Universidade do Estado do Rio de Janeiro
Instituto de Matemática e Estatística
Departamento de Informática e Ciência da Computação

ALGORITMOS GENÉTICOS E REDES DE KOHONEN NA COMPLEMENTAÇÃO DE DADOS AUSENTES

Autor(es): **Juliana Vasconcellos de Medina**

RIO DE JANEIRO
10/2012

ALGORITMOS GENÉTICOS E
REDES DE KOHONEN
NA COMPLEMENTAÇÃO DE DADOS AUSENTES

Juliana Vasconcellos de Medina

Monografia submetida ao corpo docente do Instituto de Matemática e Estatística da Universidade do Estado do Rio de Janeiro - UERJ, como parte dos requisitos necessários à obtenção do grau de Bacharel em Informática e Tecnologia da Informação.

Banca Examinadora:

_____
Prof. Jorge de Abreu Soares - Orientador
FAF/UERJ

_____
Prof. Rosa Maria Esteves Moreira Costa
IME/UERJ

_____
Prof. Ronaldo Ribeiro Goldschmidt
UFRRJ

_____
Prof. Isabel Fernandes de Souza
UFRRJ

Rio de Janeiro/RJ
30 de outubro de 2012.

A toda a minha família e amigos.

AGRADECIMENTOS

Agradeço a Deus a oportunidade de ter estudado na UERJ, onde pude aprender tanto, fazer amigos e ter um ótimo futuro. E especialmente pelos meus pais que tanto me incentivaram a não desistir ou desanimar nessa caminhada de estudos e trabalhos.

Também agradeço o total apoio que o professor Jorge de Abreu Soares me forneceu durante todo esse trabalho final de curso e também os momentos de esclarecimento que o Rafael Castaneda e a Cláudia Ferlin me proporcionaram retirando dúvidas, ajudando a retirar aquelas pedras que vira e volta aparecem no caminho.

Um especial agradecimento aos meus queridos amigos que me amparam em todos os momentos de dificuldades, na saúde e na doença, nas alegrias e nas tristezas...

Certamente esse trabalho está sendo uma grande alegria graças a todo vocês.

RESUMO

Os bancos de dados são muitos utilizados nos tempos de hoje, a quantidade de sistemas crescem cada vez mais devido a necessidade de controlar e gerenciar grande quantidade de informações. Mas e se essas informações estiverem inconsistentes? E se por causa de uma carga de dados ou pela entrada de informações num sistema algum dado for perdido? Como confiar nessas informações? A descoberta de conhecimento feito numa base de dados assim não seria nada confiável.

Muitos trabalhos já abordaram esse problema mostrando e demonstrando que a melhor solução é a complementação de dados ausentes, porém a maioria aborda a imputação simples. Este trabalho contribui ao aprimoramento da complementação de dados por desenvolver técnicas de seleção com *Algoritmos Genéticos* e agrupamento com *Redes de Kohonen* na aplicação da imputação composta além de comparar os resultados com a imputação composta (SOARES, 2007) pelas técnicas de seleção com *Análise de Componentes Principais* (*PCA - Principal Component Analysis)* e agrupamento com *K-Means*.

O objetivo maior desse trabalho é analisar a qualidade da imputação composta por seleção com *Algoritmos Genéticos* e agrupamento com *Redes de Kohonen*, então inicialmente há uma parte teórica sobre a Complementação de dados e algumas as técnicas de imputação, seleção e agrupamento. Depois, descreve-se alguns trabalhos relacionados e desenvolve-se o trabalho proposto descrevendo como foi implementada a imputação composta. Em seguida a análise dos resultados e por fim as conclusões, contribuições e ideia de trabalhos futuros.

