INSTITUTO DOCTUM DE EDUCAÇÃO E TECNOLOGIA

FACULDADES INTEGRADAS DE CARATINGA

CLÁUDIO LOPES MONTEIRO

MODELO PARA CLASSIFICAÇÃO DE DADOS SUJOS EM UMA BASE DE DADOS ESTRUTURADA

CARATINGA
2017

INSTITUTO DOCTUM DE EDUCAÇÃO E TECNOLOGIA

FACULDADES INTEGRADAS DE CARATINGA

CLÁUDIO LOPES MONTEIRO

MODELO PARA CLASSIFICAÇÃO DE DADOS SUJOS EM UMA BASE DE DADOS ESTRUTURADA

Monografia apresentada à banca examinadora da Faculdade de Ciência da Computação das Faculdades Integradas de Caratinga, como requisito parcial para obtenção do título de bacharel em Ciência da Computação, sob orientação do professor e Msc. Glauber Luiz Costa.

CARATINGA

2017

rede de ensino DOCTUM

INSTITUTO DOCTUM DE EDUCAÇÃO E TECNOLOGIA

FACULDADES INTEGRADAS DE CARATINGA

FOLHA DE APROVAÇÃO

O Trabalho de Conclusão de Curso intitulado: MODELO PARA CLASSIFICAÇÃO DE DADOS SUJOS EM UMA BASE DE DADOS ESTRUTURADA, elaborado pelo(a) aluno (a) CLÁUDIO LOPES MONTEIRO foi aprovado por todos os membros da Banca Examinadora e aceita pelo curso de Ciência da Computação das Faculdades Integradas de Caratinga, como requisito parcial da obtenção do título de

BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Caratinga, 12 de dezembro 2017

Prof. Msc. Glauber Luiz Costa
Prof. Orientador

Prof. Msc. Elias de Souza Gonçalves
Prof. Examinador 1

Maicon Ribeiro

Prof. Esp. Maicon Vinicius Ribeiro
Prof. Examinador 2

AGRADECIMENTOS

Primeiramente agradeço a Deus por ter me dado forças e iluminado meu caminho para que pudesse concluir mais esta etapa na minha vida. A minha esposa Karla, meus filhos Matheus e Anne, minha mãe Joaquina, minha irmã Luciana e minha sogra Conceição pela compreensão de minha ausência e pelo apoio incondicional durante toda esta caminhada.

Aos meus professores pelos valorosos ensinamentos nesses anos e aos colegas de turma que estiveram comigo em todos os momentos.

Agradeço também ao meu orientador Glauber por ter me guiado para que esse trabalho fosse realizado. Muito obrigado!

RESUMO

Devido ao acúmulo e crescimento constante dos dados que são gerados dentro das empresas, é certo que estes precisam estar corretos, o que nem sempre ocorre devido a vários fatores, gerando informações incorretas, inconsistentes ou até mesmo falsas. Estas anomalias podem ser definidas como dados sujos. Esse trabalho tem como objetivo buscar uma forma de identificá-los e classificá-los para que possam ser higienizados posteriormente.

Para atingir este objetivo foi realizado pesquisa na área de mineração de dados e gestão da qualidade de dados, onde foi possível criar uma forma de encontrar e separar os sujos dos corretos. Foi realizado um levantamento dos requisitos que são utilizados para sua inserção. Assim, através destes, criou-se as regras para a análise e classificação.

Com isto obteve-se os resultados esperados, sendo possível classificar todos os dados, onde gerou-se uma informação da sua qualidade. Os resultados foram expostos através de planilhas e gráficos, oferecendo aos gestores informações para o acerto ou a implementação de um manual de inserção de dados e assim podendo corrigir os que estiverem incorretos.

Palavras-chave: dados sujos, mineração, classificação, Dama Dmbok, qualidade.

ABSTRACT

Due to the accumulation and constant growth of data that is generated within entreprises, it is certain that these need to be correct, which does not Always occur due to several factors, generating incorrect, incosistent or even false information. These anomalies can be defined as dirty data. This work aims to seek a way to identify them and classify them so that they can be sanitized afterwards.

To achieve this goal was conducted research in the data mining area and data quality management, where it was possibible to create a way to find and separate the dirty from the correct ones. A survey of the requirements that are used for insertion is carried out. Thus, the rules for analysis and classification were created through these.

With this obtained the expected results, and it was possible to classify all the data, where information of its quality was generated. The results were exposed through spreadsheets and graphs, offering managers information for the hit or implementation of a data insertion manual and thus able to correct thouse that are incorrect.

Keyword: Dirty data, mining, classification, Dama Dmbok, quality.

LISTA DE FIGURAS

Figura 1: Funções Dama Dmbok (Rêgo, 2014).....**Erro! Indicador não definido.**

LISTA DE TABELAS

Tabela 1 - Estrutura do Arquivo 16
Tabela 2 - Estrutura do resultado da ferramenta..... 18
Tabela 3 - Resultado da Caracterização dos dados ... (continua)..... 18
Tabela 3 - Resultado da Caracterização dos dados ... (continuação)..... 19
Tabela 4 - Classificação dos dados sujos 20
Tabela 5 - Resultado da Caracterização dos dados – Redução ... (continua) 23
Tabela 5 - Resultado da Caracterização dos dados – Redução ... (continuação) 24
Tabela 6 - Resultado Classificação - Bairro_AL..... 26
Tabela 7 - Resultado Classificação - Bairro_RF 26
Tabela 8 - Resultado Classificação - Cep_AL..... 26
Tabela 9 - Resultado Classificação - Cep_RF 27
Tabela 10 - Resultado Classificação - Cidade_AL 27
Tabela 11 - Resultado Classificação - Cidade_RF..... 27
Tabela 12 - Resultado Classificação - Cpf_Mae 28
Tabela 13 - Resultado Classificação - Cpf_Pai 28
Tabela 14 - Resultado Classificação - DtNasc_AL..... 28
Tabela 15 - Resultado Classificação - Endcomp_AL 29
Tabela 16 - Resultado Classificação - Endcomp_RF..... 29
Tabela 17 - Resultado Classificação - Mae_AL 29
Tabela 18 - Resultado Classificação - Nome_AL..... 30
Tabela 19 - Resultado Classificação - Nome_RF 30
Tabela 20 - Resultado Classificação - Num_AL..... 30
Tabela 21 - Resultado Classificação - Num_RF 31
Tabela 22 - Resultado Classificação - Pai_AL 31
Tabela 23 - Resultado Classificação - Paìs_AL 31
Tabela 24 - Resultado Classificação - Rua_AL..... 32
Tabela 25 - Resultado Classificação - Rua_RF 32
Tabela 26 - Resultado Classificação - UF_AL..... 32
Tabela 27 - Resultado Classificação - UF_RF 33
Tabela 28 - Resultado Classificação – Geral 33

LISTA DE GRÁFICOS

Gráfico 1 - Resultado Caracterização 24
Gráfico 2 - Resultado Classificação – Geral 34

LISTA DE SIGLAS

DAMA	Data Management;
DMBOK	Data Management Body of Knowledge;
SAS	Statistical Analsys System;
KDD	Knowledge Discovery in Databases (Descoberta de Conhecimentos em Base de Dados);

SUMÁRIO

1. INTRODUÇÃO 3
2. REFERENCIAL TEÓRICO 4
2.1. Mineração de Dados 4
2.1.1. Pré-processamento 5
2.1.2. Mineração de Dados 6
2.1.3. Pós-processamento 7
2.2. Gestão de Dados 7
2.2.1. Gestão da Qualidade de Dados 9
3. METODOLOGIA 12
3.1. Levantamento de Requisitos 12
3.2. Caracterização dos Dados 13
3.3. Análise dos Dados 13
4. ESTUDO DE CASO 15
4.1. Levantamento de Requisitos 15
4.2. Caracterização dos Dados 17
4.3. Análise dos Dados 19
5. RESULTADOS 23
5.1. Caracterização dos Dados 23
5.2. Análise dos Dados 25
5.3. Resultado Geral 33
6. CONCLUSÃO 35
7. TRABALHOS FUTUROS 37
REFERÊNCIAS 38
APÊNDICE A 40
Script A – Classificação Geral – Aspecto Humanos 40
APÊNDICE B 42
Script B – Classificação Individual – Erro Aspectos Humanos - Abreviações 42
APÊNDICE C 44
Script C – Classificação Individual – Erro Aspectos Humanos - Número 44
APÊNDICE D 46

Script D – Classificação Individual – Erro Aspectos Humanos - Texto 46
ANEXO 48
Autorização para utilização dos dados.....48

1. INTRODUÇÃO

Com o crescimento exponencial na geração de dados nos dias atuais, se faz necessário que estes sejam transformados em informações coesas, para isto é necessário que as empresas tenham uma visão, o dado é o seu bem de maior valor. Segundo Rêgo (2013, p.21) “Os dados são o ativo mais importante das empresas” e segundo Ribeiro (2014, p.17) “as organizações dependem cada vez mais da qualidade de seus dados para aumentar a sua eficiência”. Diante disto, estes precisam ser gerenciados durante o seu ciclo de vida.

