com três meses, de janeiro a março; e a de curto prazo se refere a um período menor, isto é, somente às duas últimas semanas.

Para o presente projeto, consideraram-se os atributos preditivos de dados relativos à quantidade de OS e às métricas de rede. Para o atributo alvo que se pretende predizer nesse contexto, houve a seguinte simplificação: 0, se não abria OS; e 1, se abria pelo menos uma OS nas próximas duas semanas.

Foram empregados os atributos *DAT_ABR_ORD_SRV_TEC* de OS, para verificar quando ocorreu a abertura de OS; e *SNAP.CREATION_DATE*, para conhecer a data dos snapshots, ou seja, da situação da rede. Para cada estratégia, definiu-se o cálculo da quantidade de OS e das médias e dos desvios-padrão, a saber:

- Contagem de OS: a longo prazo, é feita para cada um dos três últimos meses, e, para curto prazo, do dia 1º ao 3º, 4º ao 7º e 8º ao 15º.
- Média dos sinais do modem: a longo prazo, para cada uma das quatro últimas semanas, e a curto prazo, a cada dois dias nos últimos oito dias.
- Desvio padrão dos dados do modem: a longo prazo, a cada mês, e a curto prazo, a cada cinco dias nos últimos 15 dias.

A lista completa dos atributos gerados para ambas as abordagens está presente nos Apêndices F e G.

4.4.2 Desenvolvimento dos Modelos

O desenvolvimento do modelo foi a etapa principal do escopo maior do protótipo. Mais detalhes são encontrados no seguinte artigo de Pereira et al. (2019). A melhor estratégia encontrada foi a longo prazo, e foi criado um modelo que obteve acurácia aproximadamente 88%.

A Figura 14 mostra quais etapas foram realizadas nesse trabalho e quais foram desenvolvidas no artigo em vista a todo o projeto de pesquisa.

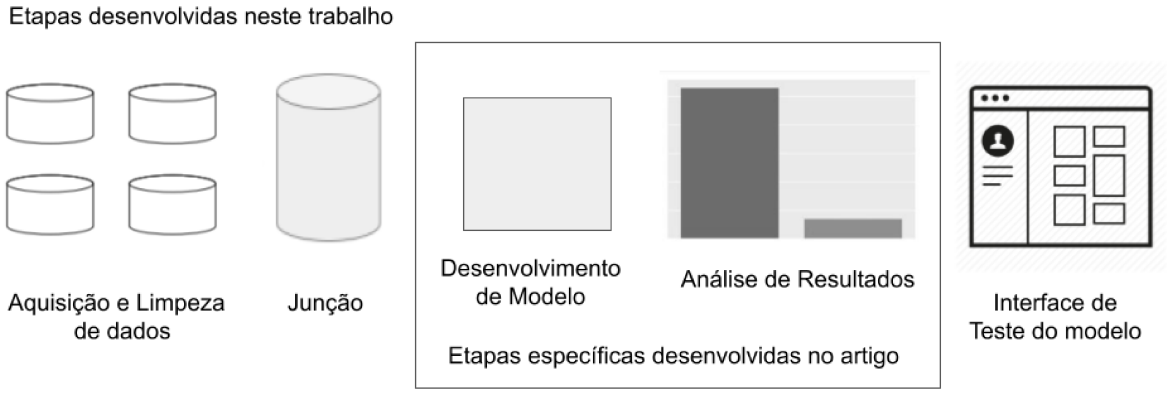

Figura 14 – Todas as etapas do projeto de pesquisa

4.5 Desenvolvimento de programa para utilização da base e predição

Após a elaboração da base Join e o modelo treinado, foi possível desenvolver um sistema para utilizá-los e predizer para quaisquer dados e clientes.

Nesse entremeio, criaram-se dois sistemas separados que, ao se unirem, permitem a utilização de tal operação: uma interface gráfica que permite inserir dados e pesquisar por clientes da base Join desenvolvido; e um programa que acessa a base e busca os dados de cliente e o modelo para realizar predições. Este último foi desenvolvido em *python* como uma Interface de Programação de Aplicação (em inglês, API) para manter seguros a base e o modelo, em se tratando de acessos externos, exceto pelo sistema.

4.5.1 Funcionalidades do sistema

O sistema desenvolvido pode ser acessado no *link* <https://order-service-web.netlify.app/>. Os repositórios contendo os códigos da interface e da API estão respectivamente em <https://github.com/rafanthx13/order-service-web> e <https://github.com/rafanthx13/order-service-api>. As funcionalidades disponibilizadas são:

- Predizer a abertura de OS para quaisquer dados inseridos.
  - São enviados os dados para o modelo que, por sua vez, retorna a predição.
- Buscar clientes da base Join e seus dados.
  - Os nomes dos clientes são listados automaticamente na página.

  - Os dados são buscados após a escolha de um cliente.

- Caso feita a predição para um cliente da base Join, por ter sido analisada, pode-se verificar se o modelo acertou ou não.

Por questões de segurança e sigilo, os nomes dos clientes foram renomeados. Nesse caso, são listados somente 200 primeiros clientes da base Join para fins de testes.

Na Figura 15 é apresentado a interface do sistema, permitindo inserir cada um dos dados ou selecionar dados da base Join. A Figura 16 apresenta o resultado quando acerta a predição de abertura de um cliente selecionado.