SUMÁRIO

Lista de Figuras

Lista de Tabelas

Lista de Siglas e Símbolos

AG	Algoritmos Genéticos
DM	*Data Marts*
DW	Data Warehouse
ETL	*Extract – Transformation – Load (Extrai – Transforma – Carrega)*
IA	Inteligência Artificial
JOONE	*Java Object Oriented Neural Engine*
KDD	Knowledge Discovery in Databases (Descoberta de Conhecimento em Base de Dados)
K-NN	*K Vizinhos Mais Próximos*
OLAP	Online Analytical processing (Processo analítico on-line)
PCA	*Principal Component Analysis (Análise de Componentes Principais)*
SGBD	Sistema Gerenciador de Banco de Dados

CAPÍTULO 1
INTRODUÇÃO

1.1 Justificativa e Posicionamento

O grande e constante avanço das tecnologias computacionais tem possibilitado cada vez mais um armazenamento maior de dados que estão sendo bem aproveitados pelas indústrias, centros de pesquisas, empresas e organizações em geral. Isso acontece devido à queda do custo de hardwares com maior capacidade e velocidade atualmente. Por exemplo, é possível comprar computadores domésticos com capacidade de armazenamento secundário de dois terabytes atualmente. Já as empresas conseguem facilmente computadores de vinte e quatro terabytes segundo sites de compra de computadores empresariais. Além disso, os processadores são cada vez mais rápidos facilitando dessa forma o armazenamento de muitos dados e a utilização desses dados para geração de informação e conhecimento.

Um interessante campo de estudo que busca esta geração de conhecimento a partir de um grande volume de dados é conhecido como *Descoberta de Conhecimento em Bases de dados* (*KDD – Knowledge Discovery in Databases*) (FAYYAD et al, 1996). O *KDD* busca descobrir de forma semiautomática padrões entre os conjuntos de dados usando conceitos de inteligência artificial (*AI - Artificial Intelligence*), estatística, aprendizado de máquina, armazenamento de dados (*Data Warehousing*), banco de dados, reconhecimento de padrões e visualização de dados. Esses padrões ajudam a explicitar o conhecimento a partir dos dados, tendo inúmeras aplicações práticas, tais como definir comportamentos dos consumidores, identificar tendências existentes em negócios, apoiar decisões, prever cenários possíveis, criar teorias entre outros. Por esta razão observamos que o tema tornou-se objeto de estudo intenso, revelado em diversos trabalhos.

Podemos dividir o processo de *KDD* em três fases: i) **preparação dos dados** ou **pré-processamento de mineração de dados**, ii) **processamento de mineração de dados** e iii) **pós-processamento** ou **visualização de padrões**. Infelizmente pouca atenção se dá para a primeira fase, já que esta etapa exige um detalhado processo de limpeza e correção de dados que toma considerável parte do tempo destinado ao projeto. Todavia, é importante salientar que dados que apresentem alto nível de ruído fomentam a geração de padrões e conseqüentes análises equivocados.

O *Pré-processamento de Mineração de Dados* prepara o conjunto de dados para que a mineração propriamente dita consiga gerar resultados coerentes. Há duas formas de preparar os dados: a forma simplista que retira os registros de dados ausentes ou sujos – que, apesar de ser uma opção a ser considerada, deve sempre que possível ser evitada; e a maneira mais cuidadosa, onde se busca, através do uso de diversas técnicas, a reparação dos dados contidos em uma base de dados. Dentre as tarefas dessa primeira fase, uma importante missão é a tentativa de "adivinhação" de valores ausentes em campos de registros. Chamamos *Complementação de Dados* ou *Imputação* a etapa onde estimamos valores substitutivos a outros ausentes nos atributos de tuplas de uma tabela. Para atingir este objetivo, variadas técnicas vem sendo utilizadas, revelando interessantes resultados.

Todavia, deseja-se saber qual técnica trará melhores resultados para a imputação de valores ausentes na fase de Pré-processamento de dados. Será que a aplicação de apenas uma técnica bastará ou um conjunto delas faz-se necessário? É preciso estudar melhor para encontrar essas respostas, testando e analisando os resultados para tentar chegar a alguma indicação.

Outra importante área de estudo que se beneficia da geração de informação em grandes bases de dados são os *Data Warehouses* (armazéns de dados) e sistemas OLAP (*On Line Analytical Processing*). Neste contexto, as bases de dados analíticas são construídas de tal modo que permitam ao usuário final construir análises que permitam a confrontação de cenários não previstos anteriormente (consultas eventuais ou *ad hoc*), com viés primordialmente histórico. Essa análise torna-se mais rica se os armazéns de dados forem construídos tomando-se por base não apenas os dados transacionais de sistemas legados, mas também os obtidos de fontes externas. Porém, para que os dados de diferentes fontes sejam armazenados num *DW* é preciso garantir que obedeçam às mesmas definições e regras, entretanto essa transformação pode acarretar danos ou até perda de dados importantes para geração de conhecimento (PUC, 2012, p.20).