Dito isto, torna-se necessário que a base de dados seja confiável e estruturada. Os gestores de tecnologia e informação enfrentam o desafio de mantê-las sempre dentro do padrão estabelecido pela empresa. Com isto eles podem gerar informações permitindo que os gestores tomem decisões de forma assertiva. Segundo Rêgo (2013, p.13) “...dados são fundamentais para as organizações.”

Baseado nestas condições a proposta deste trabalho é encontrar ou criar um método de descoberta para os dados incorretos, classificando-os de acordo com os padrões estabelecidos pela empresa.

Para isto foi utilizado a mineração de dados juntamente com a gestão de dados para que se pudesse encontrar métricas de classificação e a melhor forma de executá-la.

Dito isto, o intuito é manter a base de dados confiável para utilizá-la criando informações consistentes que possam trazer benefícios econômicos para a empresa.

Este trabalho está estruturado com quatro capítulos principais. O capítulo 2 descreve o referencial teórico, apresentando as principais funções da mineração de dados e gestão de dados que serão utilizados.

Já os capítulos 3, 4 e 5 representam respectivamente a metodologia a ser aplicada, o estudo de caso e apresentação dos resultados obtidos.

2. REFERENCIAL TEÓRICO

Neste capítulo será apresentado o referencial teórico deste trabalho, o qual discrimina as fases, técnicas e benefícios na análise de dados estruturados.

Ele está separado em duas seções, onde na primeira seção será apresentado a mineração de dados. Esta busca analisar um grande volume de dados transformando-os em conhecimento. Na seção seguinte será apresentado a gestão de dados, que busca gerenciar as informações das empresas, mantendo-as padronizadas e controlando a sua utilização.

2.1. Mineração de Dados

Com o grande volume de dados gerados pelas empresas, existe a necessidade constante de transformá-los em informações coesas, que permitam aos gestores sua utilização para assim gerir os negócios da melhor forma possível. Por muitas vezes estes são captados através de sistemas de gestão administrativa, smartphones, sistemas de compra online e mecanismos de busca na internet que vão armazenando-os em um banco de dados estruturados.

> A captação de informações pode ser feita de diversas maneiras e utilizada de igual forma em outras tantas, tentando fazer com que estas informações sejam traduzidas de modo benéfico, coerente e produtivo. As bases de dados computacionais podem trazer informações desconexas, que em um processo de tomada de decisões podem fomentar induções incorretas. (SCHIMITT, 2009, p. 9).

Uma das formas de transformar dados em informação é utilizando a mineração de dados que trabalha com estes de forma estruturada, porém é necessário que se faça algumas perguntas antes de utilizá-la, em Goldschimidt e Passos (2005, p.1) “...O que fazer com todos os dados armazenados? e como utilizar o patrimônio digital em benefício das instituições?...”.

Dentro estes aspectos, a mineração de dados, pode auxiliar no seu levantamento. Segundo Braga (2005, p.11) “A mineração de dados provê um

método automático para descobrir padrões de dados...".

> Mineração de dados é a parte de um processo maior denominado Busca de Conhecimento em Bancos de Dados (Knowledge Discovery in Database – KDD), o qual possui uma metodologia própria para preparação e exploração dos dados, interpretação de seus resultados e assimilação de conhecimento. (VIEIRA, 2014, p. 20).

Pode-se afirmar então que a mineração de dados é a busca pelo conhecimento, podendo ser feito de forma automática ou não. Ela busca informar as regras e padrões dentro de uma grande quantidade de dados.

Para um melhor esclarecimento a mineração de dados busca transformar dados em informação, processando e transformando-os em uma base de conhecimento.

Este conhecimento pode ser denominado como descoberta de conhecimento. Este é divido em três etapas: pré-processamento, que prepara os dados; mineração de dados que utiliza algoritmos para buscar o conhecimento; pós-processamento que trata o resultado da etapa anterior respondendo às perguntas, segundo Goldschimidt e Passos (2005).

2.1.1. Pré-processamento

Nesta primeira fase, utiliza-se técnicas que são aplicadas para a captação e a organização dos dados, onde os mesmos são tratados para a etapa de mineração de dados.

Segundo Goldschimidt e Passos (2005), estas tarefas são seleção de dados, limpeza, codificação e enriquecimento dos dados.

A seleção de dados realiza a redução, onde busca a compreensão da característica destes.

A limpeza realiza o tratamento dos dados, onde busca assegurar a sua qualidade, em Goldschimidt e Passos (2005, p.12) os dados devem ter as seguintes características: "completude, veracidade e integridade". Com isto é necessário que eles sejam tratados, não devem possuir campos nulos, informações erradas, como erros de digitação, assim sendo necessário seu tratamento. Segundo Camilo e Silva (2009, p.6) "A etapa de limpeza dos dados visa eliminar estes problemas de

mondo que eles não influam nos resultados dos algoritmos usados.”

A codificação dos dados tem como função codificá-los para os algoritmos da fase de mineração, possam utilizá-los sem perderem a referência sobre eles.

Já o enriquecimento busca agregar mais informações a estes, utilizando dados externos, como planilhas, banco de dados ou documentos.

2.1.2. Mineração de Dados

Trabalha com o resultado da fase anterior que é a descoberta de conhecimento, para isto é necessário a escolha de um método de busca. Entre os diversos pode-se citar o de redes neurais que em Schimitt (2009, p.17) “... são sistemas de classificação modelados segundo os princípios do sistema nervoso humano.”, regras de associação que buscam encontrar relacionamento entre os atributos que possam existir em um banco de dados e o de árvores de decisão que segundo Schimitt (2009, p.15) “São constituídas por estruturas em árvore que representam um conjunto de decisões”.

A escolha do método infere na utilização das tarefas desta fase, em Goldschimidt e Passos (2005, p.12) “A escolha da técnica depende, muitas vezes, do tipo de tarefa de KDD a ser realizada”, estas tarefas dependendo do método escolhido, podem originar novas tarefas. Entre elas destacam-se:

- Descoberta de Associação – Realiza a análise da base de dados determinando os relacionamentos entre os seus atributos.
- Classificação – Realiza o mapeamento dos registros, construindo um agrupamento dos dados, o que facilita a análise dos mesmos. O usuário define como deve ocorrer a classificação. Esta fase pode ser denominada também como caracterização.
- Descoberta de Sequências - Trabalha com os dados determinando a sua análise através do período das transações dos dados, separa os mesmos pelo tempo, modelando-os através da sua evolução.
- Sumarização – Realiza um resumo dos dados através da identificação das características dos mesmos.
- Clusterização – Transforma os dados em diversas classes, muito

diferente da classificação, pois busca de forma automática identificar estas classes.

2.1.3. Pós-processamento

Nesta última fase, todo o conhecimento adquirido é tratado e organizado. Deste são respondidas as perguntas através de diagramas, tabelas, relatórios demonstrativos e gráficos Goldschimidt e Passos (2005). Após a análise dos resultados, pode-se ainda definir uma nova busca de conhecimento, levando a uma nova investigação.

2.2. Gestão de Dados

A gestão de dados é utilizada para planejar e implementar procedimentos que são necessários para o bom gerenciamento destes e como guia para o gerenciamento, segundo Rêgo (2013, p.94) "... a função da Governança de Dados é responsável por gerir os princípios de organização e controle de dados e informação." E na opinião de Souza Jr. (2014):

> As empresas que quiserem criar um bom plano de governança de dados precisarão definir objetivos organizacionais e criar processos institucionalizados, que deverão ser implementados integrando as áreas de TI e negócios. (SOUZA JR., 2014, p. 23).

A gestão de dados utiliza o DAMA DMBOK como guia de boas práticas para a sua execução. Este estabelece dez funções primárias, onde estas não precisam ser executadas em uma ordem preestabelecida e nem implementadas na sua totalidade (Rêgo, 2014).

Sendo assim, são caracterizadas da forma exposta na figura 1:

Figura 1 - Funções Dama Dmbok

Fonte: Rêgo (2014)

- Governança de dados – Seu objetivo principal é controlar todo o planejamento das diversas disciplinas que compõe a gestão de dados. Ela planeja e monitora todas as funções que visam controlar os ativos de dados.
- Gestão da Arquitetura de dados – Entende todas as necessidades da informação na empresam, define a arquitetura de tecnologia e integração dos dados.
- Desenvolvimento dos dados – Sua função é a implementação, manutenção e a criação de projetos das soluções que visam satisfazer a real necessidade dos dados dentro da empresa.
- Gestão e Operações e Database – Ela faz o planejamento e controle dos dados estruturados no desenvolvimento de sistemas, isto dentro do ciclo de vida dos dados.
- Gestão da Segurança dos dados – Esta função defini a política e padrões de segurança de dados de acordo com o grau de confidencialidade das informações.
- Gestão de Dados Mestres e Referência – Esta função é responsável pelos dados mais precisos que são os que categorizam e classificam as informações, onde ela planeja definindo a sua arquitetura

implementando e controlando a sua utilização.

