UNIVERSIDADE FEDERAL DE SERGIPE
CAMPUS ALBERTO CARVALHO
DEPARTAMENTO DE SISTEMAS DE INFORMAÇÃO

JÉSSICA DA SILVA COSTA

MINERAÇÃO DE DADOS APLICADA À ANÁLISE DO PADRÃO DE METILAÇÃO DO GENE BRCA1, ENVOLVIDO NO CÂNCER DE MAMA

ITABAIANA
2016
UNIVERSIDADE FEDERAL DE SERGIPE
CAMPUS ALBERTO CARVALHO
DEPARTAMENTO DE SISTEMAS DE INFORMAÇÃO

JÉSSICA DA SILVA COSTA

MINERAÇÃO DE DADOS APLICADA À ANÁLISE DO PADRÃO DE METILAÇÃO DO GENE BRCA1, ENVOLVIDO NO CÂNCER DE MAMA

Trabalho de Conclusão de Curso submetido ao Departamento de Sistemas de Informação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Dr. Methanias Colaço Rodrigues Júnior

Co-Orientadora: Msc. Melline Fontes Noronha

ITABAIANA
2016

ATENÇÃO: este texto deve ser impresso no verso da contracapa.

Sobrenome, Nome.

Título do trabalho / Nome completo – Itabaiana: UFS, Ano.

99f. (indica o número de páginas do trabalho); 99 cm (indica o tamanho)

Trabalho de Conclusão de Curso (graduação) – Universidade Federal de Sergipe, Curso de [Nome do curso], Ano.

1. Assunto. 2. Área de Concentração - TCC. 3. Curso. I. Título.

JÉSSICA DA SILVA COSTA

MINERAÇÃO DE DADOS APLICADA À ANÁLISE DO PADRÃO DE METILAÇÃO DO GENE BRCA1, ENVOLVIDO NO CÂNCER DE MAMA

Trabalho de Conclusão de Curso submetido ao corpo docente do Departamento de Sistemas de Informação da Universidade Federal de Sergipe (DSIITA/UFS) como parte dos requisitos para obtenção do grau de Bacharel em Sistemas de Informação.

Itabaiana, 25 de Maio de 2016

BANCA EXAMINADORA:

_____

Prof(a) Methanias Colaço Rodrigues Junior, Doutor
Orientador
DSIITA/UFS

_____

Prof(a) Alcides Xavier Benicasa, Doutor
UFS

_____

Prof(a) Melline Fontes Noronha, Mestre
UNICAMP

Dedico

A meus pais, irmãs e amigos.

AGRADECIMENTOS

Durante estes anos que estive na Universidade Federal de Sergipe passei por muitas coisas boas e ruins, mas todas me fizeram amadurecer como pessoa e profissional. Nesse caminho encontrei muitas pessoas que estiveram comigo nessa minha jornada. Agradeço a Deus por nunca ter perdido a persistência e ter a oportunidade de realizar meu curso. Aos meus pais (João e Lúcia) e minhas irmãs (Jaqueline e Natália) que estiveram comigo sempre.

Quero agradecer aos meus orientadores Methanias e Melline que me aguentaram este tempo todo com minhas perguntas e momentos de correria. Sem vocês este trabalho não seria realizado. Seus conhecimentos me propiciaram um crescimento científico imenso, além de serem grandes amigos.

Não poderia deixar de agradecer a minha turma de amigos da UFS (não dá para citar todos, mas eles sabem demais) que incentivavam sempre com sua forma peculiar de incentivos, risos. Quero agradecer também ao pessoal do meu trabalho atual que compreendeu este meu momento mais conturbado e me incentivava sempre quando eu chegava estressada ou estava cheia de coisas para fazer.

Não poderia deixar de agradecer a todos os meus professores, desde o ensino fundamental até a faculdade. Sempre carreguei seus ensinamentos para minha vida acadêmica, profissional e pessoal. Sempre me ensinaram que não poderia desistir diante das dificuldades. Posso dizer que muito do que sou é fruto de um trabalho feito por todos os professores que tive em minha formação.

Enfim, posso agradecer a vida e os momentos por me propiciar crescer cada dia mais...

Epígrafe

"Milho de pipoca que não passa pelo fogo continua a ser milho para sempre. Assim acontece com a gente… As grandes transformações acontecem quando passamos pelo fogo." *(Rubem Alves)*

COSTA, Jéssica da Silva. **Mineração de Dados aplicada à análise do padrão de metilação do gene BRCA1, envolvido no câncer de mama**. 2016. Trabalho de Conclusão de Curso – Curso de Sistemas de Informação, Departamento de Sistemas de Informação, Universidade Federal de Sergipe, Itabaiana, 2016.

RESUMO

O uso de técnicas de mineração de dados é muito importante para extrair informações relevantes de grandes bases de dados. Sua utilização pode ser aplicada em várias áreas inclusive biologia. O câncer é uma das patologias mais estudadas atualmente e diversos experimentos são realizados, o que gera grandes bases de dados. Diante desse quadro, este trabalho objetiva a aplicação do algoritmo Árvores de decisão com rede Bayesiana em uma base de dados de um experimento sobre câncer de mama. Foram realizadas três análises diferentes que indicaram algumas faixas de valores frequentes de metilação.

Palavras-chave: Mineração de dados. Câncer. Árvores de decisão

ABSTRACT

The use of data mining techniques is very important to extract relevant information of large databases. Its use can be applied in several fields including biology. Cancer is one the most studied diseases currently and several experiments are performed, which generates large databases. Given this situation, this graduation work aims the application of Decision tree algorithm with Bayesian network in a database about breast câncer experiment. Three different analyzes that indicated some tracks frequent methylation values were performed.

Key-words: *Data Mining. Cancer. Decision Trees.*

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

NCBI	*National Center for Biotechnology Information*
UFS	Universidade Federal de Sergipe
INCA	Instituto Nacional de Câncer
DNA	Ácido dexorribonucleico
RNA	Ácido ribonucleico
DDBJ	DNA Databank of Japan
EMBL	European Molecular Bank Laboratory
CRISP-DM	*Cross Industry Standard Process for Data Mining*
AD	Ánalise discriminante
KVP	K-vizinhos mais próximos
SVM	Máquinas de vetores suporte
ACP	Análise de Componentes Principais
WBCD	*Wisconsin Breast Cancer Database*
KDD	*Knowledge Discovery in Database*

SUMÁRIO

1 INTRODUÇÃO

O câncer é uma patologia maligna que possui características como crescimento anormal e descontrolado das células, podendo invadir outros tecidos e órgãos, processo chamado de metástase. O câncer de mama é o câncer mais comum em mulheres e atinge 1 em cada 10 mulheres no mundo ocidental (ANJUM et al, 2014). No Brasil, em 2016, são estimados 57960 novos casos de câncer de mama, com risco estimado de 56,20 casos para cada 100 mil mulheres (INCA, 2016). Pesquisadores descobriram que os genes BRCA (BRCA1 e BRCA2), quando mutados devido à herança genética ou espontaneamente, podem predispor ao surgimento de câncer mamário (KOMEN, 2015). Ainda conforme ANJUM e colaboradores, variações epigenéticas também podem contribuir à suscetibilidade ao câncer (ANJUM et al, 2014).

Variações epigenéticas são mutações que não alteram a estrutura do DNA, porém interferem em seus processos, um deles é o mecanismo da metilação o qual inativa genes. Para medir os níveis de metilação em humanos são utilizados experimentos específicos que geram muitos dados. Essas análises em larga escala, assim como outros estudos na área de genética, geram milhares de dados e analisá-las manualmente é muito custoso e demorado. Por isso a bioinformática entrou como uma área de estudo interdisciplinar que visa resolver e analisar problemas biológicos através do uso de técnicas computacionais.

Uma técnica computacional que pode ser utilizada para análise de dados genômicos é a mineração de dados. Através desta técnica é possível encontrar padrões, testar hipóteses e encontrar informações ocultas que podem ser útil para resolução de problemas (SINGH e DAS, 2007). Existem várias técnicas de mineração para bioinformática, entre elas, classificação, associação, clustering ou agrupamento e regressão. A aplicação da técnica mais apropriada dependerá do objetivo e de quais dados estão disponíveis.

A proposta desde trabalho é aplicar o algoritmo de Árvores de decisão em uma base de dados de pacientes portadoras de mutações no gene BRCA1, com câncer de mama, sem o câncer de mama e não portadoras da mutação, a fim de analisar se existe algum padrão de metilação em pessoas portadoras e não portadoras de mutação no gene BRCA1 que predisponha ao surgimento de câncer mamário. A base dados de pacientes utilizados nesta proposta será formada por dados públicos de pacientes, disponíveis no banco de dados do NCBI (*National Center for Biotechnology Information).*

1.1. Motivação

O câncer de mama é uma das patologias mais estudadas atualmente devido a sua alta incidência e diversidade. A ciência evoluiu e possibilitou diversos tratamentos, no entanto existem muitas dúvidas em relação ao câncer, suas causas e evolução. A informática como área-meio para a saúde contribui para análise de dados que ajudam em diagnósticos, prognósticos e descoberta de possíveis causas.

No entanto, unir conhecimentos tão diferentes torna-se uma tarefa difícil e uma área multidisciplinar surge para fazer esta ligação. A bioinformática atua como uma área de apoio, unindo a biologia molecular e a computação para gerar informações biológicas de forma rápida, eficiente e organizada. Existem muitas formas de atuação da informática na biologia, mas uma das mais utilizadas é a mineração de dados.

Mineração de dados é uma das técnicas mais utilizadas quando há necessidade de encontrar alguma informação oculta em grande volume de dados. Experimentos da área biológica costumam resultar em muitos dados que aparentemente não fazem sentido porém ao serem examinados podem resultar em alguma informação importante. A base de dados utilizada neste experimento é grande e analisa-la seria muito custoso e complicado, daí o uso de técnicas de mineração de dados que possibilitam a descoberta de conhecimento.

1.2. Justificativa

Este trabalho visa utilizar uma base de dados pública com dados de experimentos da área de genética que normalmente estão em formato textual e transforma-la em modelo relacional. Após isso aplicar alguma técnica de mineração de dados para analisar a possibilidade de existir alguma faixa de valores de metilação que predisponha ao câncer mamário. Este conhecimento será útil para pesquisadores da área de bioinformática que poderão utilizar estes resultados para avaliações mais precisas.