A figura abaixo mostra essa transformação em quatro partes:

1. Outras bases de dados ou sistemas operacionais.
2. Data Staging Area: os dados estão passando por um conjunto de processos usualmente chamados ETL (Extract – Transformation – Load) e não podem ser acessados pelo usuário nem para consulta.

3. Área de apresentação de dados: os dados estão organizados e acessíveis para usuário consultar ou aplicações que geram relatórios ou processos de análises utilizarem.
4. Ferramentas de acesso a dados.

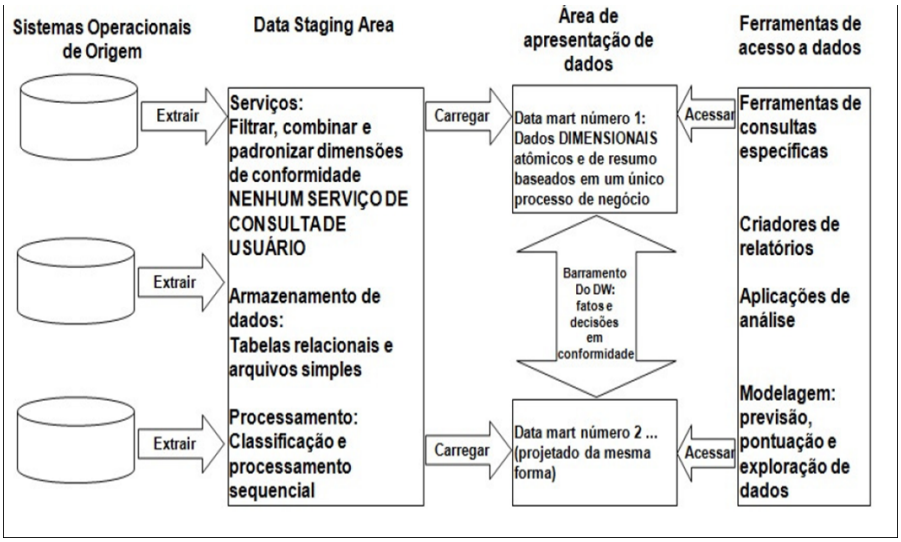


Figura 1.1 Componentes de um Data Warehouse

Fonte: KIMBALL (1998)

Durante a etapa ETL, mais especificamente na transformação dos dados, faz-se necessário um processo de preparação dos dados antes de realizar sua carga no DW. É muito interessante ver que esse preparo não só torna a base mais limpa e consistente como também fornece muita experiência aos responsáveis pelo preparo possibilitando grandes avanços nas etapas seguintes porque na preparação dos dados são feitas consultas nos DW ou DM em busca de dados desejados, consolidação de todas as informações desejadas em um único local, normalização, criação de variáveis derivadas das originais que possuam mais informação, limpeza e tratamento de erros, identificação de valores discrepantes ou ausentes e realização de ações para corrigi-los. Assim, também nessa área a preparação adequada de dados mostra-se tarefa importantíssima para a correta geração de resultados.

1.2 Motivação

O problema da pouca valorização (SOARES, 2007, p.8) da tão importante fase de *Pré-processamento de Mineração de Dados* revela-se preocupante, já que as atenções ficam mais voltadas para a fase de mineração propriamente dita. Sabendo-se que o volume de dados

armazenados atualmente pelos sistemas de informação é enorme – e só tende a crescer, a probabilidade de os dados estarem consistentes depende fundamentalmente da correta imposição de restrições de integridade nas aplicações e na base de dados, nem sempre observáveis. Ademais, dados podem ser obtidos de fontes externas, tais como planilhas ou arquivos *flat*, o que quebra a confiabilidade de garantia "certa" da consistência desses dados. Eles podem ser perdidos devido à integração de diferentes bases, carga incompleta ou mal delineada, falhas em equipamentos, problemas com fornecê-lo, dentre outros. Tudo isso pode prejudicar ou até invalidar a etapa subsequente e, por isso, há a necessidade de uma dedicada atenção no sentido de garantir uma eficiente preparação dos dados, com especial atenção à tarefa de imputação dos dados.