- Gestão de Data Warehousing e Business Intelligence – Sua principal função é o planejamento, controle e implementação das informações que dão suporte à decisão estratégica dentro da empresa.
- Gestão da Documentação e Conteúdo – Esta função cria o planejamento, implementação e controle sobre o armazenamento dos dados, gerencia o seu conteúdo e planeja a aquisição de insumos e softwares para o backup e recuperação dos dados.
- Gestão de Metadados – Ela planeja, implementa e controla as atividades de acesso aos metadados dentro das empresas.
- Gestão da Qualidade de dados – Esta função planeja e implementa atividades que medem, avaliam e otimizam a qualidade dos dados.

Portanto as funções deste guia devem ser analisadas e adaptadas à realidade de cada empresa, segundo Rêgo (2014).

2.2.1. Gestão da Qualidade de Dados

A Gestão da Qualidade de Dados trabalha com a análise e valorização dos dados, ela possui ferramentas que permite que os mesmos sejam padronizados. Durante o processo de levantamento e inserção dos dados, se ela for bem aplicada, garantirá que estes gerem informações de qualidade.

Porém a qualidade de dados não é só um problema de domínio da tecnologia da informação, isto segundo Barbieri e Farinelli (2016), é também um problema da inexistência de regras de negócio ou a má formulação das mesmas. Com isto, ela dedica-se a aplicação de técnicas que tem o propósito de medir, avaliar, melhorar e garantir a qualidade dos dados, assim conforme citado por Rêgo (2014), promovendo a importância da sua utilização dentro da empresa, seja através de treinamentos ou a conscientização dos funcionários.

E segundo Barbieri e Farinelli (2016, p.39), “A qualidade de dados deve ser um dos elementos fundamentais do arco da Governança de Dados...”, trabalhando de forma integrada, onde concilia pessoas, tecnologia e processos, permitindo que os dados sejam valorados.

A qualidade de dados é separada em três grupos, sendo:

- Qualidade dos metadados - atua diretamente nos processos de modelagem de dados, isto quando os gestores atuam nestes processos, avaliando a qualidade destes.
- Qualidade de processos - realiza ações de melhoria nas métricas e indicadores dos processos utilizados.
- Qualidade do conteúdo - utilizada diretamente no ambiente de trabalho, que segundo Rêgo (2014, p.194) “busca a tomada de ações de melhoria contínua em cima das métricas e dos indicadores dos processos controlados”. Este, trabalha diretamente com a identificação e classificação dos dados sujos.

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

Rêgo (2014) apresenta as principais causas da qualidade baixa em quatro grandes grupos, sendo:

Grupo 1 – Aspectos técnicos: São os erros cometidos em todas as etapas do processo de armazenagem.

Grupo 2 – Aspectos humanos: Trata-se de erros não intencionais na inserção de dados, podendo ser erros de digitação, treinamento ineficaz e inexistência ou processo ineficaz para a coleta de dados.

Grupo 3 – Fatores Organizacionais: Falta de conscientização sobre a base de dados, onde não se têm a importância da qualidade dos dados. A base de dados é utilizada sem controle pelos diversos departamentos.

Grupo 4 – Negligência Corporativa: Falta de conhecimento do valor em melhorar a qualidade de dados e a falta de disseminação do conhecimento, treinamento e participação dos gestores.

A qualidade de dados trabalha com alguns requisitos que estão diretamente ligados as funções e necessidades da empresa, segundo Barbieri e Farinelli (2016). Estes requisitos possibilitam avaliar o padrão, gerando indicadores sobre a qualidade dos dados, segundo Rêgo (2014). Dentre estes tempos a acurácia,

completude, consistência, atualidade, precisão, privacidade, razoabilidade, integridade referencial, unicidade e validade.

Segundo Rêgo (2014, p.200) “A seleção dos requisitos de qualidade são fundamentais para definir as questões que serão aplicadas nos processos...”, eles ajudam a definir as questões que precisam ser levantadas durante o processo de implantação da qualidade de dados, permitindo que seja estabelecido valores para a medição dos dados de baixa qualidade.

Para a medição utiliza-se o processo de perfilar os dados, onde utilizando técnicas de análise é possível o desenvolvimento de relatórios que retornem informações sobre os dados, explanando o conhecimento sobre seu conteúdo, estrutura e qualidade (Rêgo, 2014).

Para isto é necessário que se faça a análise deste resultado confrontando com as regras de negócio da empresa, onde deve-se levar em conta o volume de dados incorretos que possam ser encontrados durante a perfilagem dos dados, segundo Rêgo (2014).

Por fim é necessário que durante todo este processo seja revisto as regras de negócio que a empresa possui, levando sempre em conta que as mesmas precisam ser revistas e atualizadas. Com isto pode-se executar o processo de limpeza e correção dos dados, segundo Rêgo (2014, p.201) “... é um processo de atualização e correção dos dados, garantindo a consistência das informações”. Este processo cabe a equipe de Qualidade de Dados, que precisa aprovar o tratamento das inconsistências, garantindo que todas elas sejam sanadas.

3. METODOLOGIA

O presente trabalho teve como objetivo a criação de modelo para a classificação de dados sujos em uma base cadastral, utilizando técnicas da mineração de dados e da gestão de qualidade de dados para a sua análise e classificação.

Realizado o levantamento de bibliografias específicas com todos os conceitos que agregaram conhecimento para o levantamento de requisitos e dados, foi percebido que a mineração de dados não possui método específico para classificar os dados sujos, uma vez que na fase pré-processamento da descoberta de conhecimentos, a tarefa de limpeza de dados pede a retirada e solução dos mesmos. Então, foi utilizado como guia, a qualidade de dados da gestão de dados para gerar a classificação dos dados sujos. Ambos foram utilizados para a montagem de um modelo básico a ser implementado em qualquer base de dados estruturado, a fim demonstrar os problemas com os cadastros que porventura estejam incorretos.

3.1. Levantamento de Requisitos

Realizar entrevista no departamento responsável pela gerência e manutenção da base de dados, onde deve ser colhido as seguintes informações:

- Tipos de dados a serem analisados.
- Levantamento dos campos com preenchimento obrigatório.
- Levantamento dos campos que possuem dados oriundos de tabelas.
- Levantamento das regras de inserção dos dados.

Estas informações servirão para criar o dicionário de palavras, juntamente com a tabela de classificação, que serão utilizados nas etapas de caracterização e análise.

3.2. Caracterização dos Dados

Nesta etapa os dados serão agrupados, caracterizados e analisados. Para isto serão utilizados os softwares SAS University Edition e o Microsoft Sql Server 2016 Express Edition.

Para o agrupar e caracterizar os dados, será utilizado o SAS University Edition, fornecida pelo Instituto SAS. Esta é voltada para o público acadêmico, sendo muito utilizado nas áreas de ciências da computação, negócios, ciências da saúde, engenharia entre outras (SAS, 2017). A ferramenta a ser utilizada para a descoberta de conhecimento, Characterize Data. A sua função é criar um relatório onde tem-se o resumo e a frequência em que cada registro aparece nas tabelas analisadas. Agrupa os dados de acordo com as suas características.

Após realizado o processo de agrupamento, os dados serão exportados para dentro de um banco de dados onde serão utilizados na etapa de análise.

3.3. Análise dos Dados

Nesta etapa, serão identificados os cadastros que estão com falta de informação, erros ortográficos, dados inconsistentes e corretos, para isto será criada metodologia de acordo com o levantamento de requisitos.

Para a análise é necessário criar um dicionário com palavras baseadas nas regras de inserção dos dados. Conforme essas, pode-se ter mais de um dicionário.

A tabela de classificação é construída de acordo com as regras de inserção dos dados. Como modelo utiliza-se as principais causas de dados sujos, conforme descrito na seção 2.2.1.

A análise inicial deve ser feita de forma automática, para isto foi criado um *script* que percorrerá todo o banco de dados, buscando e classificando os dados de acordo com os dicionários criados. Para realizar esta tarefa deve ser utilizado o Script A (Apêndice A). Para a análise visual foram criados os scripts B (Apêndice B), C (Apêndice C) e D (Apêndice D). Em ambas as análises se utiliza o software

Microsoft Sql Server 2016 Express Edition para executar os scripts.

No final desta etapa, tendo os dados identificados e classificados, utiliza-se o resultado para produzir relatórios com informações quantitativas. Estes devem conter a descrição do item encontrado, isto de acordo com a tabela de classificação, como também o total de registros encontrados. Para esta tarefa é utilizado o processo de perfilar dados.

4. ESTUDO DE CASO

A Fundação Educacional de Caratinga- FUNEC foi criada em 7 de fevereiro de 1963 como uma instituição de direito privado, de caráter comunitário e sem fins lucrativos. Em meados de 1968 foram criados os primeiros cursos de graduação: Matemática, Letras, Pedagogia e História. Ela ainda é mantenedora da Escola Professora Jairo Grossi, Centro Universitário de Caratinga – UNEC e do Centro de Assistência à Saúde – CASU.

Para este estudo de caso, foi utilizado uma base de dados, onde tem-se registros dos alunos matriculados no ensino básico e superior, entre os anos de 2004 a 2017.

A instituição utiliza a solução educacional do sistema TOTVS para atender as necessidades administrativas dos departamentos ligados diretamente a área educacional. Como o sistema se trata de uma ERP – Enterpresie Resource Planning, o cadastro de alunos pode ser acessado de todos os departamentos que estão envolvidos no processo. Porém somente os departamentos da secretaria do ensino superior e ensino médio é que têm a permissão de realizar inclusões, alterações e quando necessário a exclusão dos dados.