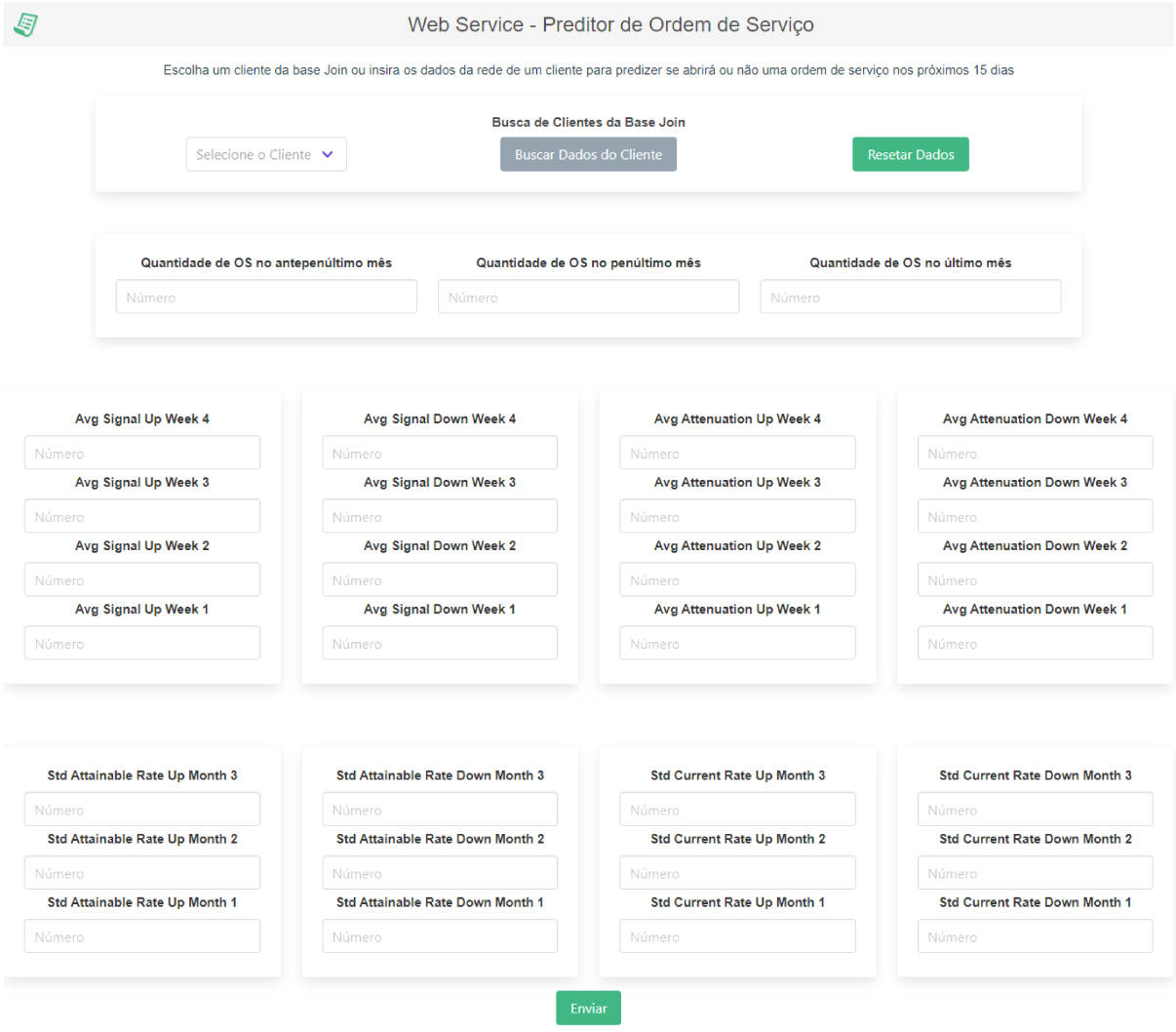

Figura 15 – Interface do sistema para teste do modelo preditivo

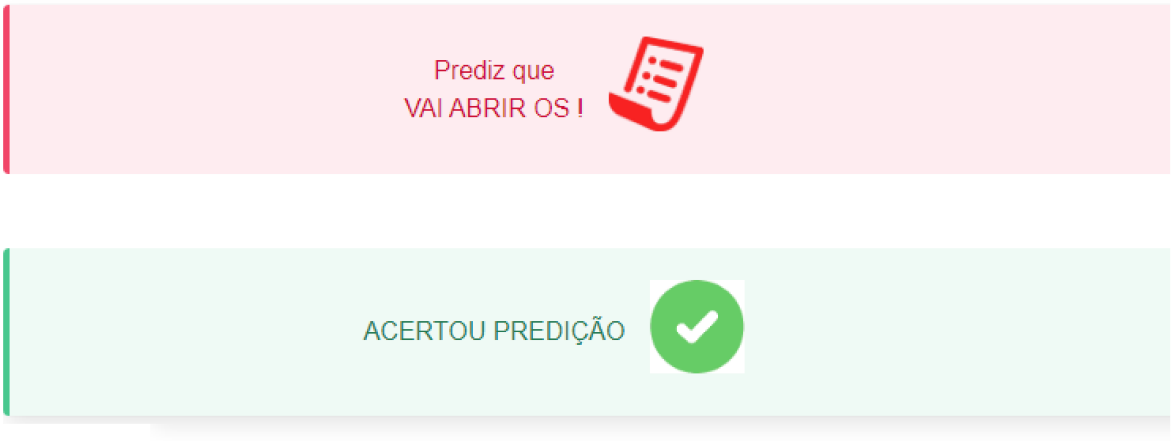

Figura 16 – Interface do sistema quando acerta a predição de abertura de OS

4.5.2 Resultados obtidos

Sobre a base Join reformulada, o modelo possui a acurácia de 88.06% dos dados (Figura 17). No que tange aos casos em que há a abertura de OS, 67,76% (Figura 18), e quando isso não ocorre, 88,55% (Figura 19). Todos os casos são expressos com a respectiva quantidade de registros e porcentagem na Figura 20.