1.3. Objetivos do Projeto

1.3.1. Objetivo Geral

O objetivo deste projeto é encontrar se existe algum padrão de metilação em genes de

pessoas com mutação no gene BRCA1 e associar essa mudança no padrão à possível predisposição e prognóstico da patologia utilizando o algoritmo Árvore de decisão.

1.3.2. Objetivos Específicos

- Descrever e comparar estudos sobre câncer de mama disponíveis na literatura;
- Criar uma base de dados de *microarray* de pacientes portadores e não portadores da mutação do gene BRCA1, transformando dados que estão em formatos textuais e planilhas para o modelo relacional.

1.4. Organização da Monografia

Este trabalho está organizado da seguinte forma:

O capítulo 2 explica alguns conceitos relevantes sobre biologia molecular que foram utilizados neste trabalho. Além disso, há algumas explicações sobre patologias cancerígenas e banco de dados biológicos e trabalhos relacionados.

O capítulo 3 explica conceitos sobre mineração de dados e a técnica de mineração de dados Árvores de decisão que foi utilizada neste trabalho.

O capítulo 4 mostra o estudo de caso do trabalho efetuado. Todos os passos são explicados de acordo com a metodologia CRISP-DM e o dicionário dos dados do modelo criado.

O capítulo 5 contém os resultados das análises realizadas pelo algoritmo, bem como os as árvores e os gráficos gerados.

O capítulo 6 mostra a conclusão da análise realizada e do trabalho como um todo. Além disso, mostra trabalhos futuros que podem ser realizados a partir do que foi feito nesta monografia.

2 BIOLOGIA MOLECULAR E CÂNCER

2.1. Conceitos Básicos de Biologia Molecular

Genética é a ciência que estuda os mecanismos dos seres vivos que transferem características a seus descendentes. A descoberta da estrutura da molécula de DNA por Watson e Crick, na década de 1950, marcou a área de genética, embora a existência dos ácidos nucleicos (DNA e RNA) já fosse conhecida há algumas décadas atrás.

Os ácidos nucleicos são polímeros formados por unidades menores chamados nucleotídeos, estes são compostos por um carboidrato, um grupo fosfato e uma base nitrogenada, as quais são adenina(A), guanina(G), citosina(C), uracila(U) ou timina(T). O DNA é uma fita dupla em formato de hélice que possui as bases mencionadas anteriormente, contudo possui timina no lugar da uracila. O RNA é uma fita simples que também possui as bases mencionadas, no entanto possui a uracila no lugar da timina. Um segmento de DNA que possui uma função especifica como a codificação de uma determinada proteína é denominado de gene (VERLI, 2014).

Estas duas moléculas possuem um papel de destaque no Dogma Central da Biologia, postulado por Francis Crick em 1958, ao ilustrar como a informação genética do DNA é transmitida e preservada. O dogma diz que a informação contida no DNA pode ser replicada e ser transcrita em RNA, o qual será traduzido em proteínas. Estas são sintetizadas no citoplasma em uma organela chamada ribossomo e para que o processo aconteça o RNA mensageiro (mRNA) faz a ligação. O mRNA funciona como um molde para que as proteínas possam ser sintetizadas no citoplasma. Todo este processo da codificação do DNA em proteínas é chamado de expressão gênica e constitui um elemento fundamental para o funcionamento do organismo dos seres vivos (REIS et al., 2011). A figura exemplifica este processo do dogma central:

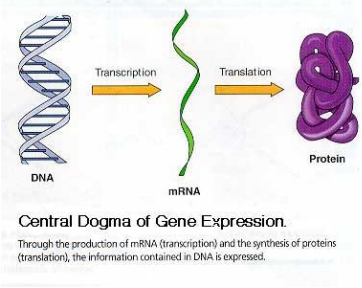

Figura1: Dogma Central da Biologia

Fonte: http://nossobioma.blogspot.com.br/2010/04/dogma-central-da-biologia.html

Expressão gênica é o processo pelo qual uma informação que está em um gene é transcrita para o RNA mensageiro (mRNA) e depois traduzida em algum tipo de proteína (BELTRAMINI et al., 2015). A expressão gênica pode variar entre organismos e até mesmo em diferentes órgãos e tecidos de um mesmo indivíduo. Para medir o padrão de expressão gênica, isto é, conhecer quais genes estão ativos e inativos entre amostras biológicas (organismos, tecidos, etc), são utilizadas algumas técnicas, dentre as mais utilizadas: o *microarray* (microarranjos) e *RNA-seq* (sequenciamento em larga escala de mRNA).

Microarray ou chips de DNA são placas que contém milhares de spots em lâminas de vidro onde são colocadas gotículas que contém DNA com sequências de genes específicos. Em sua superfície são colocadas moléculas de mRNA que atraídas pelo DNA dos spots por complementaridade. (KOUTSOS et al., 2010).

O DNA pode sofrer modificações nos cromossomos que se mantém estáveis ao longo das divisões celulares, mas que não alteram suas bases. Estas são chamadas de alterações epigenéticas. Metilação é uma dessas alterações e constitui um fator a ser analisado, pois seus padrões alterados interferem no comportamento dos genes. O processo de metilação é uma modificação química em que um grupamento metil (CH_3) é adicionado ao quinto carbono da citosina de dinucleotídeos CpG (formados pela ligação citosina e guanina). Ela ocorre em regiões do genoma ricos em dinucleotídeos CpG chamadas de "ilhas de CpG". Estas ilhas são regiões do DNA maior do que 200 pares de bases que contém aproximadamente metade das bases C e G e com a presença de aproximadamente 60% de dinucleotídeos CpG (OLIVEIRA et al., 2010). Embora o processo de metilação seja natural e importante para a regulação gênica, um exemplo é a proteção do genoma contra sequências de DNA viral que é inativado pela adição dos grupos metila, seu padrão alterado pode "silenciar" a função de determinados genes predispondo ao surgimento de patologias (MEDVEDEVA, 2011; OLIVEIRA et al., 2010; HILL et al., 2011).

Para o estudo do padrão de metilação em genes, são utilizados Chips de *microarray* específicos para padrões de metilação em humanos, como por exemplo, o BeadChipHumanMethylation450 e HumanMethylation27 da Illumina (ILLUMINA, 2015; ANJUM et al, 2014). Através destes experimentos é possível medir o nível de transcrição ou metilação dos genes estudados. Para este trabalho os dados foram obtidos através do uso do HumanMethylation27.

Figura 2: HumanMethylation27

Fonte: http://support.illumina.com/array/array_kits/infinium_humanmethylation27_beadchip_kit.html

O HumanMethylation27 é um experimento criado pela empresa Illumina (trabalha com experimentos e métodos para área de biologia molecular) para medição de níveis de metilação em genoma humano. Este experimento mede os níveis de metilação em 27578 dinucleotídeos CpG em 14495 genes. Para a realização deste experimento, é utilizada uma quantidade de DNA genômico para conversão de bissulfito para converter citosina não metilada em uracila. Esse DNA tratado passa por um processo amplificação de genoma passando por algumas enzimas. No chip do experimento é colocada a citosina metilada e a não metilada em locus específico após passarem por alguns processos específicos para o experimento. Ciclos repetidos de coloração são realizados para diferenciar as regiões. Assim o chip é digitalizado para que os níveis de fluorescência sejam analisados em cada região. Os níveis de metilação oscilam entre 0 e 1, onde 0 significa uma região com nenhuma metilação e 1 totalmente metilada. A partir destes experimentos são gerados alguns dados com a região de CpG e o seu valor de metilação no momento do experimento (ILLUMINA, 2015).

2.2. Câncer e os genes BRCA

O câncer é uma patologia maligna decorrente do crescimento anormal e desordenado das células as quais passam a se chamar células cancerosas. Em muitos casos, o câncer pode entrar em um estado de metástase, ou seja, acometer outros tecidos. As causas do câncer são as mais diversas e incluem causas internas e externas ao organismo.

As causas internas estão relacionados à questão genética e a capacidade do organismo se defender de agressões externas. No entanto, 80% a 90% dos casos de câncer estão relacionados a fatores externos como estilo de vida, exposição ao sol, envelhecimento, entre outros (INCA, 2016). Estes fatores podem predispor ao aparecimento de modificações

genéticas estruturais ou epigenéticas.

Genes BRCA (BRCA1 e BRCA2) são conhecidos como genes supressores de tumor pois impedem o crescimento desordenado de células e ajudam na reparação do genoma (AMENDOLA e VIEIRA, 2005). Estudos recentes indicam que mutações em genes BRCA podem inibir a ação destes e predispor ao aparecimento de tumores mamários (ANJUM et al, 2014). Como dito anteriormente, existem mutações não herdadas que não alteram a estrutura do DNA, por isso possuem possibilidade de serem revertidas com o uso de algumas substâncias (HILL et al., 2011). Além disso, descobrir a relação dos níveis de metilação com o aparecimento de tumores torna-se um fator importante para diagnóstico e prognóstico do câncer.

As células cancerosas podem formar uma massa de células ou tumores que podem formar vasos sanguíneos que manterão o crescimento desordenado. Estas podem se desprender do tumor e pelo sistema circulatório invadir outros tecidos. Como as células são células doentias e não cumprem suas funções normalmente, elas causam problemas no tecido em que estão como exemplo dores de cabeça devido a tumores cerebrais (INCA, 2016). Dependendo da evolução da doença, o indivíduo acometido pode ser levado a óbito em pouco tempo.

O grande número de dados gerados de biologia molecular possibilitou o desenvolvimento de várias pesquisas e a geração de vários bancos de dados. O NCBI, *National Center for Biotechnology Information*, foi criado em 1988 pelo NIH, *National Institutes of Health*, nos Estados Unidos, para abrigar dados de experimentos de biologia molecular. Atualmente, o NCBI inclui dados de outros bancos internacionais: o DDBJ (*DNA Data Bank of Japan*) e o EMBL (*European Molecular Bank Laboratory*) (PROSDOCIMI, 2007 apud Tateno et al., 2002). No NCBI existem vários estudos sobre câncer dos mais diversos tipos. A base escolhida para este trabalho especificamente trabalha com níveis metilação relacionado ao câncer de mama, mas existem outros estudos no NCBI que relacionam com outros cânceres.