Esta tarefa demanda tempo e esforço mas que fornece grande benefício à descoberta de informação ou conhecimento em bases de dados: dados inconsistentes podem substituídos e dados ausentes podem ser recuperados, assemelhando-se aos dados reais e tornando o processo mais confiável. O grande desafio da imputação é produzir uma base de dados o mais próxima possível de sua realidade.

A imputação simples (TWALA, CARTWRIGHT, SHEPPERD, 2005), desenvolvida com apenas uma técnica de regressão de dados, traz bons resultados melhorando a base de dados, porém não alcança a similaridade desejada. Já a imputação composta (SOARES, 2007), através de combinações de diferentes técnicas, pode possibilitar o que a imputação simples não consegue, pois permite que os dados sejam tralhados mais de uma vez em busca do valor a ser imputado.

Efetuado uma prévia seleção dos dados, somente as tuplas mais parecidas com a tupla com atributo ausente serão utilizados para encontrar o valor a ser imputado o que teoricamente se torna melhor do que utilizar a base de dados completa. O mesmo ocorre com o agrupamento de dados que resulta em grupos de dados que apresentam uma maior similaridade. Resta apenas encontrar qual técnica de seleção de dados e de agrupamento melhoram mais a imputação dos dados. Para tal é preciso desenvolver as possíveis implementações dessas técnicas e compará-las.

1.3 Objetivos

Observando que a imputação simples traz bons resultados, mas que não temos como nos esquecer de que se trata de um trabalho de "adivinhação", o foco desse trabalho é demonstrar

o quanto a utilização de diferentes técnicas na imputação composta pode melhorar a qualidade da imputação de dados ausentes.

Visando ampliar o conhecimento e a prática de *Complementação de Dados* ausentes utilizando a imputação composta, duas técnicas foram escolhidas, implementadas, e disponibilizadas para a utilização na imputação composta: seleção com *AG - Algoritmos Genéticos* (CONDE, 2005, FREITAS, 2002, LINDEN, 2006) e agrupamento com *Redes de Kohonen* (FERLIN, 2008, KIM, LEE, YI, 2004, KOHONEN, 1984).

Para alcançar este objetivo três trabalhos serviram como motivadores principais: SOARES (2007), que define a imputação composta e a instancia com o uso de uma técnica de seleção denominada **Análise de Componentes Principais (**PCA – *Principal Component Analysis*) (SMITH, 2002, SHLENS, 2005); COSTA (2005), que utiliza algoritmos genéticos (FREITAS, 2002, LINDEN, 2006) para a seleção de variáveis no processo de KDD em bases de dados; e FERLIN (2008), que define a ***imputação em cascata***, metodologia de preenchimento de dados ausentes em várias colunas utilizando Redes Neuronais de *Kohonen* (KIM, LEE, YI, 2004, KOHONEN, 1984) no agrupamento de dados feito antes do processo de imputação, baseado no conceito de **imputação sequencial** de RIBEIRO (2008). Mantendo as mesmas estratégias de SOARES (2007), porém criando novos *planos de imputação* ao adicionar uma nova técnica de seleção (*AG*) e uma nova técnica de agrupamento (*Redes de Kohonen*), propõe-se analisar qual o reflexo essa nova instância do workflow.

1.4 Metodologia e Resultados

As técnicas de seleção com *Algoritmos Genéticos* e agrupamento com *Redes de Kohonen,* baseadas em COSTA (2005) e em FERLIN (2008), respectivamente, foram utilizadas como técnicas de seleção e agrupamento, e seguidas de técnicas de imputação implementadas por SOARES (2007), a saber: utilização do cálculo da média, do algoritmo dos *k* vizinhos mais próximos (k-NN) (MITCHELL, 1997, AHA, KIBLER, ALBERT, 1991, DASARATHY, 1990), e o uso de redes neuronais *Back Propagation* (RUMELHART *et al*, 1986). Essas combinações de técnicas, ou melhor, estratégias formaram um total de cinco:

1) imputação;
2) seleção e imputação;
3) agrupamento e imputação;
4) seleção, agrupamento e imputação;

5) agrupamento, seleção e imputação.