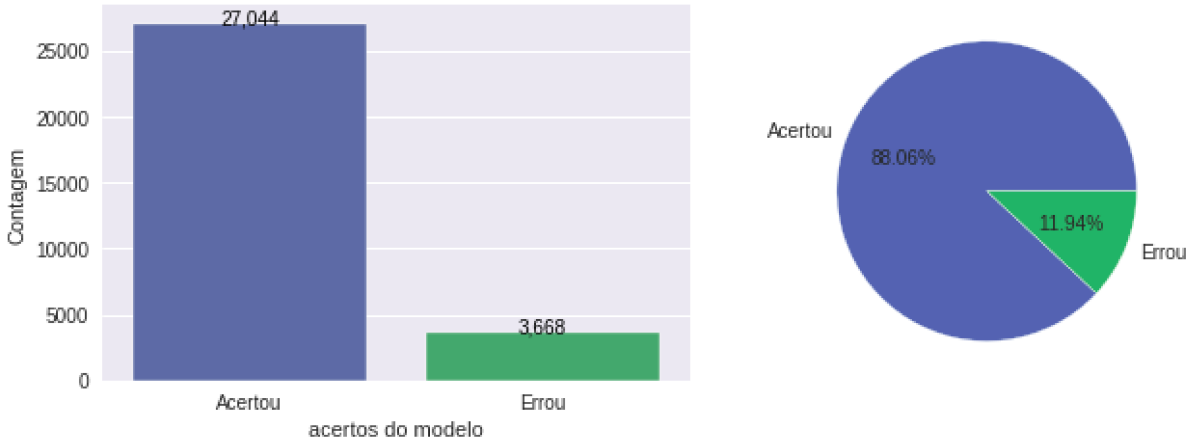

Figura 17 – Acertos do modelo

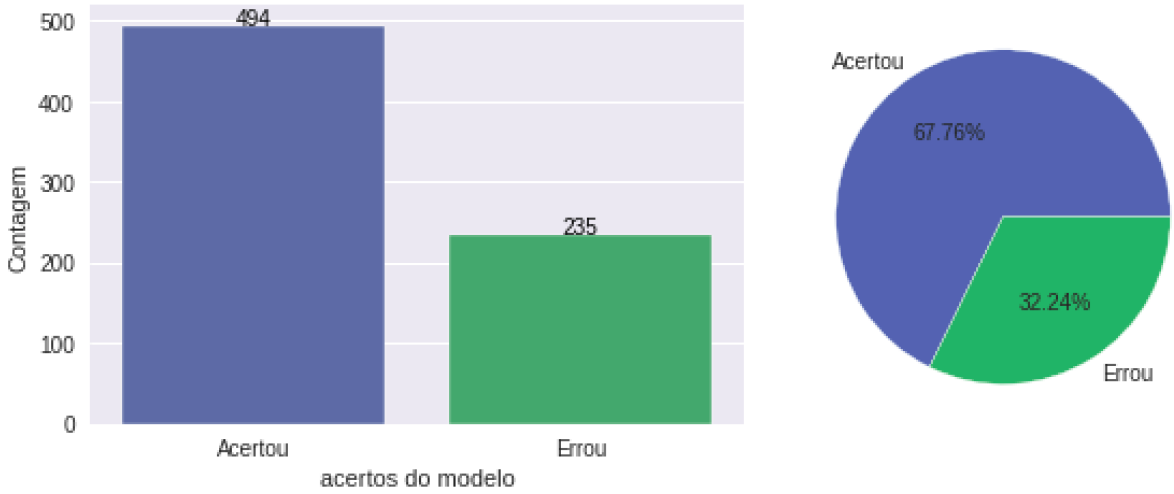

Figura 18 – Acertos do modelo quando se abre OS

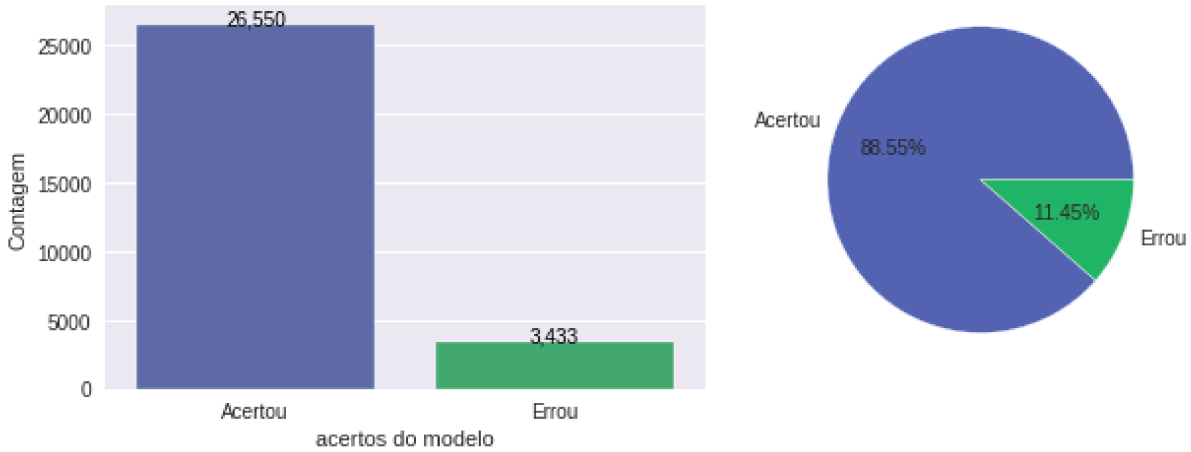

Figura 19 – Acertos do modelo quando não se abre OS

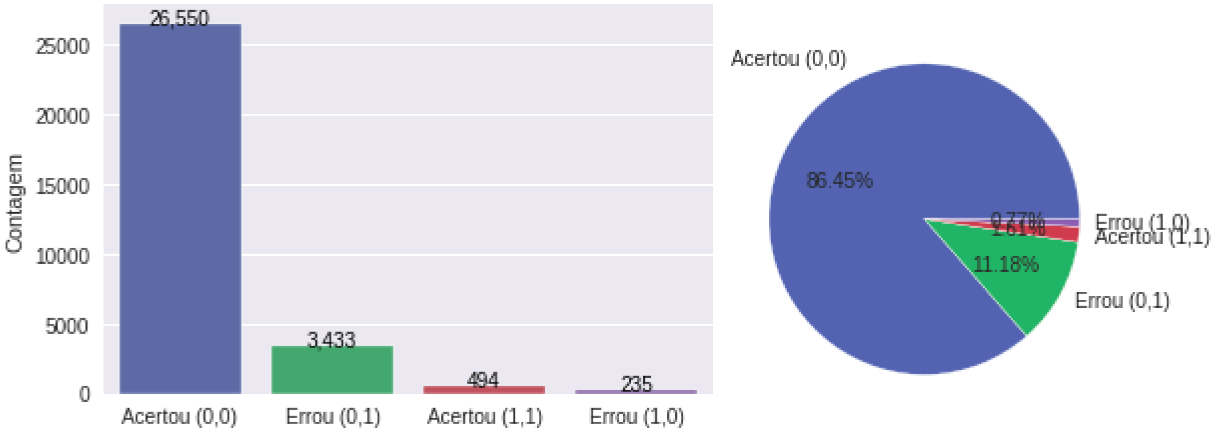

Figura 20 – Acertos e erros do modelo

5 Conclusão

Este trabalho apresentou algumas estratégias de pré-processamento de dados aplicados a dados recebidos da empresa Algar Telecom. Nesse sentido, foram desenvolvidos *scripts* e procedimentos para realizar a análise de dados, como gráficos e detecção de histórico, além da aplicação de técnicas de limpeza de dados, a exemplo da remoção de duplicatas, da substituição de valores inconsistentes e do tratamento dos dados faltantes. Essas técnicas foram empregadas em quatro bases e, após a junção das quatro tabelas da base Equipamentos em apenas uma, desenvolveu-se a junção final de todas as bases, que resultou em uma base unificada e otimizada.

O presente estudo de caso faz parte de um projeto global que requereu a remodelagem dos dados para a predição de abertura de OS em cinco dias subsequentes. Foi realizada a reestruturação para duas abordagens de curto e longo prazo e, para cada intervalo, calcularam-se médias e desvios padrão dos sinais de rede, além da contagem de OS, cujos dados fundamentam o modelo de AM desenvolvido por Pereira et al. (2019), que revela resultados satisfatórios para a predição de abertura de OS. Por fim, é desenvolvida uma página web para utilização desse modelo para predizer tanto em relação aos clientes da base de dados quanto aos novos dados.