2.3 Trabalhos Relacionados

Atualmente diversos estudos são realizados na área de biologia e informática, pois percebeu-se a necessidade de compreender melhor os dados resultantes de experimentos e dados clínicos. Estes trabalhos descritos a seguir mostram como os dados podem ser

analisados utilizando algumas técnicas de informática, mais precisamente mineração de dados. Nem todos encontraram os resultados satisfatórios, mas sugeriram melhorias na técnica ou algoritmo para que novos testes fossem realizados.

O trabalho de BEIZANUR et al. (2014) focou na descrição das técnicas de bioinformática e bancos de dados mais utilizadas em pesquisas de bioinformática aplicada ao câncer, principalmente o câncer de mama. Em seu artigo é mostrado os benefícios do uso de bioinformática em tratamentos e desenvolvimento de fármacos através do estudo genético do câncer.

Alguns destes trabalhos utilizaram bases de dados de estudos sobre genética como o trabalho de RODRIGUES e AMARAL (2012) que aplicaram cinco métodos de mineração de dados à base NCI60, provenientes de experimentos de microarray com níveis de expressão gênica de 1000 genes para classificar nove tipos de câncer. Os métodos utilizados do ambiente Weka foram o J48, Random Forest, PART, IBK e Naive Bayes. Pela análise realizada, o que classificou melhor as amostras foi o método IBK. Já SILVA e AMARAL (2011) utilizaram a mesma base do trabalho anterior (NCI60) para implementar um algoritmo genético que pudesse obter os mesmos valores de aptidão nesta base de 1000 genes que anteriormente foi utilizada com 55 genes. No entanto, ele não convergiu como a de 55 genes e os autores propuseram algumas mudanças para que o objetivo fosse atingido.

Outro trabalho que usou bases de dados de genética foi o de ANJUM et al. (2014) que analisou a possibilidade do processo de metilação do DNA em genes BRCA1 predispor ao câncer de mama não hereditário, visto que a grande maioria dos casos de câncer de mama são desta forma. Foram utilizadas neste trabalho três bases de dados sobre níveis de metilação de diferentes estudos (extraído de células sanguíneas e bucais) e aplicado um algoritmo de classificação. Neste trabalho, foram utilizadas técnicas de regressão multivariada para ajuste de idade, presença de câncer, etc. Após este tratamento, um algoritmo de classificação de rede elástica do pacote R foi utilizada para criar um classificador que comprimia 1829 CpGs com coeficientes de regressão diferentes de 0. Um dos resultados encontrados indicou que 2514 mutações no BRCA1 associadas a regiões de CpG indicavam que 57% eram hipermetiladas e 43% eram hipometiladas. Os resultados demonstraram que a assinatura de metilação de DNA derivada dos genes BRCA1 ajudam a predizer o risco de câncer de mama.

HOLSBACH et al (2014) aplica inicialmente a técnica multivariada Análise de Componentes Principais (ACP) nos dados e posteriormente as técnicas de mineração de dados K-vizinhos mais próximos (KVP) e análise discriminante (AD) para classificar a malignidade

ou benignidade do câncer mamário na base de dados WBCD (Wisconsin Breast Cancer Database) de exames clínicos de mama. A utilização do KVP obteve acurácia de 97,77%, ao reter 5,87 variáveis, enquanto que o AD obteve acurácia de 97,07%, ao reter 5,95 variáveis. SOCZEK e ORLOVSKI (2013) também usaram técnicas para classificar quanto à benignidade ou malignidade do câncer de mama. Foram utilizadas duas bases de dados e testados algoritmos de regras de classificação, árvores de decisão e Redes Neurais Artificiais. Em suas análises, o algoritmo que melhor classificou a primeira base foi o algoritmo de regras de classificação Jrip com *Split*. Ele obteve uma taxa de acerto de 97% e estatística Kappa de 0,9479, considerada muito boa. A segunda base foi melhor classificada com o algoritmo Jrip com Split, no entanto seus valores não se equipararam a primeira base, pois a segunda obteve taxa de 75% de acerto e estatística Kappa de 0,3889.

O trabalho de CHAHINE (2013) usou uma abordagem diferente, pois utilizou dados clínicos para a análise. As técnicas de mineração de dados Árvores de Decisão e Máquina de Vetores Suporte (SVM) foram utilizadas para criar um modelo com o objetivo de determinar se um câncer de próstata poderá sofrer metástase. Para isso foram utilizados dados da disciplina de urologia disponibilizados pelo Hospital das Clínicas. Segundo o autor, a menor taxa de erros foi de 13% no algoritmo em árvores de decisão. Embora o método não dispense um médico, este método ajuda médicos no processo de tomada de decisões mais rápida, conclui o autor.

3 MINERAÇÃO DE DADOS

O KDD (*Knowledge Discovery in Databases)* ou descoberta de conhecimento em base de dados é um processo muito importante e essencial para descobrir conhecimento oculto em grande volume de dados. Para FAYYAD et al. (1996), KDD refere-se a todo o processo de descoberta de conhecimento em dados e *Data Mining* ou Mineração de Dados é um fase do processo de KDD. Ainda segundo FAYYAD et al. (1996), *Data Mining* refere-se ao processo de aplicação de algoritmos específicos para extração de padrões em dados. O processo de KDD possui algumas fases essenciais, são elas: seleção dos dados, pré-processamento dos dados, transformação dos dados, mineração de dados e a avaliação ou interpretação dos resultados.

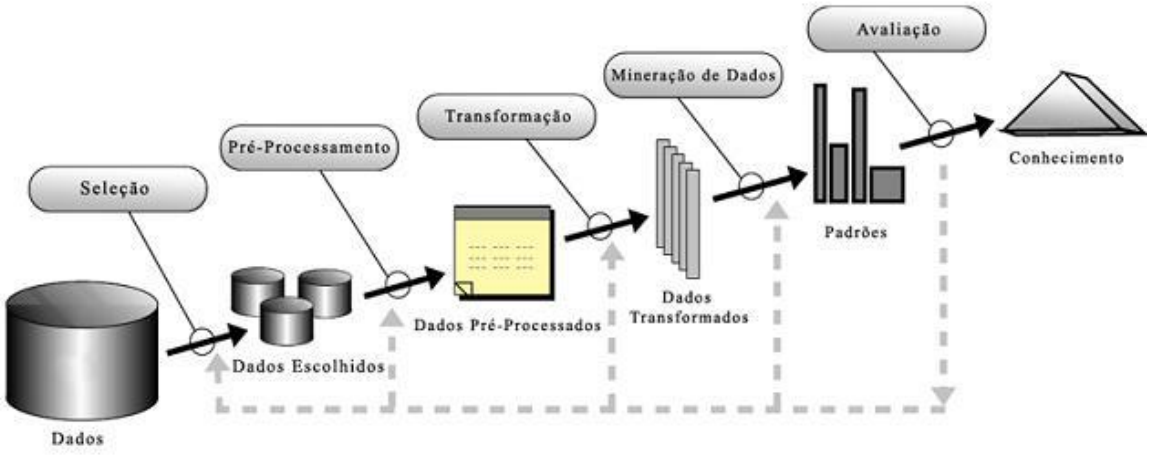

Figura 3. Processo de KDD

Fonte: http://www.devmedia.com.br/mineracao-de-dados-com-orange/31678

O processo de KDD envolve várias fases, como ilustrado na figura 3, que são muito importantes para se chegar a resultados satisfatórios. Dado um grande volume de dados é preciso selecionar os dados que serão utilizados baseados nos objetivos da análise, é primeira fase chamada de Seleção. Depois da seleção começa a fase de Pré-processamento em que entram algumas tarefas de limpeza dos dados para eliminação de ruídos e inconsistências. Muitas vezes os dados são de diversas fontes como arquivos texto, planilhas, diversas bases de dados e é preciso transformá-los e uni-los para que possam ser utilizados em conjunto para ser aplicada alguma técnica de mineração, esta é a fase de transformação dos dados. Após

estas fases começa a fase de Mineração de dados em que são aplicadas técnicas de mineração com mais diversos objetivos. Os resultados da análise precisam se transformar em conhecimento e para isso é nesse interpretar estes resultados, esta fase é chamada de Avaliação. Neste momento é preciso associar as informações obtidas com os objetivos que se pretendia com a análise.

3.1. Técnicas de Mineração de Dados

Existem diversas técnicas de mineração para descoberta de conhecimento em base de dados com mais diversos fins. Técnicas de mineração consistem na especificação dos métodos que garantem como descobrir padrões que interessam (AMO, 2004). A seguir serão explicadas algumas técnicas de mineração de dados mais utilizadas.

Classificação é uma técnica de mineração em que dados de entrada são mapeados para determinadas classes pré-definidas. Dessa forma estes atributos estão correlacionados à classe ao qual pertencem (CASTANHEIRA, 2008). A tarefa de classificação é um tipo de aprendizado supervisionado visto que as classes já são conhecidas.

Regressão é similar à classificação, no entanto ela trabalha com dados contínuos ao invés de dados discretos (CASTANHEIRA, 2008). A regressão trabalha com a ideia de estimativa em que é possível predizer o valor de um atributo baseado nos demais.

Associação caracteriza o quanto que a presença de um conjunto de dados implica na presença de outro conjunto de dados (CASTANHEIRA, 2008). Regras de associação são extremamente úteis para mecanismos de recomendação, por exemplo, é possível saber que quando um cliente compra o produto X, ele compra o produto Y. A associação trabalha com a ideia de que algo aconteça em conjunto com a outra. Isso pode ser uma estratégia muito útil para negócios nos mais diversos ramos.

Agrupamento ou Clustering é um tipo de particionamento em grupos com características similares, no entanto não há uma definição da quantidade ou quais grupos serão formados após a aplicação de algoritmos. Dessa forma os grupos serão formados proximidade de características em comum. Essa abordagem caracteriza um aprendizado não-supervisionado.

3.1.1. Árvores de Decisão

Árvores de decisão são estruturas que utilizam pequenas regras de decisão para dividir problemas em partes menores até gerarem uma previsão ou decisão. Ela é formada por um conjunto de nós de decisão que ajudam a classificar ou prever algum atributo baseado na entrada.