SOARES (2007) implementou *Análise de Componentes Principais* (*PCA – Principal Component Analysis*) (SMITH, 2002, SHLENS, 2005) como técnica de seleção de dados e *K-Means* (*K-centróides*) (MCQUEEN, 1967) como técnica de agrupamento. Já como imputação foram implementadas três técnicas: o algoritmo dos *k-vizinhos mais próximos* (*k-NN*) (MITCHELL, 1997, AHA,KIBLER, ALBERT, 1991, DASARATHY, 1990), as *Redes Neuronais* de aprendizado supervisionado *Back Propagation* (RUMELHART et al, 1986), e a *média aritmética simples*. Buscando resultados mais eficazes, combinamos estas diversas técnicas nas etapas da imputação composta, a fim de comparar a qualidade dos dados imputados pela diversificação de técnicas.

O sistema, *Appraisal*, desenvolvido por SOARES (2007) em linguagem de programação Java para executar os planos de imputação de SOARES (2007), foi estendido para executar também os planos de imputação contendo *Algoritmos Genéticos* e *Redes de Kohonen*. Não foi aproveitada a parte de Comitê implementada no *Appraisal* por SOARES (2007), por não ser o foco deste trabalho.

As mesmas bases de dados de SOARES (2007) foram utilizadas no banco de dados MySQL versão 5, bases estas do repositório de aprendizado de máquina da Universidade da Califórnia Irvine (NEWMAN et al, 1998). Esse repositório serviu e serve para diversos trabalhos de descoberta de dados. As bases escolhidas foram:

1ª *Iris Plants* com 150 registros, quatro atributos numéricos e um classificador.

2ª *Pima Indians Diabetes* com 392 registros, oito atributos numéricos e uma classe.

3ª *Wisconsin Breast Cancer* com 682 registros, nove atributos numéricos e uma classe.

A primeira base de dados registra classificações de plantas; a segunda dados médicos de pacientes com diabetes ou não; e a terceira registra características clínicas de pacientes com câncer de mama ou não.

Seguindo SOARES (2007), os valores ausentes nessas três bases completas foram gerados por um módulo do *Appraisal* de forma completamente aleatória (*MCAR – Missing Completely At Random*). Cada atributo alvo sofreu de 10% a 50% de ausência nos seus registros, sendo que o intervalo de ausência foi de 10%. Não geramos valores ausentes em mais de um atributo por não se tratar de uma imputação multivariada, como feito em FERLIN (2008).

O *Appraisal* também possui um módulo de relatórios dos resultados gerados pelas execuções das estratégias, e assim os resultados foram analisados para avaliar cada estratégia em função da base, atributo e percentual de ausência.

1.5 Organização do texto

A organização foi feita da seguinte forma: no Capítulo 2, aborda-se a complementação de dados ausentes. No Capítulo 3, encontra-se uma descrição dos trabalhos relacionados e que serviram de base para a proposta deste trabalho. No Capítulo 4 desenvolve-se o que está sendo proposto definindo as bases de dados, técnicas e configurações utilizadas. Já no Capítulo 5 há uma análise dos resultados obtidos nos testes. E por fim, no Capítulo 6, tece-se considerações finais avaliando os resultados obtidos nesse trabalho e citando ideias para trabalhos futuros.

CAPÍTULO 2
COMPLEMENTAÇÃO DE DADOS AUSENTES

Neste capítulo abordaremos os conceitos de imputação de dados ausentes, bem como apresentaremos os trabalhos disponíveis na literatura que tratam sobre o tema.