Dentre as dificuldades encontradas para a realização deste trabalho, destaca-se o grande esforço requerido para o acesso aos dados e para a compreensão do que eles representam. Isso se deveu, principalmente, à inevitável rotatividade ao longo do tempo de atividades exercidas por funcionários de grandes empresas como a Algar Telecom. Tal rotatividade impede que um mesmo funcionário acompanhe o histórico da evolução das bases de dados da empresa. Apesar disso, vale ressaltar o grande apoio recebido de vários funcionários da empresa que dedicaram muito de seu tempo para tornar possível o cumprimento dos objetivos da presente proposta.

5.1 Trabalhos Futuros

Diante dos resultados e da experiência adquirida, é possível indicar alguns trabalhos futuros ou outros caminhos para este projeto, com vistas a continuar as atividades ora estabelecidas:

- Desenvolver e analisar os comentários de técnicos durante o tratamento de OS para obtenção de algum conhecimento via processamento de linguagem natural.
- Adicionar outros atributos ou métricas à base final reformulada.

- Criar outros conjuntos de dados por meio de técnicas de redução de dimensionalidade, por exemplo.
- Utilizar o modelo sobre os dados atuais da empresa, coletar resultados e propor melhorias para o modelo, como otimização de parâmetros do modelo de AM.
- Testar técnicas mais avançadas de AM, como algoritmos de aprendizagem profunda.
- Desenvolver um *pipeline* de dados, ou seja, automatizar o pré-processamento realizado neste estudo.