Para a execução da análise do banco de dados de metilação foi utilizada a árvore de decisão da Microsoft. O algoritmo desta árvore trabalha com dados discretos e contínuos. Para a previsão de dados discretos, ele analisa o relacionamento dos dados de entrada, no entanto para dados contínuos, ele utiliza regressão linear para detectar em que momento a árvore irá se dividir (MSDN, 2016).

Além disso o algoritmo aplica abordagem Bayesiana para aprender os modelos de interação causal o que lhe permite obter distribuições aproximadas dos modelos. Através do uso de redes Bayesianas o algoritmo calcula as probabilidades posteriores de acordo com os dados de treinamento. Para a avaliação das informações para o aprendizado é utilizado a suposição da equivalência de probabilidade. Esta afirma que os dados não devem ajudar a discriminar estruturas de rede que de outra forma representam as mesmas asserções de independência condicional. Baseado nisso cada caso terá uma única rede bayesiana e uma medida de confiança para a rede (HECKERMAN et al., 1995).

Para pontuar o ganho de informação o algoritmo Árvores de Decisão da Microsoft permite a utilização de três técnicas para atributos discretos: Entropia de Shannon, Bayeasiana com K2 a priori e Bayesiana Dirichlet com uniforme a priori. Para atributos contínuos é calculado somente o interesse, pois o mesmo suporta somente valores contínuos e o mesmo não calcula ganho de pontuação pelas técnicas mencionadas anteriormente, neste casos ele utiliza regressão linear (MSDN, 2016).

A Árvore de decisão da Microsoft possui algumas peculiaridades em relação a parâmetros de configuração da Árvore de Decisão. O parâmetro COMPLEXITY_PENALTY inibe o crescimento da árvore quando seu valor é alto e aumenta a divisão da árvore quando ele é baixo. O padrão é 0,5 se há de um a nove atributos, 0,9 se há de 10 a 99 atributos e 0,99 se há 100 ou mais atributos. O FORCE_REGRESSOR força que determinadas colunas virem regressoras da árvore, está função somente está disponível na versão Enterprise do SQL Server. O SCORE_METHOD indica o método de pontuação que será utilizado: 1 para Entropia, 3 para Bayesiano com K2 a priori e 4 para Bayesiano Dirichlet Equivalente com Uniforme a priori.

4 ESTUDO DE CASO

Neste capítulo são apresentados todos os passos e métodos utilizados para a execução deste trabalho. Na seção 4.1 é explicada a metodologia usada para todo o processo de mineração de dados. Na seção 4.2 é explicada os passos realizados na execução e como eles seguiram a metodologia proposta.

4.1. Objetivos e Metodologia CRISP-DM

Este projeto visa aplicar um algoritmo de mineração de dados para extrair e relacionar informações biológicas sobre processos de metilação nos genes. Foi utilizada a metodologia CRISP-DM (Cross Industry Standard Process for Data Mining) que foi criada em 1996 com o objetivo de padronizar os processos de mineração de dados. Ela compreende seis fases, são elas:

- Business Understanding (Entendimento do negócio) – esta fase é relativa ao entendimento do negócio, dos objetivos e requerimentos do projeto;
- Data Understanding (Entendimento dos dados) – esta fase visa compreender , verificar os problemas e os primeiros insights nos dados;
- Data Preparation (Preparação dos dados)– esta fase é responsável pelo tratamento dos dados (seleção, limpeza e transformação) para a fase posterior;
- Modeling (Modelagem) – esta fase é responsável pela aplicação das técnicas (modelagem e algoritmos) de mineração de dados;
- Evaluation (Avaliação) – esta fase consiste na construção de um modelo de alta qualidade para análise dos resultados;
- Deployment (Implantação) – fase da aplicação do conhecimento gerado na tomada de decisões. (CHAPMAN et al, 2000).

A seguir serão explicados como estas fases da metodologia CRISP-DM foram aplicadas em todos os passos do projeto. Além disso temos um tópico com um dicionário de todos os dados.

4.2. Fases da metodologia

Seguindo as fases da metodologia CRISP-DM, a primeira fase foi o Entendimento do negócio para isso foi feita uma pesquisa exploratória em artigos e materiais de instituições

que trabalham com pesquisas relativas aos seguintes temas: genética-epigenética e patologias cancerígenas. Na área computacional, os estudos serão feitos sobre os seguintes temas: mineração de dados e algoritmos de mineração de dados.

A segunda fase foi o Entendimento dos dados para isso foi selecionada uma base de dados de um experimento disponível publicamente no NCBI [Número de acesso GSE57285]. Esse experimento é composto por dados de perfis de metilação de DNA no gene BRCA1 de aproximadamente 27000 ilhas de CpGs, originadas de amostras de sangue de 84 pacientes mulheres. Dessas 84 pacientes, 42 não possuíam a mutação no gene BRCA1, 7 possuíam a mutação, mas nenhuma patologia cancerígena e 35 possuíam a mutação e desenvolveram o câncer de mama. Os dados foram obtidos do experimento com o Illumina Infinium 27K Human DNA Methylation Beadchip v1.2.

O experimento é feito com a linkagem de uma sequência de DNA metilado (M) e outra não metilada (U). Para cada região de CpG, o status de metilação é calculado pelo raio dos sinais de fluorescência $\beta = Max(M,0)/[Max(M,0)+Max(U,0)+100]$. O valores β são variáveis contínuas entre 0 (sem metilação) e 1 (completamente metilada) representando o raio do alelo metilado com a intensidade do locus combinado (Informações Suplementares, ANJUM, 2014).

A terceira fase foi a Preparação dos dados para o modelo proposto, para isso foi selecionado alguns dados da planilha da empresa Illumina sobre o experimento e arquivos individuais das 84 pacientes. A base resultante foi modelada em três entidades que representam a Pessoa, o Status da Mutação e a Região de CpG. A entidade Pessoa possui dados descritivos do indivíduo que se submeteu ao experimento: idade, sexo e o tipo de tecido. A entidade que representa o Status da Mutação descreve qual mutação a Pessoa possui e se está com câncer. A entidade Região de CpG representa as 27 578 regiões de CpG. Várias pessoas possuem várias regiões de CpG e para representar este relacionamento, uma nova entidade Pessoa e Região de CpG liga o indivíduo a suas ilhas de CpG e seus respectivos valores.

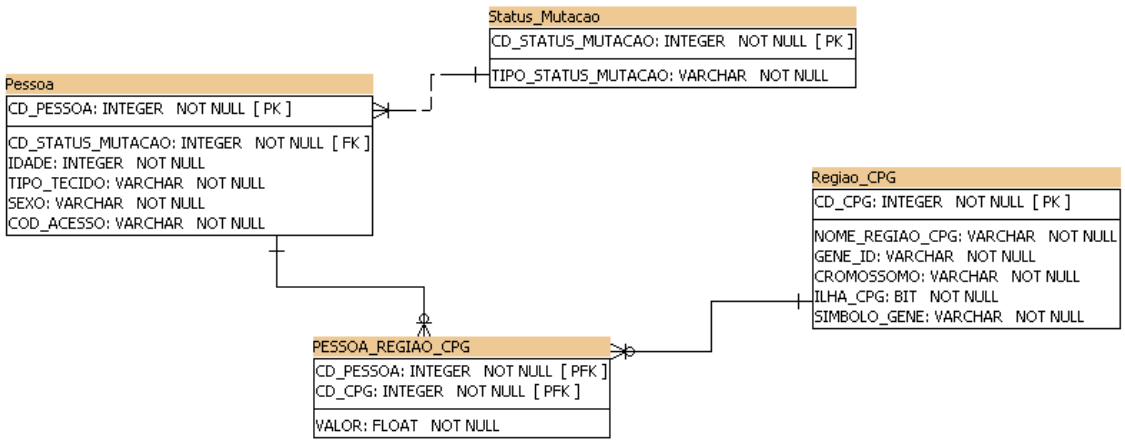

Figura. 4 – Modelagem dos dados

Após realizada a modelagem, os dados das pessoas foram inseridos nas tabelas. Os dados estavam em formato txt e em planilhas em formato .xls. Para efetuar a importação desses dados para o Banco de Dados SQL Server, foi necessária a utilização da ferramenta de importação de dados do Integration Services. Todos os arquivos individuais foram importadas um a um e conferidos para que os dados estivessem íntegros após a importação. Da planilha foram selecionados os seguintes dados: o identificador do cg, o identificador do gene, o cromossomo, um atributo para identificar se o local é uma região de CpG e o símbolo do gene. Estes ficaram na tabela de REGIAO_CPG. Foram selecionados dados dos arquivos texto que possuem relevância para a análise, os dados de identificação das pessoas (sexo, idade e o tecido analisado) que foram colocados na tabela PESSOA e a tabela principal das 27578 regiões de cg e seus respectivos valores de metilação. Os valores de metilação foram colocados na tabela PESSOA_REGIAO_CPG que liga a tabela PESSOA com REGIAO_CPG.

A quarta fase ou Modelagem foi aplicação do algoritmo para análise dos dados. O algoritmo escolhido foi Árvore de decisão e para isso foram feitas três análises diferentes com mudança dos dados de entrada e dos previsíveis. Para isso foi criada uma view ou visão dos dados com a seleção dos seguintes atributos: ID_ARTIFICIAL (representa o número registro na view), CD_PESSOA, NOME_REGIAO_CPG, VALOR, ILHA_CPG, CROMOSSOMO, SIMBOLO_GENE, IDADE e TIPO_STATUS_MUTACAO. Durante a criação da view foram eliminados registros que possuíam valores nulos nos níveis de metilação pois não é recomendável prever estes valores através de algum algoritmo porque estes valores vieram de

um experimento.

A quinta fase ou Avaliação é o resultado das análises realizadas e os gráficos que representam como o algoritmo se comportou. Há um capítulo específico que mostra os resultados obtidos e suas respectivas representações gráficas. A sexta fase ou Implantação seria o uso dos resultados obtidos e como aplica-los em algum contexto. No entanto, essa parte não pode ser realizada devido à necessidade de análises posteriores.