Os dados fornecidos geram informações para os departamentos financeiro e tesouraria que são responsáveis pela geração de boletos, cadastros de bolsas, relatórios de inadimplência e de recebimentos e outros serviços pertinentes as suas funções. Com estes dados são geradas informações para o Ministério da Educação como o censo escolar.

Considera-se que estes dados geram informação de extrema importância para a instituição, estes precisam estar coesos e livres de qualquer problema com o cadastro. Com isto o intuito deste estudo de caso é classificar os dados sujos norteando a instituição para a sua correção.

4.1. Levantamento de Requisitos

Para o levantamento de requisitos foi realizado entrevista com o analista de

sistemas do departamento de informática da Fundação Educacional de Caratinga, foi informado que não existe documentação formal para a inserção de dados no sistema educacional, mas que existe um treinamento onde são ensinados aos usuários boas práticas para a inserção destes.

Os dados que serão caracterizados, analisados e classificados têm sua origem do módulo educacional do sistema TOTVS que é utilizado na instituição. Para isto o departamento de informática cedeu o uso da seguinte estrutura com 30.273 (trinta mil e duzentos e setenta e três) registros para este estudo de caso conforme exposta na Tabela 1.

Tabela 1 - Estrutura do Arquivo

Campo	Descrição
Bairro_AL	Bairro do aluno
Bairro_RF	Bairro do responsável financeiro
Cep_AL	Cep do aluno
Cep_RF	Cep do responsável financeiro
Cidade_AL	Cidade do Aluno
Cidade_RF	Cidade do responsável financeiro
Cpf_Mae	Cpf da mae do aluno
CPF_Pai	Cpf do pai do aluno
DtNasc_AL	Data de nascimento do aluno
EndComp_AL	Complemento do endereço do aluno
EndComp_RF	Complemento do endereço do responsável financeiro
Mae_AL	Nome da mãe do aluno
Nome_AL	Nome do aluno
Nome_RF	Nome do responsável financeiro
Num_AL	Número do imóvel do aluno
Num_RF	Número do imóvel do responsável financeiro
Pai	Nome do pai do aluno
País_AL	País de origem do aluno
Rua_AL	Descrição do logradouro do aluno
Rua_RF	Descrição do logradouro do responsável financeiro
UF_AL	Estado do aluno
UF_RF	Unidade federativa do responsável financeiro

Fonte: Próprio autor

Com isto foi feito o levantamento dos seguintes requisitos de regras de inserção de dados:

- Nos campos como nome do aluno, nome do pai e mãe, responsável financeiro e endereço não podem ser utilizados sinais de pontuação, acentos, caracteres matemáticos e abreviações.
- Nos campos como cep, número do imóvel, cpf só podem ter caracteres numéricos, mesmo que estes campos permitam a inserção de caracteres alfabéticos.
- Não existem campos configurados no sistema com a sua inserção obrigatória, por ser tratar de cadastro de alunos. Nem sempre o aluno possui todos os documentos no ato da sua matrícula. Com isto estes acertos e inserções de dados é feito a posterior.
- Os campos estado e país são oriundos de tabelas pre-cadastradas.

Para a caracterização dos dados é necessário saber os atributos de cada campo. O analista de sistemas informou que o campo DtNasc_AL é no formato data, todos os outros são configurados com o atributo nvarchar, o qual aceita qualquer tipo de caractere.

4.2. Caracterização dos Dados

Após realizado o levantamento de requisitos, os dados serão preparados para o agrupamento e caracterização. Para realizar esta etapa será utilizado o SAS University Edition. Este processo foi separado nas seguintes etapas:

- Importação dos dados – Utilizando a ferramenta de importação, este transforma os dados no formato que o programa requer.
- Caracterização – Agrupa os registros que serão analisados. A ferramenta a ser utilizada é a Characterize Data, esta irá criar uma tabela para cada campo analisado conforme exposta na tabela 2:

Tabela 2 - Estrutura do resultado da ferramenta.

Campo	Descrição
Variables (nome do campo)	Apresenta os registros tal como exportados
Frequency (frequência)	Quantidade de vezes que o registro foi encontrado
Percent (percentual)	Valor em percentual sobre a quantidade de vezes que o registro foi encontrado
Cumulative Frequency (frequência acumulada)	soma da frequência
Cumulative Percent (percentual acumulado)	soma do percentual

Fonte: Próprio autor

A importação de dados é um processo que não requer o tratamento dos dados que serão analisados. Estes foram entregues no formato xlsx do programa Microsoft Excel 2016, formato compatível com o SAS University Edition. Com isto bastou apenas abri-lo para que fosse apresentado o script de importação, este script faz a leitura e cria uma tabela com todos os registros, estes não são alterados pelo script o que assegura a legitimidade para a análise.

Após a importação, foi feita a caracterização dos dados utilizando a ferramenta Characterize Data, escolhe-se então quais os campos que irão ser analisados por ela. Foram utilizados todos os campos conforme descrito na seção 4.1. Após incluindo todos os campos o script de importação é executado. Com a execução da ferramenta, esta apresentou as novas tabelas com os dados caracterizados, em todas elas obtiveram-se uma considerável redução de registros, o que irá facilitar a análise e classificação dos mesmos na próxima etapa. A tabela 3 descreve como ficou a quantidade de registros para cada campo:

Tabela 3 - Resultado da Caracterização dos dados ... (continua)

Campo	Quantidade	Redução
Bairro_AL	2.023	93,32%
Bairro_RF	1.977	93,47%
Cep_AL	3.207	89,41%

Tabela 4 - Resultado da Caracterização dos dados ... (continuação)

Cep_RF	3.009	90,06%
Cidade_AL	926	96,94%
Cidade_RF	654	97,84%
Cpf_Mae	2.684	91,13%
Cpf_Pai	2.188	92,77%
DtNasc_AL	11.398	62,35%
EndComp_AL	1.505	95,03%
EndComp_RF	1.347	95,55%
Mae_AL	16.646	45,02%
Nome_AL	27.697	8,51%
Nome_RF	26.281	13,19%
Num_AL	1.705	94,37%
Num_RF	1.672	94,48%
Pai_AL	15.799	47,81%
País-AL	36	99,88%
Rua_AL	11.131	63,23%
Rua_RF	10.668	64,76%
UF_AL	28	99,91%
UF_RF	29	99,90%

Fonte: Próprio autor

Cada campo irá ser tratado como uma nova tabela, o software SAS University Edition apresenta o resultado da caracterização em uma página .html que foi exportada para um arquivo no formato .xlsx, onde foram acrescentados os campos código e categoria. O campo código foi preenchido com valore únicos para cada registro, criando uma identidade para cada registro. O campo categoria será utilizado para a etapa de classificação.

Após criado o arquivo com os dados caracterizados, estes foram exportados para dentro de um banco de dados utilizando o Microsoft Sql Server 2016 Express Edition, onde tem-se vinte duas tabelas preparadas para a classificação.

4.3. Análise dos Dados

Antes de começar a classificação é necessário criar algumas informações que serão de vital importância durante este processo. Então este será dividido em

4 etapas. Deve-se criar uma tabela de classificação e dicionários de palavras baseado nas regras de inserção.

A primeira etapa é a criação da tabela de classificação e dos dicionários. A tabela 4 foi baseada nas causas da baixa qualidade dos dados como descrito na seção 2.2.1

Tabela 5 - Classificação dos dados sujos

Código	Nome	Descrição
1	Aspecto Técnico	Dados não preenchidos.
2	Aspecto Humano	Erros na inserção de dados.
3	Fatores Organizacionais	Base de dados utilizada por diversos departamentos sem restrições.
4	Dados para análise	Não é possível validar os dados, requer análise dos dados.
5	Dados Corretos	Os dados estão de acordo com as regras de inserção.

Fonte: Próprio autor

As causas da baixa qualidade dos dados podem ser alteradas de acordo com os requisitos levantados. Para este estudo de caso foi inserido dados para análise e dados corretos. A classificação de dados para análise será utilizada quando não for apresentado documentos ou planilhas que possam garantir que os dados estão corretos. Estes podem ter a sua escrita correta, porém não tem como garantir que ele foi digitado corretamente, a citar como exemplo um nome de aluno, sua escrita está correta, porém não existe um documento que comprove se o nome foi digitado na sua totalidade. Por isto é necessário a criação de dicionários para a classificação.

Segundo as regras levantadas nos requisitos, para a base em análise deve-se levar em consideração que não se pode ter sinais de pontuação, acentos, caracteres matemáticos e abreviações, existem ainda campos que só podem ter valores numéricos e em outros valores somente *strings*. Com isto foram criados quatro dicionários distintos:

- Dicionário de Caracteres – Neste tem-se os sinais de pontuação, caracteres matemáticos e acentos. Servirá para classificar os dados do código 2 da tabela de classificação.