Referências

ANATEL. *Banda larga fixa foi o único serviço de telecomunicações que apresentou crescimento em 2016*. 2017. Portal da Anatel. Disponível em: <http://www.anatel.gov.br/institucional/component/content/article?id=1512>. Acesso em: 24 maio 2018. Citado na página 13.

BATISTA, G. E. d. A. P. A. *Pré-processamento de dados em aprendizado de máquina supervisionado*. Monografia (Tese Doutorado) — Universidade de São Paulo, São Carlos, 2003. Disponível em: <https://www.teses.usp.br/teses/disponiveis/55/55134/tde-06102003-160219/pt-br.php>. Acesso em: 20 fev 2022. Citado 3 vezes nas páginas 22, 23 e 24.

BRAZILIENSE, C. *Brasileirão tem a pior média de gols entre 17 ligas nacionais da Europa e América do Sul*. 2021. Correio Braziliense. Disponível em: <https://blogs.correiobraziliense.com.br/dribledecorpo/brasileirao-tem-a-pior-media-de-gols-entre-17-ligas-nacionais-da-europa-e-america-do-sul/>. Acesso em: 29 fev 2022. Citado na página 21.

CANALTECH. *Estas são as melhores operadoras de telecom do Brasil segundo a Anatel*. 2019. CanalTech. Disponível em: <https://canaltech.com.br/telecom/estas-sao-as-melhores-operadoras-de-telecom-do-brasil-segundo-a-anatel-110054/>. Acesso em: 29 outubro 2019. Citado na página 12.

DATAMEER. *Para empresas de telecom, big data é o "petróleo"da nova economia*. 2013. InforGrafico. Disponível em: <https://www.datameer.com/wp-content/uploads/2015/09/State_of_the_Industry.pdf>. Acesso em: 08 de novembro 2019. Citado na página 12.

DELVA, P. *LSQL JOIN: Aprenda INNER, LEFT, RIGHT, FULL e CROSS*. 2020. Alura. Disponível em: <https://www.alura.com.br/artigos/join-em-sql>. Acesso em: 01 de fev 2022. Citado 2 vezes nas páginas 6 e 17.

ELMASRI, R.; NAVATHE, S. B. *Fundamentals of Database Systems*. 7th. ed. [S.l.]: Pearson, 2015. ISBN 0133970779. Citado 2 vezes nas páginas 15 e 16.

ESTADAO. *Velho mercado, novos concorrentes - Em um mercado estável, crescimento fica por conta do surgimento de novos concorrentes, principalmente na área de banda larga fixa*. 2019. Estadão - Empresa Mais. Disponível em: <https://publicacoes.estadao.com.br/empresasmais2018/setor/telecom/>. Acesso em: 29 outubro 2019. Citado na página 12.

EXPERIAN, S. 2018. Serasa Experian. Disponível em: <https://www.serasaexperian.com.br/amplie-seus-conhecimentos/blog/mais-de-90\-das-empresas-brasileiras-consideram-dados-na-definicao-da-estrategia-de-negocios\-aponta-experian>. Acesso em: 10 outubro 2019. Citado na página 12.

FACELI, K. *Inteligência artificial: uma abordagem de aprendizado de máquina*. Grupo Gen - LTC, 2011. ISBN 9788521618805. Disponível em: <https://books.google.com.br/books?id=4DwelAEACAAJ>. Citado 5 vezes nas páginas 20, 21, 22, 23 e 24.

FERREIRA, R. M. *Pré-processamento de dados de trajetórias para mineração de dados e análise de similaridade*. Monografia (Monografia) — Universidade Federal de Santa Catarina, 2017. Disponível em: <https://repositorio.ufsc.br/handle/123456789/182211>. Acesso em: 20 fev 2022. Citado na página 26.

FRANCISCHELLI, R. A. *Estudo sobre a influência da extração de dados na web na descoberta de conhecimento estratégico relevante*. Monografia (Monografia) — Universidade de Caxias do Sul, 2013. Disponível em: <https://repositorio.ucs.br/xmlui/handle/11338/1216>. Acesso em: 20 fev 2022. Citado na página 26.

GANTZ, J.; REINSEL, D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. In: . [S.l.: s.n.], 2012. Citado na página 12.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123814790. Citado 7 vezes nas páginas 6, 18, 19, 20, 21, 22 e 25.

KANG, E. *Customer experience matters in telecom*. 2013. LivePerson. Disponível em: <https://www.liveperson.com/connected-customer/posts/ideal-online-experience-what-it-takes-consumers-click-not-abandon>. Acesso em: 29 outubro 2019. Citado na página 12.

KIMBALL, R.; ROSS, M. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. 2nd. ed. USA: John Wiley and Sons, Inc., 2002. ISBN 0471200247. Citado na página 18.

KLEMZ, L. *What Is Customer Experience?* 2019. Resultados Digitais. Disponível em: <https://resultadosdigitais.com.br/blog/o-que-e-churn/>. Acesso em: 29 outubro 2019. Citado na página 12.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. *Artif. Intell.*, Elsevier Science Publishers Ltd., GBR, v. 97, n. 1–2, p. 273–324, dec 1997. ISSN 0004-3702. Disponível em: <https://doi.org/10.1016/S0004-3702(97)00043-X>. Citado na página 24.

MAURO, A. D.; GRECO, M.; GRIMALDI, M. What is big data? a consensual definition and a review of key research topics. In: . [S.l.: s.n.], 2014. Citado na página 12.

MEYER, C.; SCHWAGER, A. Understanding customer experience. *Harvard business review*, v. 85, p. 116–26, 157, 03 2007. Citado na página 12.