Fica evidente a tipificação deste trabalho como um estudo de caso, pois o mesmo possui um caráter investigativo adequado a problemas complexos em que estão envolvidos diversos fatores (ARAÚJO et al, 2008). O trabalho investigará um conjunto de dados na busca de um padrão que interfira em um processo de importância significativa, neste caso a metilação de regiões do DNA.

4.3. Dicionário dos dados

Os dados utilizados no modelo descrito são resultantes da utilização de uma base de dados disponível no NCBI cujo número de acesso é GSE57285 e de uma planilha de dados da empresa Illumina, responsável pelo experimento que gerou os dados apresentados anteriormente.

Na tabela PESSOA existe os códigos cd_pessoa e cd_status_mutacao que identificam a pessoa e o código da mutação o qual identifica o tipo de mutação, 0 para BRCA1 sem mutação e saudável, 1 para BRCA1 mutante e saudável e 2 para BRCA1 mutante com câncer de mama, estas são as descrições do atributo tipo_status_mutacao que se encontram na tabela STATUS_MUTACAO. Ainda na tabela PESSOA existem a idade, tipo_tecido e cod_acesso que identificam faixa etária, o tecido do corpo da pessoa e número de acesso individual da tabela de valores de metilação da pessoa, respectivamente.

Na tabela de REGIAO_CPG foram utilizados dados da planilha da empresa Illumina que descreviam o CpG. O código cd_cpg é um sequencial das regiões das 27578 regiões de CpG. Os atributos nome_regiao_cpg, gene_id, cromossomo, ilha_cpg e símbolo_gene identificam o nome da região de cpg estudada (exemplo cg00000292), a identificação do gene que representa em que local se encontra a região, a identificação do cromossomo que contém o gene, se esta região é considerada uma ilha de CpG (atributo 0 ou 1, com um significado booleano) e o símbolo do gene que contém a região, respectivamente.

A tabela REGIAO_CPG e PESSOA possuem um relacionamento N:N e para isso foi criada uma tabela que simplificava o relacionamento entre as duas. A tabela PESSOA_REGIAO_CPG contém os códigos de pessoa e do cpg (estas formam a chave primária da tabela) e o valor de metilação de uma pessoa com sua região de cpg.

5 RESULTADO DA ANÁLISE DOS DADOS

Para analisar estes dados com Árvore de Decisão foram utilizadas três abordagens para dados de entrada e dado previsível. Nas duas análises realizadas que previam o atributo discreto foram testadas os três tipos de ganhos de pontuação: Entropia de Shannon, Bayeasiana com K2 e Bayesiana Dirichlet com uniforme a priori, os resultados foram muito similares. A primeira abordagem baseia-se na ideia de passar algumas entradas como cromossomo, nome_região_cpg e o valor. O atributo previsível foi o tipo_status_mutação que indicava três possibilidades: BRCA1 mutante com câncer, BRCA1 mutante e saudável e BRCA1 sem mutação e saudável. Após a análise, a probabilidade de previsão obtida foi de 50, 46%. As cores das barras dos nós da árvore indicam a probabilidade de ser um dos três tipos de status de mutação baseados nas faixas de valores.

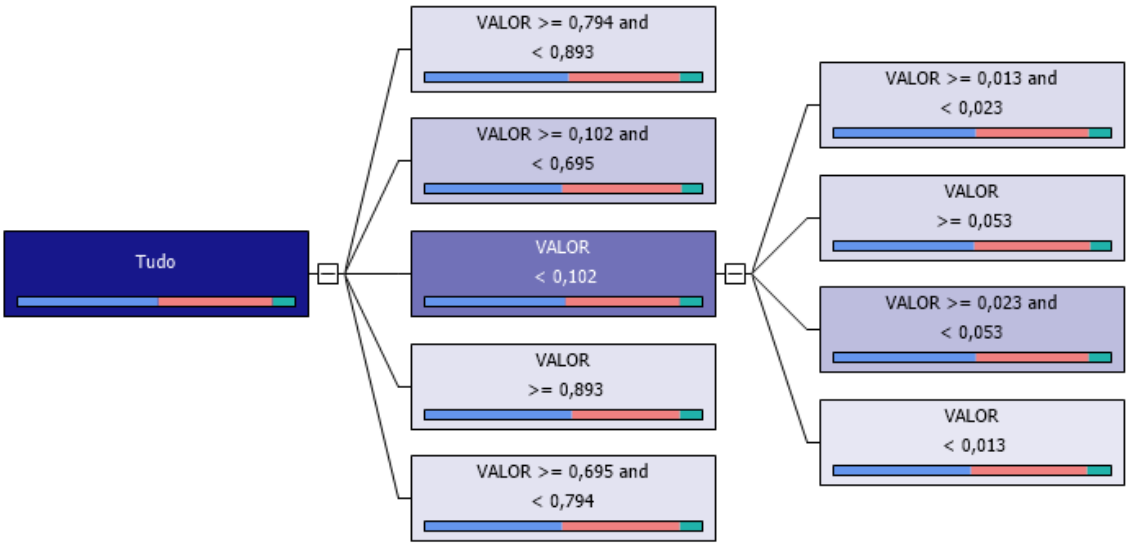

Figura 5. Análise de metilação com a primeira abordagem

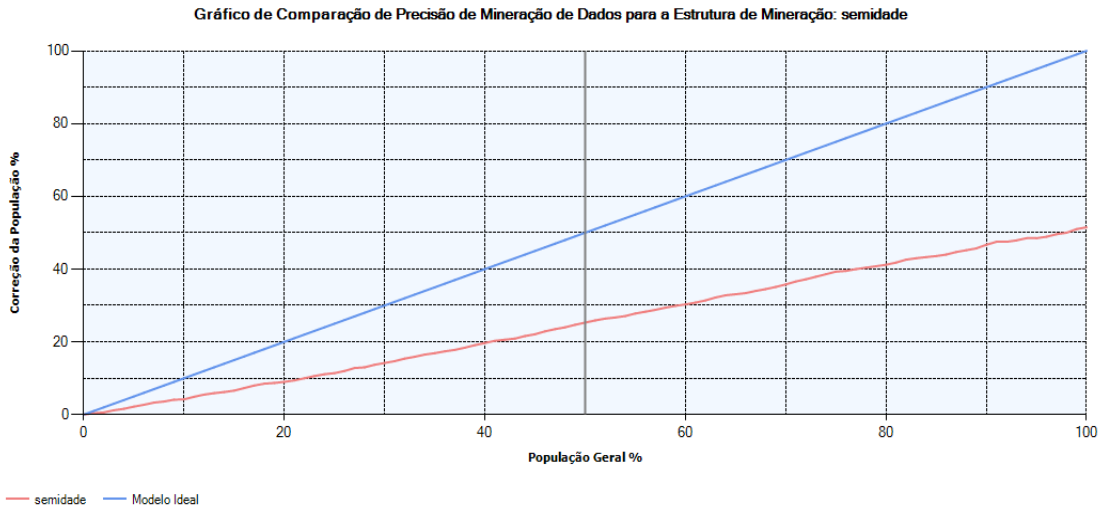

Gráfico 1. Gráfico de previsão da primeira abordagem

A segunda abordagem foi utilizada uma previsibilidade invertida. Desta vez o valor seria o atributo previsível e as entradas foram nome_região_cpg, valor, cromossomo e tipo_status_mutacao. Embora o dado previsível (classe alvo) fosse contínuo, as entradas eram todas discretas o que fez com o algoritmo tratasse tudo como discreto. Quando existem dados contínuos como entrada, o algoritmo trabalha com regressão linear. Pela análise gerada, ele destacou um determinado caminho na árvore gerada dando um coeficiente de 0.279. Para gerar esta árvore, o parâmetro COMPLEXITY_PENALTY foi alterado para 0.1 para que a árvore obtivesse mais subdivisões. A árvore abaixo mostra o modelo gerado depois da análise e seu respectivo gráfico de previsão.

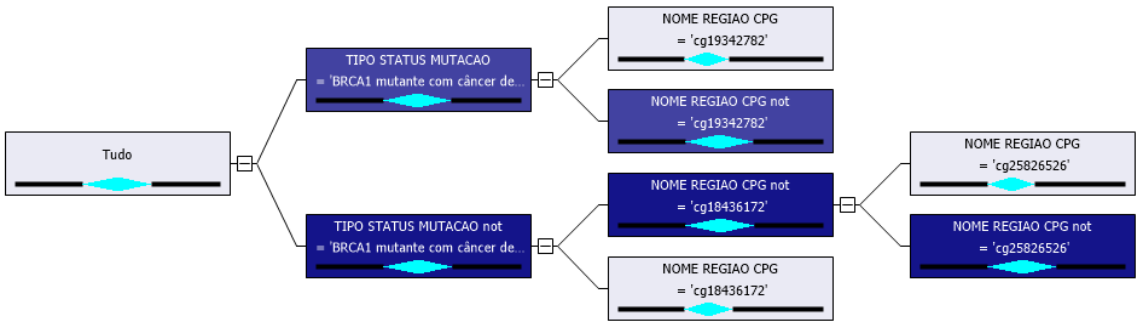

Figura 6. Análise de metilação com a segunda abordagem

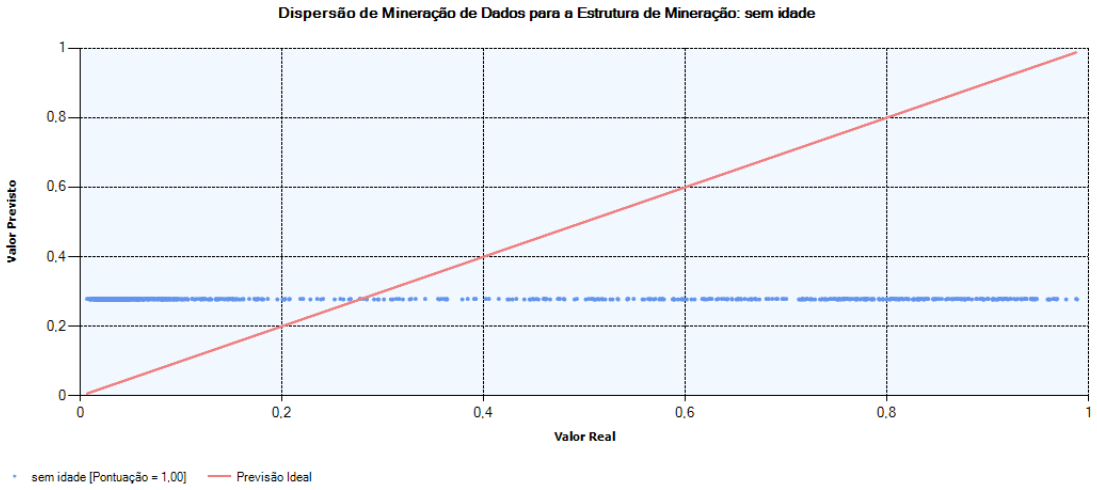

Gráfico 2 . Gráfico de previsão com a segunda abordagem

A terceira abordagem utilizou alguns atributos para prever o tipo_status_mutação. Desta vez os atributos escolhidos para a entrada foram idade, cromossomo, nome_região_cpg e o valor. Ao utilizar a idade percebeu-se que a amostra não era muito distribuída em relação a este atributo. Quando existiam vários indivíduos com a mesma idade, o algoritmo comparava as faixas de valores de metilação, no entanto em alguns casos existia somente uma pessoa com determinada idade o que impossibilitava a comparação com outros indivíduos. Isso acontecia principalmente com os registros que possuíam o status_mutacao_cancer que indicava a existência da mutação e saúde, pois eram registros em menor quantidade. A probabilidade de previsão obtida foi 74,59%. As figuras a seguir mostram a árvore de decisão gerada e seu respectivo gráfico de previsão.