- Dicionário de Abreviações – Neste tem-se as abreviações mais comuns, se durante a fase de classificação for encontrado algum dado com abreviação inexistente neste dicionário, a mesma deve ser incluída. Servirá para classificar os dados do código 2 da tabela de classificação.
- Dicionário Numérico – Neste tem-se os valores numéricos de 0 a 9, servirá para classificar os dados onde não se pode ter estes valores.
- Dicionário de Strings – Neste tem-se todas as letras do alfabeto, servirá para classificar os dados onde não se pode ter estes valores.

Nas próximas etapas, será utilizado o software Microsoft Sql Server 2016 Express Edition para executar todos os scripts e análises visuais que forem necessárias.

A segunda etapa é a classificação que será executada de forma automática. Para isto, foi utilizado o Script A (Apêndice A) juntamente com o dicionário de caracteres, que realizou a leitura de dados em todas as tabelas do banco de dados, classificando-os quando encontrados e registrando com o código 2 no campo categoria.

Após esta classificação, foi feita uma análise visual nos dados que não foram classificados, encontrou-se registros com duplo espaço entre as palavras nas tabelas nome_rf e nome_al, então este foi incluso no dicionário de caracteres e repetiu-se a execução do script A, classificando novamente os dados.

Para classificar as abreviações será utilizado o script B (Apêndice B) juntamente com o dicionário de abreviações que classificará com o código 2, este script é executado manualmente, precisa-se trocar o nome da tabela, o motivo é que pode ter tabelas em que as abreviações são autorizadas, neste estudo de caso não se existe esta tabela, porém para se ter validade do processo é necessário que seja feito desta forma. Feita a análise visual dos dados classificados, foram encontrados diversos dados que não possuíam abreviações, estes foram classificados de forma errônea pelo script B. O problema estava na forma em que os usuários digitavam, grande parte das abreviações é precedida por ponto (.), mas as que foram encontradas estavam precedidas de espaço. Então foram inclusas no dicionário as abreviações precedidas de espaço. O Script B (Apêndice B) foi executado novamente.

Na terceira etapa será utilizado o dicionário numérico e o de strings. Estes

não podem ser utilizados em todas as tabelas. A classificação será de acordo com as regras de inserção, podendo variar de acordo com o que for encontrado.

O dicionário numérico foi utilizado juntamente com o Script C (Apêndice C), nas tabelas Bairro_Al, Bairro_RF, Cidade_AL, Cidade_RF, Pais_Al, Rua_Al, Rua_RF, Uf_Al, e Uf_RF. Os dados encontrados foram classificados com o código 2. Nas tabelas de nomes próprios como Nome_Rf, Nome_Al, Pai e Mae foi utilizado apenas um script de seleção buscando os registros que não foram classificados, quando encontrado algum caractere numérico, foi realizado a análise visual, foram encontrados registros com valores numéricos, porém como não existe regra para esta classificação estes foram classificados com o código 4.

O dicionário de strings foi utilizado juntamente com o script D, nas tabelas Cep_Al, Cep_Rf e Dt_Nasc_Al. Os dados encontrados foram classificados com o código 2.

Na quarta etapa foi utilizado análise visual em todas as tabelas. Neste ponto grande parte dos dados já foram classificados o que diminuiu consideravelmente os registros para análise.

Com isto foi encontrado dados com o valor nulo, oriundos dos registros que não foram realizados no momento do cadastro e nem foram acertados posteriormente, então estes foram classificados com o código 1.

Então as tabelas foram separadas em dois grupos:

- Grupo 1 - Dados oriundos de tabelas pre-cadastradas.
- Grupo 2 - Dados oriundos de tabelas que tem seus dados inseridos por diversos usuários

As tabelas do grupo 1 são Pais_Al, Uf_Rf e Uf_Al, onde os dados que não foram classificados, receberam o código 5. Já as tabelas restantes foram para o grupo 2 e tiveram os seus dados classificados com o código 4, pois estes dependem da validação dos documentos das pessoas registradas.

Os resultados destas classificações serão apresentados no capítulo 5, onde os resultados serão explanados.

5. RESULTADOS

Neste capítulo serão apresentados os resultados obtidos do capítulo 4. Onde foram analisados e classificados 666.006 (seiscentos e sessenta e seis mil e seis) registros. Ele está separado em três seções, onde na primeira seção será apresentado o resultado da caracterização dos dados. Na seção seguinte será explanado a classificação de cada tabela e ao final o resultado final da classificação dos dados.

5.1. Caracterização dos Dados

Com a importação de 30.273 (trinta mil, duzentos e setenta e três) registros de 22 (vinte e dois) campos, obtivemos a soma de 666.006 (seiscentos e sessenta e seis mil e seis) a serem agrupados, conforme demonstrado na Tabela 5.

Tabela 6 - Resultado da Caracterização dos dados – Redução ... (continua)

Nome Tabela	Registros		
	Total	Redução	Redução %
Bairro_AL	30.273	2.023	93,32%
Bairro_RF	30.273	1.977	93,47%
Cep_AL	30.273	3.207	89,41%
Cep_RF	30.273	3.009	90,06%
Cidade_AL	30.273	926	96,94%
Cidade_RF	30.273	654	97,84%
Cpf_Mae	30.273	2.684	91,13%
Cpf_Pai	30.273	2.188	92,77%
DtNasc_AL	30.273	11.398	62,35%
EndComp_AL	30.273	1.505	95,03%
EndComp_RF	30.273	1.347	95,55%
Mae_AL	30.273	16.646	45,02%
Nome_AL	30.273	27.697	8,51%

Tabela 7 - Resultado da Caracterização dos dados – Redução ... (continuação)

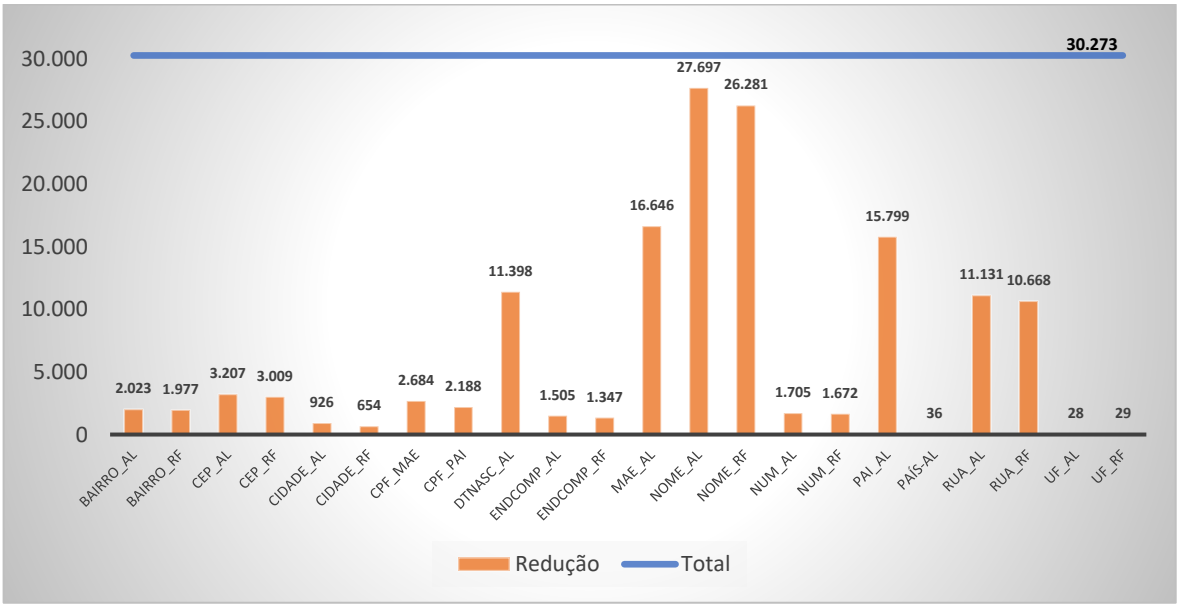
Nome_RF	30.273	26.281	13,19%
Num_AL	30.273	1.705	94,37%
Num_RF	30.273	1.672	94,48%
Pai_AL	30.273	15.799	47,81%
País-AL	30.273	36	99,88%
Rua_AL	30.273	11.131	63,23%
Rua_RF	30.273	10.668	64,76%
UF_AL	30.273	28	99,91%
UF_RF	30.273	29	99,90%
Total	**666.006**	**142.610**	**21,41 %**

Fonte: Próprio autor

Obteve-se então 142.610 (cento e quarenta e dois mil, seiscentos e dez) registros a serem analisados e classificados, o que representa 21,41% do total dos dados.

No Gráfico 1 está demonstrado de maneira mais clara a redução de cada campo. Observa-se que nos campos onde os dados são obtidos de tabelas pre-cadastradas obteve-se uma redução considerável. Os campos País_AL reduziu de 30.273 (trinta mil, duzentos e setenta e três) para 36 (trinta e seis) com redução de 99,88%, UF_AL 28 (vinte e oito) com redução de 99,91% e UF_RF 29(vinte e nove) com redução de 99,90%.

Gráfico 1 - Resultado Caracterização



Fonte: Próprio autor

Em média obteve-se uma redução de 78,59% de todos os registros caracterizados. Esta redução impactou diretamente na fase de análise dos dados, onde foi possível realizar a análise visual de forma rápida.