PEREIRA, F. et al. Feature-based time series classification for service request opening prediction in the telecom industry. In: _____. [S.l.: s.n.], 2019. p. 120–132. ISBN 978-3-030-30243-6. Citado 2 vezes nas páginas 41 e 46.

RAMAKRISHNAN, R.; GEHRKE, J. *Database Management Systems*. 2nd. ed. USA: McGraw-Hill, Inc., 2000. ISBN 0072440422. Citado 2 vezes nas páginas 15 e 17.

SILVA, D. A. d. *Aplicação de técnicas de pré-processamento e agrupamento na base de dados de benefícios previdenciários do ministério público do trabalho*. Monografia (Monografia) — Universidade Federal de Uberlândia, 2018. Disponível em: <https://repositorio.ufu.br/handle/123456789/22118>. Acesso em: 20 fev 2022. Citado na página 26.

TAN, P.-N. et al. *Introduction to Data Mining (2nd Edition)*. 2nd. ed. [S.l.]: Pearson, 2018. ISBN 0133128903. Citado 2 vezes nas páginas 21 e 23.

TIINSIDER, R. *IDC prevê que big data e analytics valerão US$ 215,7 bilhões em 2021*. 2021. TI Insider. Disponível em: <https://tiinside.com.br/18/08/2021/idc-preve-que-big-data-e-analytics-valerao-us-2157-bilhoes-em-2021/>. Acesso em: 10 de outubro 2021. Citado na página 12.

TURBAN, E. et al. *Business Intelligence: Um enfoque gerencial para a inteligência do negócio*. Grupo A - Bookman, 2009. ISBN 9788577804252. Disponível em: <https://books.google.com.br/books?id=_Uvqyr32hlMC>. Citado na página 18.

UMEZAWA, C. O. *Automação do recolhimento e da disponibilização das informações de Automação de Processos Robóticos*. Monografia (Monografia) — Universidade Federal de Santa Catarina, 2021. Disponível em: <https://repositorio.ufsc.br/handle/123456789/228603>. Acesso em: 20 fev 2022. Citado na página 26.

Apêndices

APÊNDICE A – todos os valores não-nulos possíveis para a coluna *EDUCATION_LEVEL*

Possui valores incompatíveis com o nível educacional, como -1, 3 e -5.

	EDUCATION_LEVEL	count	percentage
0	-3	708803	57.19%
1	-1	333954	26.94%
2	2 GRAU COMPLETO	65188	5.26%
3	1 GRAU INCOMPLETO (PRIMARIO)	40096	3.23%
4	SUPERIOR COMPLETO	31126	2.51%
5	1 GRAU COMPLETO (PRIMARIO E GINASIO)	24621	1.99%
6	2 GRAU INCOMPLETO	15013	1.21%
7	SUPERIOR INCOMPLETO	14232	1.15%
8	POSGRADUACAO	2898	0.23%
9	-5	2600	0.21%
10	SEM INSTRUCAO	952	0.08%

APÊNDICE B – Valores não-nulos da coluna *EDUCATION_LEVEL* após processo de limpeza

Após o processo de limpeza a coluna não apresenta valores que não indiquem o nível educacional.

	EDUCATION_LEVEL	count	percentage
0	2 GRAU COMPLETO	65188	33.58%
1	1 GRAU INCOMPLETO (PRIMARIO)	40096	20.65%
2	SUPERIOR COMPLETO	31126	16.03%
3	1 GRAU COMPLETO (PRIMARIO E GINASIO)	24621	12.68%
4	2 GRAU INCOMPLETO	15013	7.73%
5	SUPERIOR INCOMPLETO	14232	7.33%
6	POSGRADUACAO	2898	1.49%
7	SEM INSTRUCAO	952	0.49%

APÊNDICE C – Listagem de valores possíveis para coluna *NUMBER_OF_HOUSEMATES* da base Cliente

	NUMBER_OF_HOUSEMATES	count	percentage
0	-3	708803	78.75%
1	3 A 5	117221	13.02%
2	1 A 2	45212	5.02%
3	3 a 5	10286	1.14%
4	1 a 2	9632	1.07%
5	+5	8344	0.93%
6	ACIMA de 5	450	0.05%
7	Acima de 5	116	0.01%
8	3+a+5	8	0.00%
9	CASADO	4	0.00%
10	MASCULINO	4	0.00%
11	Null	2	0.00%
12	2	2	0.00%
13	ATE 09	2	0.00%
14	FEMININO	2	0.00%
15	03	2	0.00%
16	4	2	0.00%
17	2 GRAU INC	2	0.00%

APÊNDICE D – Coluna *NUMBER_OF_HOUSEMATES* com valores corrigidos

Valores corrigidos, onde *NaN* indica valor nulo.

	NUMBER_OF_HOUSEMATES	count	percentage
0	Nan	339389	63.96%
1	3 a 5	127515	24.03%
2	1 a 2	54844	10.33%
3	Acima de 5	8910	1.68%
4	2	2	0.00%
5	4	2	0.00%
6	3	2	0.00%
7	Ate 09	2	0.00%

APÊNDICE E – Listagem da variabilidade de registros para um atributo com valor específico de OS

No número de protocolo 173274782, por exemplo, há 92 registros com esse número. A ideia do procedimento é informar, para cada coluna, se seus valores se modificam ou não entre os 92 registros. *UNIQUE* indica que não mudou; *DIFF* demonstra que sim e, por isso, é responsável por haver mais de um registro para um mesmo número de protocolo. Caso a quantidade de valores é menor que 10, cita cada um dos valores; se não, apenas a quantidade de valores distintos e sua porcentagem em relação ao número total de registros duplicados.