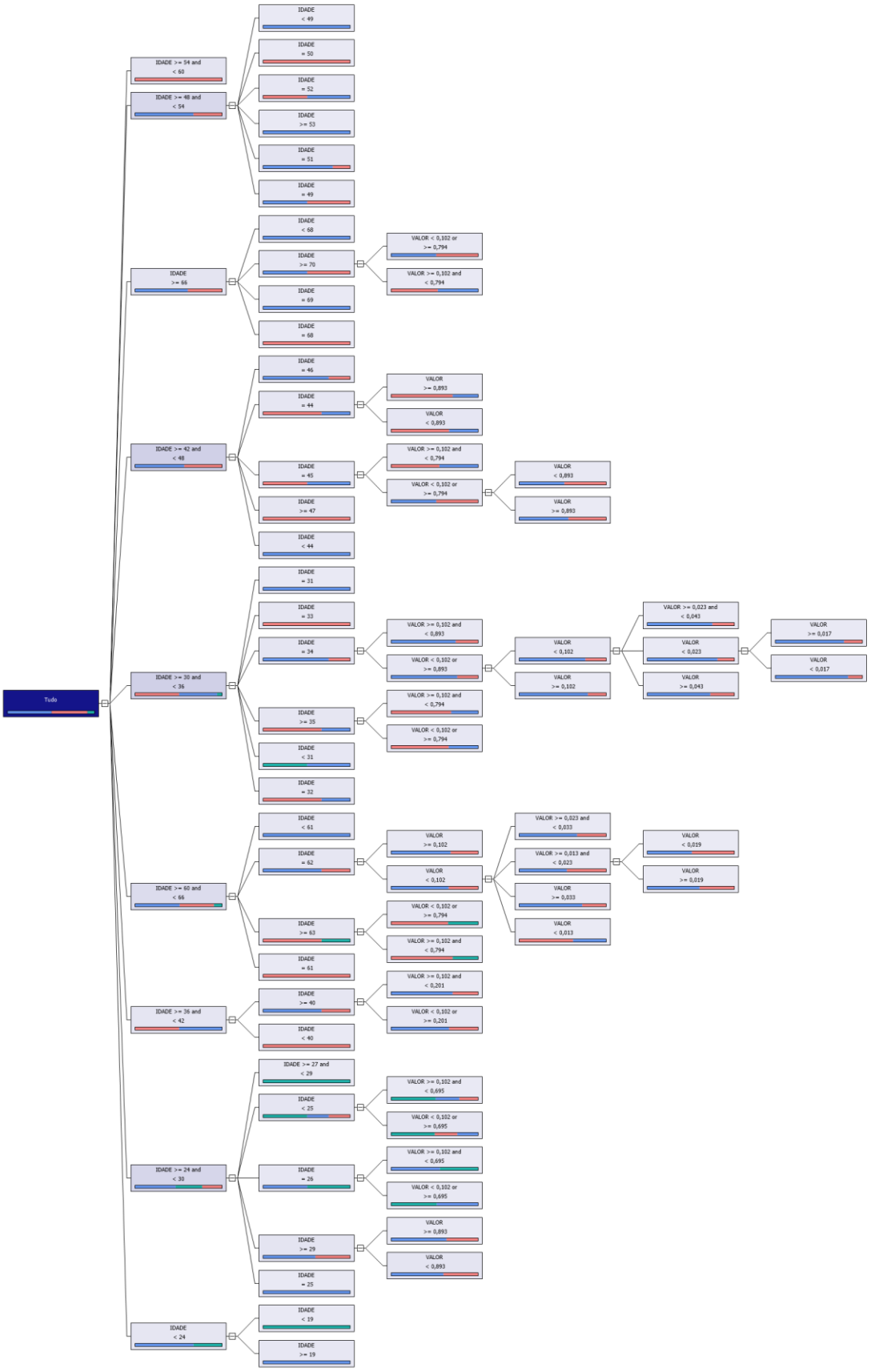

Figura 7. Análise de metilação com a terceira abordagem

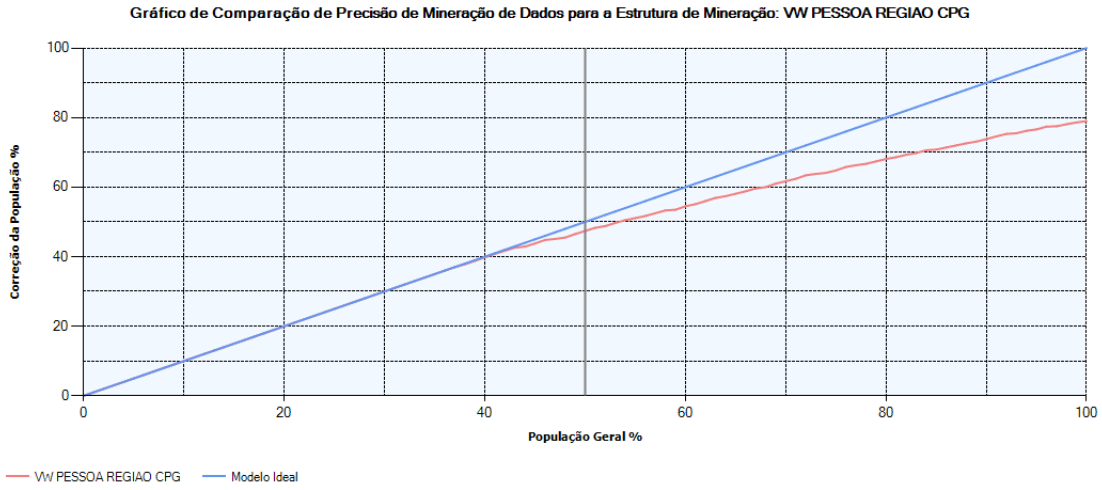

Gráfico 3. Gráfico de previsão com a terceira abordagem

6 Conclusão

A análise realizada não obteve uma faixa específica de metilação que indicasse a possiblidade de surgimento de um câncer mamário. O algoritmo obteve algumas faixas de valores que indicavam uma determinada quantidade de pessoas que eram propensas a serem de cada um dos três casos de status de mutação. Quando a idade estava entre as entradas o algoritmo separava os dados pelas idades e comparava valores ao criar faixas de valores que se tornavam frequentes em determinada idade. No entanto, ele não utilizava a região de cpg para comparar pessoa a pessoa com a existência da idade.

Ao realizarmos a análise com o valor como previsível percebemos que o algoritmo tratava o valor como atributo discreto devido a todas as entradas serem discretas embora a classe alvo fosse contínua. A análise poderia ser feita em formato horizontal analisando as regiões CG, no entanto a quantidade de regiões de CpG impossibilitava o uso do algoritmo por gerar 27578 colunas.

Para trabalhos futuros seria relevante aprofundar o estudo com as árvores de decisão, testar outras técnicas como redes neurais ou até outros algoritmos de Aprendizado Máquina que pudessem analisar esta base de dados sob outras perspectivas e obter outros resultados. Desta forma seria possível realizar algumas comparações com os resultados obtidos e talvez encontrar algum padrão de nível de metilação que indicasse uma possibilidade de câncer. Outra forma interessante seria analisar ouras bases de dados sobre metilação para compararmos resultados e termos outras amostras.

REFERÊNCIAS

- INCA, Instituto Nacinal do Câncer. **Estimativa 2016 Incidência de Câncer no Brasil. 2016**. Disponível em http: //www.inca.gov.br/estimativa/2016/sintese-de-resultados-comentarios.asp

- AMENDOLA, Luiz Cláudio Belo; VIEIRA, Roberto. **A contribuição dos genes BRCA na predisposição hereditária ao câncer de mama.** Revista Brasileira de Cancerologia. INCA: 2005. Disponível em: http://www1.inca.gov.br/rbc/n_51/v04/pdf/revisao3.pdf

- KOUTSOS, Anastasios; MANAIA, Alexandra; WILLINGALE-THEUNE, Julia. **Introdução aos Microarrays de ADN**. EMBL: 2010. Disponível em: http://www.embl.it/training/scienceforschools/teacher_training/teachingbase/microarray_port/introduction_microarrays_port.pdf

- MEDVEDEVA, Yulia A. (2011). **Algorithms for CpG Islands** Search: New Advantages and Old Problems, Bioinformatics - Trends and Methodologies, Dr. Mahmood A. Mahdavi (Ed.), ISBN: 978-953-307-282-1, InTech, DOI: 10.5772/22883. Disponível em: http://www.intechopen.com/books/bioinformatics-trends-and-methodologies/algorithms-for-cpg-islands-search-new-advantages-and-old-problems

- BELTRAMINI, Leila Maria; SILVA, Aparecido Rodrigues da; COSTA, Gislaine. **O processo de expressão gênica**. 2015, Material de Apoio para Oficina. Disponível em: http://cbme.usp.br/files/mat_apoio/Roteiro_para_oficina_sobre_Expressao_Genica.pdf

- ILLUMINA. **Dna Methylation Analysis**. 2015, Material da empresa. Disponível em: http://www.illumina.com/Documents/products/datasheets/datasheet_dna_methylation_analysis.pdf

- VERLI, Hugo. **Bioinformática da Biologia à Flexibilidade Molecular**. 1ª edição,

São Paulo, 2014, 282 p. Disponível em: http://www.ufrgs.br/bioinfo/ebook/ .