2.1 Abordagens de Imputação

Segundo SCHÖNER (2004) e GOLDSCHMIDT e PASSOS (2005), citado por RIBEIRO (2007, p.21), "imputação é qualquer procedimento automático ou semiautomático, capaz de preencher valores ausentes encontrados em bases de dados", ou seja, complementar os dados ausentes auxiliado por alguma técnica computacional. Porém é importante atentar que essa atribuição precisa ser feita com um valor mais próximo possível ao valor original, ou ao menos com uma baixa taxa de erro, a fim de que estes valores gerados não gerem uma grande perturbação futura quando de seu uso. Aplicações comuns da técnica de imputação são o processo de Extração, Transformação e Carga (ETL) em Data Warehouses, na fase de *Pré-processamento de Mineração de Dados* em descoberta de conhecimento (*KDD*), na "limpeza" e consistência das informações nas áreas de estatística (RUBIN, 1988), medicina/genética (CELTON et al, 2010), processos industriais (LAKSHMINARAYAN, HARP, SAMAD, 1999), problemas de Engenharia de Software (SONG, SHEPPERD, CARTWRIGHT, 2005), entre outros.

Um processo importante na complementação de dados ausentes é entender qual o tipo de ausência os dados sofreram pois que o sucesso da imputação irá depender disso. Alguns autores já mapearam alguns tipos de ausência, WAYMAN (2003) citado por GRAHAM e DONALDSON (1993) classificaram esses tipos em mecanismos de dados em **acessíveis** e **inacessíveis**, no primeiro a causa da ausência pode ser explicada (MCAR e MAR) e no segundo a causa é desconhecida (NMAR e outros tipos ainda não encontrados).

No mecanismo de **ausência completamente aleatória** (*MCAR – Missing Completely At Random*) não sabemos qual motivo da perda do dado. Já a **ausência aleatória** (MAR – *Missing At Random*) é em função a um valor de outro atributo. E a **ausência não aleatória** (*NMAR – Not Missing At Random*, ou IM – *Ignorable Missing*) ocorre quando um atributo depende de outro atributo que está ausente (LITTLE, RUBIN,1987). Ainda há um

quarto mecanismo de ausência: **valores fora da faixa esperada para um atributo (**BROWN, KROSS, 2003a, 2003b, SCHAFER, GRAHAM, 2002).

Existem diversas propostas para tratar valores ausentes, a de MAGNANI (2004) a principio parece ser a mais completa, pois define três tipos de remoção de dados ausentes (registro, em pares, colunas), três tipos de imputação (global baseada no atributo ausente, global baseada nos atributos não ausentes e local), estimativa de parâmetros e gerenciamento dos dados ausentes.

Entretanto, a classificação de MAGNANI (2004) é restrita a métodos estatísticos e por isso se baseia na estimativa de parâmetros. Por isso SOARES (2007) complementou muito bem essa classificação de tratamento de dados ausentes da seguinte forma:

1. Métodos convencionais
    a. Remoção dos registros com dados ausentes
    b. Remoção em pares de dados ausentes
    c. Remoção de colunas com dados ausentes
2. Imputação
    a. Imputação Global Baseada no Atributo Ausente
    b. Imputação Global Baseada nos Atributos não Ausentes
    c. Imputação local
3. Modelagem de dados
    a. Métodos de Verossimilhança
    b. Modelos Bayesianos
4. Gerenciamento direto dos dados ausentes
5. Métodos Híbridos
    a. Imputação Múltipla
    b. Imputação Composta

2.2 Trabalhos Relacionados

BATISTA e MONARD (2003) propõem formas de tratar dados ausentes: ignorar e descartar dados sem ausência completamente aleatória, analisar se os atributos são suficientemente relevantes para não se descartar registros com muitos dados ausentes, estimar parâmetros a partir dos atributos vizinhos para completar os dados ausentes e imputação dos valores estimados.

Já MYRTVEIT, STENSRUD e OLSSON (2001) definem o descarte de registros incompletos, as técnicas baseadas em imputação, as técnicas baseadas em designações de pesos e as técnicas baseadas em modelos para tratar dados ausentes.

TWALA, CARTWRINGHT e SHEPPERD (2005) propõem realizar análise dos dados completos, imputação e procedimentos baseados em modelos.

HRUSCHKA, HRUSCHKA JR. E EBECKEN (2003a, 2003b) desenvolvem mais métodos para tratamento de dados ausentes: ignorar os registros com valores ausentes, preencher manualmente, substituir o valor ausente por uma constante, usar a média ou moda e atribuir o valor mais provável.