Como é possível observar para esse caso, as combinações de *CRIADO _COMENTARIO* e *COMENTARIO_TECNICO_CAMPO* foram os atributos principais para haver registros duplicados. Isso significa que, para a mesma OS, foram registrados contendo comentários de técnico de campo diferentes; criando assim mais registros mas com pouca informação adicional.

```
QTD Rows 92

SRK_ORD_SRV                IS UNIQUE   [nan]
CHAVE_CLIENTE              IS UNIQUE   [11266304]
CHAVE_PRODUTO              IS UNIQUE   [243131]
CHAVE_CONTRATO             IS UNIQUE   [5524633]
SRK_DAT_ABR_ORD_SRV_TEC    IS UNIQUE   [20180205]
SRK_DAT_FCH_ORD_SRV_TEC    IS UNIQUE   [20180206]
SRK_DAT_EXE_ORD_SRV_TEC    IS UNIQUE   [20180206]
SRK_DAT_VNC_ORD_SRV_TEC    IS UNIQUE   [20180206]
DAT_ABR_ORD_SRV            IS UNIQUE   [nan]
DAT_FCH_ORD_SRV            IS UNIQUE   [nan]
DAT_EXE_ORD_SRV            IS UNIQUE   [nan]
DAT_VNC_ORD_SRV            IS UNIQUE   [nan]
DAT_ABR_ORD_SRV_TEC        IS UNIQUE   ['05/02/2018 13:20:49']
DAT_FCH_ORD_SRV_TEC        IS UNIQUE   ['06/02/2018 10:26:15']
DAT_EXE_ORD_SRV_TEC        IS UNIQUE   ['06/02/2018 10:25:00']
DAT_VNC_ORD_SRV_TEC        IS UNIQUE   ['06/02/2018 13:18:00']
DESC_SEGMENTO              IS UNIQUE   ['EMPRESARIAL']
COD_FECHMENTO              IS UNIQUE   ['BV000']
COD_ABERTURA               IS UNIQUE   ['11J7C']
NUM_PROCOLO                IS UNIQUE   [173274782]
DESC_TIPO_OS               IS UNIQUE   ['Manutenção Corretiva de Produt']
DESC_TIPO_FECHAMENTO       IS UNIQUE   ['CAUSA ENCERRADO SEM REPARO']
DESCRICAO_PRODUTO          IS UNIQUE   ['BANDA LARGA 10 MBPS']
DESC_FAMILIA_PRODUTO       IS UNIQUE   ['BANDA LARGA 10 MB']
STATUS_OS                  IS UNIQUE   ['FECHADO']
DATA_ULTIMO_AGENDAMENTO    IS UNIQUE   [nan]
DATA_REFERENCIA_OS         IS DIFF     ['05/02/2018 00:00:00'
                                        '06/02/2018 00:00:00']
FLAG_OS_TRATADA_TEC_CAMPO  IS UNIQUE   ['SIM']
NUM_PROTOCOLO_UNICO        IS UNIQUE   [nan]
CRIADOR_COMENTARIO         IS DIFF     ['luizgfs' 'gabrielsouza' 'RONALDOAM_SOM'
                                        'amandas' 'SOM' 'SysMobile' 'SYSMobile']
CRIADO_COMENTARIO          IS DIFF IN  43    == 46.74% OF TOTAL
COMENTARIO_TECNICO_CAMPO   IS DIFF IN  34    == 36.96% OF TOTAL
DESC_AREA_CONHEC_TEC_CAMPO IS UNIQUE   [nan]
DESC_SEGMENTO_TEC_CAMPO    IS UNIQUE   [nan]
ATTENDANT_KEY              IS UNIQUE   [1631256]
DATA_INTERACAO             IS UNIQUE   ['05/02/2018 12:03:56']
DATA_ATENDIMENTO           IS UNIQUE   [nan]
DATA_CONCLUSAO_ATENDIMENTO IS UNIQUE   [nan]
```

APÊNDICE F – Lista de atributos criados para a estratégia de longo prazo

	Coluna	Descrição
1	SR_month_3	Número de OSs abertas no antepenúltimo mês
2	SR_month_2	Número de OSs abertas no penúltimo mês
3	SR_month_1	Número de OSs abertas no último mês
4	avg_signal_up_week_4	Média do sinal de Upload na pré-antepenúltima semana
5	avg_signal_up_week_3	Média do sinal de Upload na antepenúltima semana
6	avg_signal_up_week_2	Média do sinal de Upload na penúltima semana
7	avg_signal_up_week_1	Média do sinal de Upload na última semana
8	avg_signal_down_week_4	Média do sinal de Download na pré-antepenúltima semana
9	avg_signal_down_week_3	Média do sinal de Download na antepenúltima semana
10	avg_signal_down_week_2	Média do sinal de Download na penúltima semana
11	avg_signal_down_week_1	Média do sina de Download na última semana
12	avg_attenuation_up_week_4	Média da atenuação de Upload na pré-antepenúltima semana
13	avg_attenuation_up_week_3	Média da atenuação de Upload na antepenúltima semana
14	avg_attenuation_up_week_2	Média da atenuação de Upload na penúltima semana
15	avg_attenuation_up_week_1	Média da atenuação de Upload na última semana