- ANJUM S, FOURKALA E-O, Zikan M, et al. **A *BRCA1*-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival**. *Genome Medicine*. 2014; 6(6):47. doi:10.1186/gm567. Disponível em: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4110671/.

- OLIVEIRA, Naila Francis Paulo de; PLANELLO Aline Cristiane; ANDIA, Denise Carleto; PARDO, Ana Paula de Souza. **Metilação de DNA e Câncer.** Revista Brasileira de Cancerologia, 2010. Disponível em: http://www.inca.gov.br/rbc/n_56/v04/pdf/11_revisao_metilacao_dna_cancer.pdf.

- HILL, Victoria K.;RICKETTS, Christopher; BIECHE, Ivan; VACHER, Sophie; GENTLE, Dean; LEWIS, Cherry; MAHER, Eamonn R., LATIF, Farida. **Genome-Wide DNA Methylation Profiling of CpG Islands in Breast Cancer Identifies Novel Genes Associated with Tumorigenicity,** Cancer Res April 15, 2011 71:2988-2999; Published OnlineFirst March 1, 2011;doi:10.1158/0008-5472.CAN-10-4026

- BARRETT T. **Gene Expression Omnibus (GEO)** 2013 May 19. In**:** The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013. Disponível em: http://www.ncbi.nlm.nih.gov/books/NBK159736/

- INCA. **O que é câncer?** 2016. Disponível: http://www1.inca.gov.br/conteudo_view.asp?id=322

- CHAPMAN, Pete; CLINTON, Julian; KERBER, Randy; KHABAZA, Thomas; REINARTZ, Thomas; SHEARER, Colin; WIRTH, Rüdiger. **CRISP-DM 1.0, Step-by-step data mining guide**. 2000. Disponível em: https://www.the-modeling-agency.com/crisp-dm.pdf.

- ARAÚJO, Cidália et al. **Estudo de Caso. Métodos de Investigação em Educação. Instituto de Educação e Psicologia**, Universidade do Minho, 2008. Disponível em:http://grupo4te.com.sapo.pt/estudo_caso.pdf

- AMO, Sandra de. **Técnicas de Mineração de dados**. 2004: Universidade Federal de

Uberlândia. Disponível em: http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf.

- CASTANHEIRA, Luciana Gomes. **Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões. Belo Horizonte**, 2008. 95 p. Dissertação (Mestrado em Engenharia Elétrica) – Programa de Pós-Graduação em Engenharia Elétrica - PPGEE, Universidade Federal de Minas Gerais, 2008. Disponível em: http://www.ppgee.ufmg.br/defesas/349M.PDF

- FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. ***From Data Mining to Knowledge Discovery in Databases***. 1996. Disponível em http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230. Acesso em 15 de Jul. 2015.

- MSDN, Portal. **Algoritmo Árvores de Decisão da Microsoft**, 2016. Disponível em: https://msdn.microsoft.com/pt-br/library/ms175312(v=sql.120).aspx

- Heckerman D., Geiger, D., Chickering, D.M., **Learning Bayesian Networks: The Combination of Knowledge and Statistical Data,** in *Machine Learning*, vol. 20, no. MSR-TR-94-09, pp. 197-243, Morgan Kaufmann Publishers [March 1995]. Disponível em: http://research.microsoft.com/apps/pubs/default.aspx?id=65088

- MSDN, Portal. **Referência técnica do algoritmo Árvores de Decisão da Microsoft,** 2016. Disponível em: https://msdn.microsoft.com/pt-br/library/cc645868(v=sql.120).aspx