TSENG, WANG, LEE (2003) abordam apenas duas formas de complementação de dados: baseada em imputação para dados numéricos e baseada em mineração de dados para dados categóricos.

Similarmente a fase de *Processamento de Mineração de Dados*, a *Complementação de Dados Ausentes* pode fazer uso das mesmas técnicas. Estas técnicas estão divididas em duas categorias: *Descriptive data mining* e *Predictive data mining* **(**FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1996**)**. A primeira descreve os dados resumidamente mostrando as propriedades importantes e a segunda busca prever o comportamento de novos conjuntos de dados a partir dos dados existentes. Algumas técnicas estão listadas a seguir:

1. Sumarização (identificar características comuns entre os dados)
2. Associação (encontrar itens que ocorram simultaneamente)
3. Classificação (descobrir função que mapeia dados em classes)
4. Regressão (descobrir função que mapeia dados em valores)
5. Seleção (selecionar dados com características comuns)
6. Agrupamento (agrupar dados com características comuns)
7. Detecção de Desvios (identificar dados fora das características)
8. Detecção de sequencias (buscar itens frequentes)

Por isso, quando a fase de Pré-processamento de Mineração de dados não é tratada de forma simplista, são aplicadas técnicas de regressão de dados, precedidas ou não por outras técnicas como de seleção e/ou agrupamento, para que os dados ausentes sejam imputados com menor erro possível.

Uma das técnicas de regressão é a *Média* ou a *Moda*, que é uma forma simplista e mais comumente utilizada, entretanto também temos os *K-Vizinhos Mais Próximos* (*K-Nearest Neighbors*) (MITCHELL, 1997, AHA, KIBLER, ALBERT, 1991, DASARATHY,

1990) e o *Back Propagation* (RUMELHART et al, 1986), além de outras técnicas que não iremos tratar nesse trabalho.

E como já foi ressaltado anteriormente, é preciso buscar a técnica mais adequada, a que gera resultados mais eficientes, então pode ser que a combinação de tal ou qual técnica de regressão traga melhores resultados se for precedida de uma técnica de seleção como os *AGs* (*Algoritmos Genéticos*) (LINDEN, 2006) ou *PCA* (*Análise de Componentes Principais*) (SMITH, 2002, SHLENS, 2005), entre outras técnicas conhecidas que não fazem parte desse trabalho. Ou pode ser que uma das técnicas de regressão seja mais eficaz quando precedida por uma técnica de agrupamento como *K-Means* (MCQUEEN, 1967) ou *Redes Kohonen* (KOHONEN, 1984), entre outras técnicas que também não serão abordadas aqui.

2.3 Técnicas de Regressão

A regressão é uma das técnicas mais usadas na *Mineração de Dados*, e especificamente no pré-processamento de dados. Ela busca explicar uma ou várias variáveis de interesse em função de outras variáveis.

Os algoritmos de regressão, lineares ou não lineares, são utilizados em toda área do conhecimento, como por exemplo, computação, engenharias, biologia, medicina, agronomia, administração, sociologia, entre outros. (OGLIARI *et al*, 2003, p.3)

2.3.1 Média ou Moda

O resultado da *Média* será sempre o valor que mais aparece entre todos os registros do banco de dados. A *Moda* segue o mesmo raciocínio, a diferença é que esta trata de atributos categóricos. Ambas então possuem uma grande vantagem e ao mesmo tempo uma grande desvantagem: o rápido desempenho e se basear na tendência de valores preenchidos.

Essa desvantagem afeta consideravelmente o resultado da imputação que em grande parte das vezes não tem relação com o valor que foi mais preenchido. Segundo SOARES (2007, p.50) “algumas alternativas que podem ser adotadas para evitar que a média não distorça excessivamente os dados são a aplicação de uma perturbação aleatória ao valor da média. Esta perturbação pode adicionar ou remover um valor Δm da média, tentando reduzir os efeitos descritos anteriormente.”