16	avg_attenuation_down_week_4	Média da atenuação de Download na pré-antepenúltima semana
17	avg_attenuation_down_week_3	Média da atenuação de Download na antepenúltima semana
18	avg_attenuation_down_week_2	Média da atenuação de Download na penúltima semana
19	avg_attenuation_down_week_1	Média da atenuação de Download na última semana
20	std_attainable_rate_up_month_3	Desvio Padrão da taxa de atenuação de Upload no no antepenúltimo mês
21	std_attainable_rate_up_month_2	Desvio Padrão da taxa de atenuação de Upload no penúltimo mês
22	std_attainable_rate_up_month_1	Desvio Padrão da taxa de atenuação de Upload no último mês
23	std_attainable_rate_down_month_3	Desvio Padrão da taxa de atenuação de Download no antepenúltimo mês
24	std_attainable_rate_down_month_2	Desvio Padrão da taxa de atenuação de Download no penúltimo mês
25	std_attainable_rate_down_month_1	Desvio Padrão da taxa de atenuação de Download no último mês
26	std_current_rate_up_month_3	Desvio Padrão da taxa de Upload no antepenúltimo mês
27	std_current_rate_up_month_2	Desvio Padrão da taxa de Upload no penúltimo mês
28	std_current_rate_up_month_1	Desvio Padrão da taxa de Upload no último mês
29	std_current_rate_down_month_3	Desvio Padrão da taxa de Download no antepenúltimo mês
30	std_current_rate_down_month_2	Desvio Padrão da taxa de Download no penúltimo mês
31	std_current_rate_down_month_1	Desvio Padrão da taxa de Download no último mês

APÊNDICE G – Lista de atributos criados para a estratégia de curto prazo

	Coluna	Descrição
1	SR_day_15	Número de OS do oitavo ao décimo quinto anteriores
2	SR_day_7	Número de OSs do quarto ao sétimo dia anteriores
3	SR_day_3	Número de OSs do último ao antepenúltimo dia
4	avg_signal_up_day_7_8	Média do sinal de Upload no sétimo ao oitavo dia anteriores
5	avg_signal_up_day_5_6	Média do sinal de Upload no quinto ao sexto dia anteriores
6	avg_signal_up_day_3_4	Média do sinal de Upload no terceiro ao quarto dia anteriores
7	avg_signal_up_day_1_2	Média do sinal de Upload no penúltimo e último dia
8	avg_signal_down_day_7_8	Média do sinal de Download no sétimo ao oitavo dia anteriores
9	avg_signal_down_day_5_6	Média do sinal de Download no quinto ao sexto dia anteriores
10	avg_signal_down_day_3_4	Média do sinal de Download no terceiro ao quarto dia anteriores
11	avg_signal_down_day_1_2	Média do sinal de Download no penúltimo e último dia
12	avg_attenuation_up_day_7_8	Média da taxa de atenuação de Upload no sétimo ao oitavo dia anteriores
13	avg_attenuation_up_day_5_6	Média da taxa de atenuação de Upload no quinto ao sexto dia anteriores
14	avg_attenuation_up_day_3_4	Média da taxa de atenuação de Upload no terceiro ao quarto dia anteriores
15	avg_attenuation_up_day_1_2	Média da taxa de atenuação de Upload no penúltimo e último dia
16	avg_attenuation_down_day_7_8	Média da taxa de atenuação de Download no sétimo ao oitavo dia anteriores
17	avg_attenuation_down_day_5_6	Média da taxa de atenuação de Download no quinto ao sexto dia anteriores
18	avg_attenuation_down_day_3_4	Média da taxa de atenuação de Download no terceiro ao quarto dia anteriores
19	avg_attenuation_down_day_1_2	Média da taxa de atenuação de Download no penúltimo e último dia

20	std_attainable_rate_up_day_11_15	Desvio Padrão da taxa de atenuação de Upload do décimo primeiro ao décimo quinto dias anteriores
21	std_attainable_rate_up_day_6_10	Desvio Padrão da taxa de atenuação de Upload do sexto ao décimo dia anteriores
22	std_attainable_rate_up_1_5	Desvio Padrão da taxa de atenuação de Upload nos últimos cinco dias
23	std_attainable_rate_down_day_11_15	Desvio Padrão da taxa de atenuação de Download do décimo primeiro ao décimo quinto dias anteriores
24	std_attainable_rate_down_6_10	Desvio Padrão da taxa de atenuação de Download do sexto ao décimo dia anterior
25	std_attainable_rate_down_1_5	Desvio Padrão da taxa de atenuação de Download nos último cinco dias
26	std_current_rate_up_day_11_15	Desvio Padrão da taxa corrente de Upload do décimo primeiro ao décimo quinto dias anteriores
27	std_current_rate_up_6_10	Desvio Padrão da taxa corrente de Upload dos dias do sexto ao décimo dia anteriores
28	std_current_rate_up_1_5	Desvio Padrão da taxa corrente de Upload nos últimos cinco dias
29	std_current_rate_down_day_11_15	Desvio Padrão da taxa corrente de Download do décimo primeiro ao décimo quinto dias anteriores
30	std_current_rate_down_6_10	Desvio Padrão da taxa corrente de Download do sexto ao décimo dias anteriores
31	std_current_rate_down_1_5	Desvio Padrão da taxa corrente de Download nos últimos cinco dias