5.2. Análise dos Dados

Nesta seção estão explanados os resultados da análise e classificação dos dados. Foram 22 (vinte e duas) tabelas utilizadas nesta fase.

Para a classificação foram utilizados os dados da Tabela 2 do capítulo 4, seção 4.3. Com isto temos o seguinte:

- Aspecto Técnico – foram classificados todos os registros encontrado em branco ou com valor nulo.
- Aspecto Humano – foram classificados todos os registros que estavam com erros de digitação conforme os requisitos levantados no capítulo 4, seção 4.1. Para classificá-los foram utilizados os dicionários de caracteres, abreviações, números e strings.
- Dados para análise – Classificados todos os registros que passaram pela análise visual e não possuem documentos ou tabelas que comprovem a sua veracidade.
- Dados corretos – Classificados todos os registros que passaram pela análise visual e possuem documentos ou tabelas que comprovem a sua veracidade.

A seguir estão explanados os resultados obtidos em cada tabela.

Tabela 8 - Resultado Classificação - Bairro_AL

Bairro_AL			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	5.132	16,952%
2	Aspecto Humano	3.177	10,495%
4	Dados para análise	21.964	72,553%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 9 - Resultado Classificação - Bairro_RF

Bairro_RF			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	6.239	20,609%
2	Aspecto Humano	3.015	9,959%
4	Dados para análise	21.019	69,432%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 10 - Resultado Classificação - Cep_AL

Cep_AL			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	5.325	17,590%
2	Aspecto Humano	17.426	57,563%
4	Dados para análise	7.522	24,847%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 11 - Resultado Classificação - Cep_RF

Cep_RF			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	6.398	21,134%
2	Aspecto Humano	17.894	59,109%
4	Dados para análise	5.981	19,757%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 12 - Resultado Classificação - Cidade_AL

Cidade_AL			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	4.749	15,687%
2	Aspecto Humano	3.872	12,790%
4	Dados para análise	21.652	71,522%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 13 - Resultado Classificação - Cidade_RF

Cidade_RF			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	4.749	15,687%
2	Aspecto Humano	3.872	12,790%
4	Dados para análise	21.652	71,522%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 14 - Resultado Classificação - Cpf_Mae

Cpf_Mae			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	26.542	87,675%
2	Aspecto Humano	3.731	12,325%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 15 - Resultado Classificação - Cpf_Pai

Cpf_Pai			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	27.161	89,720%
2	Aspecto Humano	3.112	10,280%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 16 - Resultado Classificação - DtNasc_AL

DtNasc_AL			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	4.428	14,627%
2	Aspecto Humano	1	0,003%
4	Dados para análise	25.844	85,370%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 17 - Resultado Classificação - Endcomp_AL

Endcomp_AL			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	21.839	72,140%
2	Aspecto Humano	3.569	11,789%
4	Dados para análise	4.865	16,070%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 18 - Resultado Classificação - Endcomp_RF

EndComp_RF			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	22.703	74,994%
2	Aspecto Humano	3.185	10,521%
4	Dados para análise	4.385	14,485%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 19 - Resultado Classificação - Mae_AL

Mae_AL			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	8.874	29,313%
2	Aspecto Humano	5.840	19,291%
4	Dados para análise	15.559	51,396%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 20 - Resultado Classificação - Nome_AL

Nome_AL			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	4	0,013%
2	Aspecto Humano	9.549	31,543%
4	Dados para análise	20.720	68,444%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 21 - Resultado Classificação - Nome_RF

Nome_RF			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	8	0,026%
2	Aspecto Humano	8.808	29,095%
4	Dados para análise	21.457	70,878%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 22 - Resultado Classificação - Num_AL

Num_AL			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	6.530	21,570%
2	Aspecto Humano	1.197	3,954%
4	Dados para análise	22.546	74,476%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 23 - Resultado Classificação - Num_RF

Num_RF			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	7.664	25,316%
2	Aspecto Humano	1.152	3,805%
4	Dados para análise	21.457	70,878%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 24 - Resultado Classificação - Pai_AL

Pai_Al			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	9.320	30,787%
2	Aspecto Humano	8.627	28,497%
4	Dados para análise	12.326	40,716%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 25 - Resultado Classificação - Paìs_AL

País_AL			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	2.974	9,824%
2	Aspecto Humano	104	0,344%
4	Dados para análise	18	0,059%
5	Dados Corretos	27.177	89,773%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 26 - Resultado Classificação - Rua_AL

Rua_AL			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	4.693	15,502%
2	Aspecto Humano	11.859	39,174%
4	Dados para análise	13.721	45,324%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 27 - Resultado Classificação - Rua_RF

Rua_RF			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	5.822	19,232%
2	Aspecto Humano	12.051	39,808%
4	Dados para análise	12.400	40,961%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 28 - Resultado Classificação - UF_AL

UF_AL			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
1	Aspecto Técnico	3.937	13,005%
2	Aspecto Humano	6	0,020%
5	Dados Corretos	26.330	86,975%
Total		30.273	100,000%

Fonte: Próprio autor

Tabela 29 - Resultado Classificação - UF_RF

UF_RF			
Código	**Descrição**	**Quantidade de Registros**	**Resultado**
2	Aspecto Humano	2	0,007%
5	Dados Corretos	30.271	99,993%
Total		30.273	100,000%

Fonte: Próprio autor

5.3. Resultado Geral

Nesta seção está explanado o resultado geral da análise e classificação dos dados. Seguindo a gestão de qualidade de dados obteve-se a classificação dos dados sujos, conforme demonstrado na Tabela 28 nos itens com o código 1 e 2.

Estes apresentam um resultado, onde temos 46% dos registros apresentam problemas, sejam eles com a falta de informação ou com erros de digitação. Na análise de requisitos pode-se levantar que a instituição não possui um manual para a inserção de dados e sim boas práticas para a inserção dos dados.

Tabela 30 - Resultado Classificação – Geral

Código	Descrição	Quantidade de Registros	Resultado
1	Aspecto Técnico	185.091	28%
2	Aspecto Humano	122.049	18%
4	Dados para análise	275.088	41%
5	Dados Corretos	83.778	13%
Total		666.006	100%

Fonte: Próprio autor

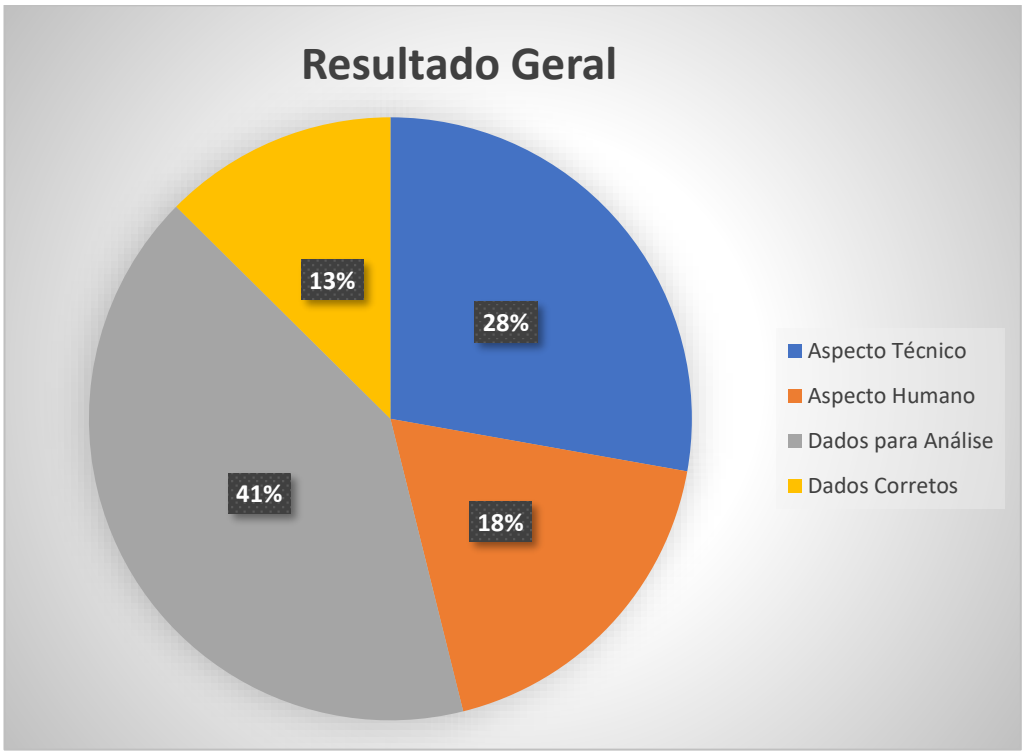
Os dados classificados com o código 4 na Tabela 28, representam 41% de dados classificados para análise, como dito anteriormente, não se pode confirmar

a veracidade dos mesmos, mas não quer dizer que estes estão errados.

Os dados classificados com o código 5 na Tabela 28, representa 13% de dados que não precisam ser higienizados, eles estão corretos.

Os resultados da Tabela 28 estão apresentados no Gráfico 2.

Gráfico 2 - Resultado Classificação – Geral



Fonte: Próprio autor

Ainda de acordo com os resultados obtidos, a metodologia utilizada neste estudo de caso, apresentou que é possível utilizar uma tarefa da mineração de dados juntamente com a gestão de dados para obter-se um resultado satisfatório na classificação dos dados sujos.