ANEXOS

Script SQL para criação do banco de dados

```
CREATE DATABASE BD_METILACAO_CANCER

USE BD_METILACAO_CANCER

CREATE TABLE PESSOA(
	CD_PESSOA INTEGER IDENTITY (1,1) NOT NULL,
	CD_STATUS_MUTACAO INTEGER NOT NULL,
	IDADE INTEGER NOT NULL,
	TIPO_TECIDO VARCHAR (30) NOT NULL,
	SEXO VARCHAR NOT NULL,
	PRIMARY KEY (CD_PESSOA)
)

CREATE TABLE STATUS_MUTACAO(
	CD_STATUS_MUTACAO INTEGER NOT NULL,
	TIPO_STATUS_MUTACAO VARCHAR(40) NOT NULL,
	PRIMARY KEY (CD_STATUS_MUTACAO)
)


CREATE TABLE REGIAO_CPG(
	CD_CPG INTEGER IDENTITY (1,1) NOT NULL,
	NOME_REGIAO_CPG VARCHAR(10) NOT NULL,
	GENE_ID VARCHAR(20) NOT NULL,
	CROMOSSOMO VARCHAR(10) NOT NULL,
	ILHA_CPG BIT NOT NULL,
	SIMBOLO_GENE VARCHAR (10) NULL,
	PRIMARY KEY (CD_CPG)
)


CREATE TABLE PESSOA_REGIAO_CPG(
	CD_PESSOA INTEGER NOT NULL,
	CD_CPG INTEGER NOT NULL,
	VALOR FLOAT NULL,
	PRIMARY KEY (CD_PESSOA, CD_CPG)
)

ALTER TABLE REGIAO_CPG ADD CONSTRAINT CK_ILHA_CPG CHECK (ILHA_CPG IN (0,1))

ALTER TABLE REGIAO_CPG DROP CONSTRAINT CK_ILHA_CPG

ALTER TABLE PESSOA ADD CONSTRAINT FK_STATUS_MUTACAO FOREIGN KEY (CD_STATUS_MUTACAO)
REFERENCES STATUS_MUTACAO(CD_STATUS_MUTACAO)

ALTER TABLE PESSOA_REGIAO_CPG ADD CONSTRAINT FK_CD_PESSOA FOREIGN KEY (CD_PESSOA)
REFERENCES PESSOA(CD_PESSOA)

ALTER TABLE PESSOA_REGIAO_CPG ADD CONSTRAINT FK_CD_CPG FOREIGN KEY (CD_CPG)
REFERENCES REGIAO_CPG(CD_CPG)

ALTER TABLE PESSOA ADD COD_ACESSO VARCHAR(15) NOT NULL

-----Tabela temporária

create table TEMP_HUMANMETHYLATION(
	NOME VARCHAR(10),
```

```
	GENE_ID VARCHAR(20),
	CROMOSSOMO VARCHAR (10),
	ILHA_CPG_VARCHAR VARCHAR(10),
	ILHA_CPG BIT,
	SIMBOLO_GENE VARCHAR(10)
)

---Inserts nas tabelas

INSERT INTO REGIAO_CPG(NOME_REGIAO_CPG, GENE_ID, CROMOSSOMO, ILHA_CPG, SIMBOLO_GENE)
SELECT NOME, GENE_ID, CROMOSSOMO, ILHA_CPG, SIMBOLO_GENE FROM TEMP_HUMANMETHYLATION
ORDER BY NOME

INSERT INTO STATUS_MUTACAO VALUES (0,'BRCA1 sem mutação e saudável')
INSERT INTO STATUS_MUTACAO VALUES (1,'BRCA1 mutante e saudável')
INSERT INTO STATUS_MUTACAO VALUES (2,'BRCA1 mutante com câncer de mama')

INSERT INTO PESSOA VALUES(2,59, 'Sangue', 'F', 'GSM1378474')
INSERT INTO PESSOA VALUES(2,47, 'Sangue', 'F', 'GSM1378475')
INSERT INTO PESSOA VALUES(0,35, 'Sangue', 'F', 'GSM1378476')
INSERT INTO PESSOA VALUES(0,41, 'Sangue', 'F', 'GSM1378477')
INSERT INTO PESSOA VALUES(2,46, 'Sangue', 'F', 'GSM1378478')
INSERT INTO PESSOA VALUES(2,63, 'Sangue', 'F', 'GSM1378479')
INSERT INTO PESSOA VALUES(0,53, 'Sangue', 'F', 'GSM1378480')
INSERT INTO PESSOA VALUES(0,46, 'Sangue', 'F', 'GSM1378481')
INSERT INTO PESSOA VALUES(2,33, 'Sangue', 'F', 'GSM1378482')
INSERT INTO PESSOA VALUES(0,51, 'Sangue', 'F', 'GSM1378483')
INSERT INTO PESSOA VALUES(0,51, 'Sangue', 'F', 'GSM1378484')
INSERT INTO PESSOA VALUES(1,24, 'Sangue', 'F', 'GSM1378485')
INSERT INTO PESSOA VALUES(2,57, 'Sangue', 'F', 'GSM1378486')
INSERT INTO PESSOA VALUES(0,26, 'Sangue', 'F', 'GSM1378487')
INSERT INTO PESSOA VALUES(0,71, 'Sangue', 'F', 'GSM1378488')
INSERT INTO PESSOA VALUES(0,62, 'Sangue', 'F', 'GSM1378489')
INSERT INTO PESSOA VALUES(2,29, 'Sangue', 'F', 'GSM1378490')
INSERT INTO PESSOA VALUES(1,24, 'Sangue', 'F', 'GSM1378491')
INSERT INTO PESSOA VALUES(2,34, 'Sangue', 'F', 'GSM1378492')
INSERT INTO PESSOA VALUES(2,63, 'Sangue', 'F', 'GSM1378493')
INSERT INTO PESSOA VALUES(0,60, 'Sangue', 'F', 'GSM1378494')
INSERT INTO PESSOA VALUES(0,43, 'Sangue', 'F', 'GSM1378495')
INSERT INTO PESSOA VALUES(2,70, 'Sangue', 'F', 'GSM1378496')
INSERT INTO PESSOA VALUES(0,42, 'Sangue', 'F', 'GSM1378497')
INSERT INTO PESSOA VALUES(2,35, 'Sangue', 'F', 'GSM1378498')
INSERT INTO PESSOA VALUES(0,46, 'Sangue', 'F', 'GSM1378499')
INSERT INTO PESSOA VALUES(2,52, 'Sangue', 'F', 'GSM1378500')
INSERT INTO PESSOA VALUES(2,33, 'Sangue', 'F', 'GSM1378501')
INSERT INTO PESSOA VALUES(0,34, 'Sangue', 'F', 'GSM1378502')
INSERT INTO PESSOA VALUES(0,44, 'Sangue', 'F', 'GSM1378503')
INSERT INTO PESSOA VALUES(1,28, 'Sangue', 'F', 'GSM1378504')
INSERT INTO PESSOA VALUES(2,44, 'Sangue', 'F', 'GSM1378505')
INSERT INTO PESSOA VALUES(2,39, 'Sangue', 'F', 'GSM1378506')
INSERT INTO PESSOA VALUES(0,41, 'Sangue', 'F', 'GSM1378507')
INSERT INTO PESSOA VALUES(0,52, 'Sangue', 'F', 'GSM1378508')
INSERT INTO PESSOA VALUES(0,46, 'Sangue', 'F', 'GSM1378509')
INSERT INTO PESSOA VALUES(2,51, 'Sangue', 'F', 'GSM1378510')
INSERT INTO PESSOA VALUES(0,29, 'Sangue', 'F', 'GSM1378511')
INSERT INTO PESSOA VALUES(2,57, 'Sangue', 'F', 'GSM1378512')
INSERT INTO PESSOA VALUES(2,33, 'Sangue', 'F', 'GSM1378513')
INSERT INTO PESSOA VALUES(1,18, 'Sangue', 'F', 'GSM1378514')
INSERT INTO PESSOA VALUES(0,67, 'Sangue', 'F', 'GSM1378515')
INSERT INTO PESSOA VALUES(0,31, 'Sangue', 'F', 'GSM1378516')
INSERT INTO PESSOA VALUES(2,61, 'Sangue', 'F', 'GSM1378517')
```

```
INSERT INTO PESSOA VALUES(0,24, 'Sangue', 'F', 'GSM1378518')
INSERT INTO PESSOA VALUES(0,32, 'Sangue', 'F', 'GSM1378519')
INSERT INTO PESSOA VALUES(2,44, 'Sangue', 'F', 'GSM1378520')
INSERT INTO PESSOA VALUES(0,43, 'Sangue', 'F', 'GSM1378521')
INSERT INTO PESSOA VALUES(2,47, 'Sangue', 'F', 'GSM1378522')
INSERT INTO PESSOA VALUES(2,62, 'Sangue', 'F', 'GSM1378523')
INSERT INTO PESSOA VALUES(0,25, 'Sangue', 'F', 'GSM1378524')
INSERT INTO PESSOA VALUES(0,60, 'Sangue', 'F', 'GSM1378525')
INSERT INTO PESSOA VALUES(0,29, 'Sangue', 'F', 'GSM1378526')
INSERT INTO PESSOA VALUES(2,49, 'Sangue', 'F', 'GSM1378527')
INSERT INTO PESSOA VALUES(1,63, 'Sangue', 'F', 'GSM1378528')
INSERT INTO PESSOA VALUES(0,34, 'Sangue', 'F', 'GSM1378529')
INSERT INTO PESSOA VALUES(2,58, 'Sangue', 'F', 'GSM1378530')
INSERT INTO PESSOA VALUES(2,24, 'Sangue', 'F', 'GSM1378531')
INSERT INTO PESSOA VALUES(0,42, 'Sangue', 'F', 'GSM1378532')
INSERT INTO PESSOA VALUES(0,69, 'Sangue', 'F', 'GSM1378533')
INSERT INTO PESSOA VALUES(0,45, 'Sangue', 'F', 'GSM1378534')
INSERT INTO PESSOA VALUES(2,45, 'Sangue', 'F', 'GSM1378535')
INSERT INTO PESSOA VALUES(2,54, 'Sangue', 'F', 'GSM1378536')
INSERT INTO PESSOA VALUES(0,48, 'Sangue', 'F', 'GSM1378537')
INSERT INTO PESSOA VALUES(0,34, 'Sangue', 'F', 'GSM1378538')
INSERT INTO PESSOA VALUES(0,60, 'Sangue', 'F', 'GSM1378539')
INSERT INTO PESSOA VALUES(2,41, 'Sangue', 'F', 'GSM1378540')
INSERT INTO PESSOA VALUES(2,50, 'Sangue', 'F', 'GSM1378541')
INSERT INTO PESSOA VALUES(0,19, 'Sangue', 'F', 'GSM1378542')
INSERT INTO PESSOA VALUES(0,51, 'Sangue', 'F', 'GSM1378543')
INSERT INTO PESSOA VALUES(1,26, 'Sangue', 'F', 'GSM1378544')
INSERT INTO PESSOA VALUES(2,35, 'Sangue', 'F', 'GSM1378545')
INSERT INTO PESSOA VALUES(2,32, 'Sangue', 'F', 'GSM1378546')
INSERT INTO PESSOA VALUES(2,29, 'Sangue', 'F', 'GSM1378547')
INSERT INTO PESSOA VALUES(0,30, 'Sangue', 'F', 'GSM1378548')
INSERT INTO PESSOA VALUES(0,23, 'Sangue', 'F', 'GSM1378549')
INSERT INTO PESSOA VALUES(0,49, 'Sangue', 'F', 'GSM1378550')
INSERT INTO PESSOA VALUES(2,68, 'Sangue', 'F', 'GSM1378551')
INSERT INTO PESSOA VALUES(1,30, 'Sangue', 'F', 'GSM1378552')
INSERT INTO PESSOA VALUES(0,62, 'Sangue', 'F', 'GSM1378553')
INSERT INTO PESSOA VALUES(2,47, 'Sangue', 'F', 'GSM1378554')
INSERT INTO PESSOA VALUES(2,32, 'Sangue', 'F', 'GSM1378555')
INSERT INTO PESSOA VALUES(0,51, 'Sangue', 'F', 'GSM1378556')
INSERT INTO PESSOA VALUES(0,29, 'Sangue', 'F', 'GSM1378557')

--- Inserções na tabela pessoa_regiao_cpg
insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 1, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378474 G
on (G.ID_REF = C.NOME_REGIAO_CPG)

insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 2, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378475 G
on (G.ID_REF = C.NOME_REGIAO_CPG)

insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 3, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378476 G
on (G.ID_REF = C.NOME_REGIAO_CPG)

insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 4, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378477 G
on (G.ID_REF = C.NOME_REGIAO_CPG)
```

```
insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 5, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378478 G
on (G.ID_REF = C.NOME_REGIAO_CPG)

insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 6, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378479 G
on (G.ID_REF = C.NOME_REGIAO_CPG)

insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 7, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378480 G
on (G.ID_REF = C.NOME_REGIAO_CPG)

insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 8, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378481 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 9, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378482 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 10, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378483 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 11, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378484 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 12, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378485 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 13, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378486 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 14, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378487 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 15, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378488 G
```

```
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 16, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378489 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 17, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378490 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 18, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378491 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 19, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378492 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 20, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378493 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 21, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378494 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 22, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378495 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 23, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378496 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 24, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378497 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 25, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378498 G
on (G.ID_REF = C.NOME_REGIAO_CPG)
```

```
insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 26, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378499 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 27, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378500 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 28, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378501 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 29, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378502 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 30, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378503 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 31, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378504 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 32, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378505 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 33, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378506 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 34, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378507 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 35, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378508 G
on (G.ID_REF = C.NOME_REGIAO_CPG)
```

```
insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 36, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378509 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 37, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378510 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 38, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378511 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 39, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378512 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 40, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378513 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 41, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378514 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 42, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378515 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 43, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378516 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 44, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378517 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 45, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378518 G
on (G.ID_REF = C.NOME_REGIAO_CPG)
```

```
insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 46, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378519 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 47, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378520 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 48, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378521 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 49, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378522 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 50, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378523 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 51, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378524 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 52, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378525 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 53, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378526 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 54, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378527 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 55, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378528 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
```

```
select 56, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378529 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 57, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378530 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 58, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378531 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 59, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378532 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 60, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378533 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 61, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378534 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 62, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378535 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 63, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378536 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 64, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378537 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 65, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378538 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 66, C.cd_cpg,G.VALUE from REGIAO_CPG C
```

```
inner join GSM1378539 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 67, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378540 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 68, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378541 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 69, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378542 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 70, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378543 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 71, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378544 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 72, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378545 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 73, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378546 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 74, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378547 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 75, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378548 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 76, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378549 G
```

```
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 77, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378550 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 78, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378551 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 79, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378552 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 80, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378553 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 81, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378554 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 82, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378555 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 83, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378556 G
on (G.ID_REF = C.NOME_REGIAO_CPG)


insert into PESSOA_REGIAO_CPG (CD_PESSOA, CD_CPG, VALOR)
select 84, C.cd_cpg,G.VALUE from REGIAO_CPG C
inner join GSM1378557 G
on (G.ID_REF = C.NOME_REGIAO_CPG)

----Visão utilizada para análise
CREATE VIEW VW_PESSOA_REGIAO_CPG_DETALHADA
AS
SELECT ROW_NUMBER() OVER(ORDER BY P.CD_PESSOA) AS ID_ARTIFICIAL,
P.CD_PESSOA,
 RC.NOME_REGIAO_CPG,
 PRP.VALOR, RC.ILHA_CPG, RC.CROMOSSOMO, RC.SIMBOLO_GENE,
P.IDADE,
SM.TIPO_STATUS_MUTACAO
FROM PESSOA_REGIAO_CPG PRP
JOIN PESSOA P
```

```
ON (P.CD_PESSOA = PRP.CD_PESSOA)
JOIN STATUS_MUTACAO SM
ON(SM.CD_STATUS_MUTACAO = P.CD_STATUS_MUTACAO)
JOIN REGIAO_CPG RC
ON (PRP.CD_CPG = RC.CD_CPG)
WHERE VALOR IS NOT NULL
```