6. CONCLUSÃO

Com o presente trabalho buscou-se uma forma de utilizar a mineração de dados juntamente com a gestão de dados para a identificação dos dados sujos que porventura possam existir em uma base de dados estruturada. Para que isto fosse feito, buscou-se conhecimentos específicos em ambas as áreas de estudo, porém foi constatado que na primeira fase da mineração de dados que é de pré-processamento, existe a tarefa de limpeza de dados. Esta exige que estes sejam higienizados e com isto dar prosseguimento nas fases seguintes que são as de mineração de dados e a de pós-processamento. A partir deste fato, não foi possível utilizar um método de conhecimento em banco de dados para a classificação dos dados sujos.

Porém, na fase de mineração de dados, existe a tarefa de classificação que pôde ser utilizada de forma separada. Com isto foi possível construir a metodologia aplicada neste trabalho. Utilizou-se a qualidade de dados como guia para que fosse possível, aplicá-la.

A metodologia foi aplicada em três fases: levantamento de requisitos que buscou e apresentou as regras de inserção dos dados; caracterização dos dados que agrupou e reduziu a massa de dados; análise de dados que utilizou os resultados das duas primeiras fases para a construção do conhecimento e a aplicação da classificação dos dados sujos.

Na fase de levantamento de requisitos, apurou-se que mesmo que não existisse regras de inserção padronizadas, existia uma boa prática para a inserção, onde não era permitido a utilização de sinais de pontuação, acentos, caracteres matemáticos e abreviações nos campos nome do aluno, pai, mãe, responsável financeiro e endereço e nos campos cep, número do imóvel, cpf só podem ter caracteres numéricos.

Para a fase de caracterização dos dados, foi utilizado a tarefa de classificação da mineração, que agrupou, reduzindo em média 78,59% da massa de dados utilizada. Os campos que tinham sua informação, oriunda de um pré-cadastro tiverem uma redução em mais de 99%.

O resultado das duas primeiras fases, foi utilizado na análise de dados. Com

o resultado da primeira, construiu-se a tabela de classificação dos dados e os dicionários que foram utilizados na análise e classificação. Com o resultado da segunda fase, aplicou-se a classificação utilizando o dicionário de caracteres, este foi executado de forma automática em todas as tabelas onde 18% dos dados classificados apresentaram erro na digitação. Assim foi possível realizar a análise visual, onde encontrou-se 28% de registros sem nenhuma informação e apenas 13% dos dados puderam ser considerados como corretos. Os dados para análise obtiveram 41%.

Com isto, mesmo não utilizando um método de conhecimento em banco de dados, foi possível mesclar a gestão de dados com a mineração de dados. Com este resultado obtido, pode-se apresentar um norteamento para que os gestores possam acertar ou implementar um manual de inserção de dados, que evite que os mesmos continuem sendo incluídos da forma incorreta.

7. TRABALHOS FUTUROS

Como direção para trabalhos futuros pode-se sugerir o seguinte:

- Criar metodologia para a correção automática dos erros levantados no grupo aspecto humano.
- Estudo de caso para implementar o grupo de gestão de qualidade de dados, aplicando a metodologia deste trabalho para que se possa validar os dados que ficaram em análise.
- Utilizar este trabalho para construir um método para medir e monitorar continuamente os dados inseridos.
- Aplicar este trabalho na fase de pré-processamento na tarefa de limpeza de dados, da mineração de dados.

REFERÊNCIAS

BARBIERI, Carlos; FARINELLI, Fernanda. **Uma visão sintética e comentada do Data Mangement Body of Knowledge (DMBOK)**. Belo Horizonte, 2016. FumSoft. Disponível em < http://www.fumsoft.org.br/comunica/arquivos/uma_visao_sintetica_e_comentada_do_dmbok_fumsoft_carlos_barbieri.pdf >. Acesso em: 22/02/2017.

BRAGA, Luís Paulo Vieira. **Introdução à Mineração de Dados**. 2ª ed. Rio de Janeiro: Editora E-PAPERS SERVIÇOS EDITORIAIS LTDA, 2005.

CAMILO, Cássio Oliveira; SILVA, João Carlos. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**. Goiânia, 2009. Instituto de Informática Universidade Federal de Goiás Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf. Acessado em 19/06/2017.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Mining: Um guia prático**. 2ª ed. São Paulo: Editora CAMPUS, 2005.

SOUZA JÚNIOR, Geraldo Magela de . **Arquitetura de Dados: modelo conceitual e abordagem para criação e manutenção**. Belo Horizonte, 2014. Mestrado (Ciências Empresariais)-Universidade Fumec. Disponível em < https://www.lume.ufrgs.br/bitstream/handle/10183/127437/000968916.pdf?sequence=1>. Acesso em: 22/02/2017.

RÊGO, Bergson Lopes. **Gestão e Governança de Dados, Promovendo dados como ativo de valor nas empresas**. 1ª ed. Rio de Janeiro: Editora BRASPORT LIVROS E MULTIMÍDIA LTDA, 2013.

RIBEIRO, Ana Paula Tolentino. **Cenário de Gestão de Dados no Serpro comparado à referência conceitual do Guia DMBOK**. Porto Alegre, 2014. TCC (Especialização em Gestão Pública)-Universidade Federal do Rio Grande do Sul. Disponível em <

https://www.lume.ufrgs.br/bitstream/handle/10183/127437/000968916.pdf?sequence=1 >. Acesso em: 22/02/2017.

SAS, Institute. SAS, 2017. Disponível em <https://www.sas.com>. Acesso em: 13/09/2017.

SCHIMITT, Leandro Lopes. **Explorando Data Mining com Data Quality**. Canoas, 2009. Monografia (Ciência da Computação)-Centro Universitário La Salle. Disponível em <http://biblioteca.unilasalle.edu.br/docs_online/tcc/graduacao/ciencia_da_computacao/2009/llschimitt.pdf>. Acesso em: 22/02/2017.

VIEIRA, Mario Henrique Paes. **Aplicação de técnicas de mineração em um programa de concessão de benefícios ao consumidor: o caso do Programa Nota Legal do Distrito Federal**. Brasília, 2014. Dissertação (Pós-Graduação em Computação Aplicada) -Universidade de Brasília. Disponível em < http://repositorio.unb.br/bitstream/10482/17391/1/2014_M%C3%A1rioHenriquePaesVieira.pdf>. Acesso em: 12/12/2016.

APÊNDICE A

Script A – Classificação Geral – Aspecto Humanos

```
/*
Objeto: Classificação Geral - Aspecto Humanos
Autor: Claudio Lopes Monteiro
Data Criação: 03/09/2017
Data Modificação: 20/10/2017

Função
    Este script análise todas as tabelas que estão registradas na tabela <TABELA>
gerando a classificação de acordo com os dados que estão na tabela <TOKEN>.
    Após a classificação o mesmo gera o log na tabela <TOKENREG>.

Parâmetros
    A variável @Categoria precisa ser alterada de acordo com a tabela de erros.

*/



-- Variáveis
DECLARE @Tabela NVARCHAR(255);
DECLARE @Tabela2 NVARCHAR(255);
DECLARE @Token1 NVARCHAR(50);
DECLARE @Token2 NVARCHAR(50);
DECLARE @V1 NVARCHAR(1) = '%'; -- Caractere coringa
DECLARE @V2 NVARCHAR(1) = ''''; -- Caractere coringa
DECLARE @Cont1 AS SMALLINT;
DECLARE @Cont2 AS SMALLINT;
DECLARE @Rgtotal INT;
DECLARE @Categoria INT;

-- Codigo de Classificação
Set @Categoria = 2

-- Inicia Contador
SET @Cont1 = 1



WHILE @Cont1 <= (SELECT COUNT(*) FROM Tabela)

            BEGIN


                        -- Armazena Tabela que será utilizada
                        SET @Tabela = (SELECT Descricao FROM Tabela WHERE Codigo =
@Cont1);

                        -- Contador
```

```
                        SET @Cont1 += 1;

                        -- Inicia Contador Token
                        SET @Cont2 = 1

                        WHILE @Cont2 <= (SELECT COUNT(*) FROM TOKEN)

                                    BEGIN
                                                -- Zera o valor de Rgtotal
                                                SET @Rgtotal = 0;

                                                -- Token Utilizado
                                                SET @Token1 = (SELECT Token From
Token WHERE Codigo = @Cont2)

                                                -- Gera update e executa a
classificação na tabela
                                                SET @Tabela2 = 'Update '+@Tabela+'
SET Categoria = '+CONVERT(VARCHAR,@Categoria)+' from '+@Tabela+' where descricao
like '+@V2+@V1+@Token1+@V1+@V2;
                                                EXEC(@Tabela2);

                                                -- Gera o Log na tabela TokenReg
                                                DECLARE @T TABLE (VALOR INT)
                                                SET @TABELA2 = 'SELECT COUNT(*) FROM
'+@TABELA+' where descricao like '+@V2+@V1+@Token1+@V1+@V2;
                                                INSERT INTO @T EXEC(@TABELA2)
                                                SET @Rgtotal=  (SELECT CONVERT(INT,
VALOR) FROM @T);
                                                DELETE @T


                                                INSERT INTO TokenReg VALUES (@CONT2,
@TOKEN1, @TABELA, @Rgtotal);



                                                -- Contador
                                                SET @Cont2 += 1;
                                    END

            END
```

APÊNDICE B

Script B – Classificação Individual – Erro Aspectos Humanos - Abreviações

```
/*
Objeto: Classificação Individual - Erro Aspectos Humanos - Abreviações
Autor: Claudio Lopes Monteiro
Data Criação: 03/09/2017
Data Modificação: 20/10/2017

Função
    Este script gera a classificação de acordo com os dados que estão na tabela
<TOKENABR>.
    Após a classificação, o mesmo gera o log na tabela <TOKENREG>.

Parâmetros
    A variável @Tabela precisa ser alterada com o nome da tabela a ser classificada.
    A variável @Categoria precisa ser alterada de acordo com a tabela de erros.
            Alterar o nome das tabelas em Classificação da Tabela

*/

-- Variáveis
DECLARE @TOKEN NVARCHAR(50);
DECLARE @V1 NVARCHAR(1) = '%'; -- Caractere coringa
DECLARE @Rgtotal INT;
DECLARE @Contador AS SMALLINT;
DECLARE @Tabela NVARCHAR(30);
DECLARE @Categoria INT;

-- Inicia Contador
SET @Contador = 1

-- Tabela ser classificada
SET @Tabela = 'Uf_Resp$'

-- Codigo de Classificacao
SET @Categoria = 2


WHILE @Contador <= (SELECT COUNT(*) FROM TOKENABR)

            BEGIN

                        -- Zera o valor de Rgtotal
                        SET @Rgtotal = 0;

                        -- Token Utilizado
                        SET @TOKEN = (SELECT TOKEN FROM TOKENABR WHERE CODIGO =
@CONTADOR);

                        -- Classificação da Tabela
```

```
                        Update Uf_Resp$  SET Categoria = @Categoria from Uf_Resp$
where Descricao like @TOKEN+@V1;

                        -- Gera o Log na tabela TokenReg
                        SET @Rgtotal = (select COUNT(*) from Uf_Resp$ where
Descricao like @TOKEN+@V1);
                        Insert INTO TokenReg  VALUES (@Contador ,@Token,
@Tabela,@Rgtotal);

                        -- Contador - Condição de Saida
                        SET @CONTADOR += 1;

            END
```

APÊNDICE C

Script C – Classificação Individual – Erro Aspectos Humanos - Número

```
/*
Objeto: Classificação Individual - Erro Aspectos Humanos - Numero
Autor: Claudio Lopes Monteiro
Data Criação: 03/09/2017
Data Modificação: 20/10/2017

Função
    Este script gera a classificação de acordo com os dados que estão na tabela
<TOKENNUM>.
    Após a classificação, o mesmo gera o log na tabela <TOKENREG>.

Parâmetros
    A variável @Tabela precisa ser alterada com o nome da tabela a ser classificada.
    A variável @Categoria precisa ser alterada de acordo com a tabela de erros.
            Alterar o nome das tabelas em Classificação da Tabela
*/

-- Variáveis
DECLARE @TOKEN NVARCHAR(50);
DECLARE @V1 NVARCHAR(1) = '%'; -- Caractere coringa
DECLARE @Rgtotal INT;
DECLARE @Contador AS SMALLINT;
DECLARE @Tabela NVARCHAR(30);
DECLARE @Categoria INT;
-- Inicia Contador
SET @Contador = 1

-- Tabela a ser Classificada
SET @Tabela = 'RespFinanceiro$'

-- Codigo de Classificacao
SET @Categoria = 4


WHILE @Contador <= (SELECT COUNT(*) FROM TOKENNUM)

            BEGIN

                  -- Zera o valor de Rgtotal
                        SET @Rgtotal = 0;

                        -- Token Utilizado
                        SET @TOKEN = (SELECT TOKEN FROM TOKENNUM WHERE CODIGO =
@CONTADOR);

                        -- Classificação da Tabela
                        Update RespFinanceiro$  SET Categoria = @Categoria from
RespFinanceiro$ where Descricao like @V1+@TOKEN+@V1 AND CATEGORIA IS NULL;
```

```
                        -- Gera o Log na tabela TokenReg
                        SET @Rgtotal = (select COUNT(*) from RespFinanceiro$ where
Descricao like @V1+@TOKEN+@V1 AND CATEGORIA IS NULL);
                        Insert INTO TokenReg  VALUES (@Contador ,@Token,
@Tabela,@Rgtotal);

                        -- Contador - Condição de Saida
                        SET @CONTADOR += 1;

            END
```

APÊNDICE D

Script D – Classificação Individual – Erro Aspectos Humanos - Texto

```
/*
Objeto: Classificação Individual - Erro Aspectos Humanos - Texto
Autor: Claudio Lopes Monteiro
Data Criação: 03/09/2017
Data Modificação: 20/10/2017

Função
    Este script gera a classificação de acordo com os dados que estão na tabela
<TOKENNUM>.
    Após a classificação, o mesmo gera o log na tabela <TOKENREG>.

Parâmetros
    A variável @Tabela precisa ser alterada com o nome da tabela a ser classificada.
    A variável @Categoria precisa ser alterada de acordo com a tabela de erros.
            Alterar o nome das tabelas em Classificação da Tabela
*/

-- Variáveis
DECLARE @TOKEN NVARCHAR(50);
DECLARE @V1 NVARCHAR(1) = '%'; -- Caractere coringa
DECLARE @Rgtotal INT;
DECLARE @Contador AS SMALLINT;
DECLARE @Tabela NVARCHAR(30);
DECLARE @Categoria INT;

-- Inicia Contador
SET @Contador = 1

-- Codigo de Classificacao
SET @Tabela = 'Data_Nascimento$'

-- Codigo de Classificacao
SET @Categoria = 2




WHILE @Contador <= (SELECT COUNT(*) FROM TokenText)

            BEGIN

                        -- Zera o valor de Rgtotal
                        SET @Rgtotal = 0;

                        -- Token Utilizado
                        SET @TOKEN = (SELECT TOKEN FROM TokenText WHERE CODIGO =
@CONTADOR);
```

```
                        -- Classificação da tabela
                        Update Data_Nascimento$  SET Categoria = @Categoria from
Data_Nascimento$ where Descricao like @V1+@TOKEN+@V1 AND CATEGORIA IS NULL;

                        -- Gera o Log na Tabela TokenReg
                        SET @Rgtotal = (select COUNT(*) from Data_Nascimento$ where
Descricao like @V1+@TOKEN+@V1 and categoria is null);
                        Insert INTO TokenReg  VALUES (@Contador ,@Token,
@Tabela,@Rgtotal);

                        -- Contador - Condição de Saida
                        SET @CONTADOR += 1;

            END
```

ANEXO

Autorização para utilização dos dados

Solicitação de Dados

Odon Faiçal Loures para mim 28 de nov

Conforme contato com o Magnífico Reitor Antonio Fonseca da Silva, está liberado a base de dados para estudos.
Att
Odon Faiçal Loures
Analista de Sistemas

De: Claudio Monteiro [mailto:monteiroclaudio@gmail.com]
Enviada em: segunda-feira, 16 de outubro de 2017 13:49
Para: Odon - Unec
Assunto: Solicitação de Dados
…

Venho através deste, solicitar a V.Sa. a permissão para trabalhar com os dados cadastrais de pessoas que estão inclusas no sistema TOTVS.
A finalidade desta autorização é para que eu possa utiliza-los no meu Trabalho de Conclusão de Curso. O trabalho proposto esta descrito abaixo. Meu nome é Claudio Lopes Monteiro, nascido em Caratinga/MG, aluno do Curso de Ciências da Computação da Rede Doctum, estou atualmente no 8º Período.

Os dados que serão analisados e trabalhados são os descritos abaixo:
1 - Dados Cadastrais
2 - Dados que serão analisados: Nome, Endereço completo, cpf, data nascimento, sexo, telefone, nome do pai e mae .
O intuito e analisar estes dados, verificar os que estão incorretos e propor uma solução para acerta-los e mante-los da forma correta. Para isto será utilizado técnicas de trabalho da Mineração de Dados e da Gestão e Governança de Dados.

A Mineração de Dados é utilizada para a definição de um problema, após isto ela prepara, explora estes dados validando eles para a utilização.

A Gestão e Governança de Dados é utilizada para gerir os dados de uma empresa, validando processos e organizando os dados para se tornarem informações valiosas para a empresa.

Os dados solicitados só serão utilizados para testar e confirmar a metodologia que sera utilizada. Será entregue uma cópia do resultado da análise e irei propor o acerto dos dados da maneira mais prática possível.

Tema Proposto - Modelo para corrigir e gerenciar cadastros

Objetivo Geral
Utilizar a *Mineração de Dados* para sanitizar uma base de dados e aplicar os métodos da *Gestão e Governança de Dados* para a padronização e manutenção permanente dos dados sanatizados.

Desde já agradeço a atenção de V.sa..

Conto com sua colaboração,

Atenciosamente,

Claudio Lopes Monteiro