A perturbação pode ser uma alternativa sim, mas também é fundamental escolher a média mais adequada ao problema em questão:

a) Média Geométrica (somente quando todos os dados são positivos)

$$\bar{x} = \sqrt[n]{x_1 * x_2 * \ldots * x_n} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

b) Media Aritmética ponderada (w_i é o peso de cada componente x_i)

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i * x_i}{\sum_{i=1}^{n} w_i}$$

c) Média Harmônica (capacidade de um elemento substituir cada n)

$$H = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n} \right)} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}$$

Além de outras médias como Minkowski, Lehmer e a Generalizada Phillips que são combinações das médias aqui citadas.

Algoritmo de Média ou Moda

Dado um ou mais registros com atributo ausente e um conjunto de registros com o mesmo atributo completo, o algoritmo aplica a função do tipo de média ou moda escolhido e retorna o valor a ser imputado. A imputação é feita nesse(s) registro(s) com o valor retornado.

2.3.2 k-NN - K Vizinhos Mais Próximos

O resultado do *K Vizinhos Mais Próximos* (MITCHELL, 1997, AHA, KIBLER, ALBERT, 1991, DASARATHY, 1990) será os registros mais próximos ao registro com dados ausentes em questão, e este resultado é gerado através de uma função de similaridade que pode ser, por exemplo, Euclidiana ou Manhattan.

a) **Euclidiana**: é a mais comumente aplicada, ela calcula a distância entre os registros ti e tj da seguinte forma, onde k é o número de colunas da tabela:

$$d(a,b) = \sqrt[2]{\sum_{i=1}^{n} (a_i - b_i)^2} = \|a - b\|_2$$

b) **Manhattan**: é uma variação da distância Euclidiana, onde o resultado é a soma da diferença absoluta entre um registro a e b.

$$d(a,b) = \sum_{i=1}^{n} |a_i - b_i| = \|a - b\|_1$$

c) **Malahanobis**: se baseia nas relações entre os registros podendo ser um registro conhecido e outro desconhecido. O conjunto a ser comparado deve ter o mesmo número de colunas mas não necessariamente o mesmo número de registros.

$$d_{Ma}(a,b) = \sqrt{(a-b)^T Cov^{-1}(a-b)} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_{ij}(a_i - b_i)(a_i - b_i)}$$

Apesar de gerar melhores resultados ao comprar conceitualmente com a *Média* ou a *Moda*, o algoritmo dos *K Vizinhos Mais Próximos* também apresenta um problema: o de ser tornarem desproporcionais se alguns registros possuírem valores muito maiores dos demais. SOARES (2007, p.45) sugere uma idéia de solução que seria a normalização dos valores dos atributos, dividindo cada um pelo maior valor da tabela ou o maior valor daquele atributo, mas o próprio SOARES conclui que essa solução pode ser prejudicada quando existe uma forte correlação linear entre os atributos dos objetos. SOARES (2007) apresenta um comparativo desses algoritmos:

1) Vantagens:
    a. É robusto, mesmo em dados com ruídos;
    b. Possui boa execução até em tabelas com muitos registros;
    c. Serve tanto para atributos numéricos quanto para categóricos;
    d. Não precisa especificações diferenciadas para cada atributo.
2) Desvantagens:
    a. Alta infra-estrutura devido a necessidade de calcular a distância de cada registro com outro registro;
    b. Encontrar o melhor valor para o k para ser complicado;
    c. O tipo de distância escolhido pode afetar o desempenho ao comparar com outro tipo de distância;
    d. Não se sabe se é melhor usar todos os atributos para calcular a distância.

Algoritmo dos K Vizinhos Mais Próximos

Dado um registro xi com atributo t ausente e um número k que indica a quantidade de vizinhos do registro xi, o algoritmo aplica a função do tipo de distância escolhido calculando assim a distância entre o registro xi e demais registros xj, i ≠ j, ou seja, que possuem o atributo t completo. Depois o algoritmo ordena de forma crescente os registros de acordo com as distâncias encontradas e retorna os k primeiros registros.

A imputação do atributo t ausente em xi é feita com o valor resultante da média desses k primeiros registros retornados.

2.3.3 Back Propagation

Os sistemas conexionistas consideram o cérebro como uma rede de elementos computacionais tornando assim possível a modelagem e resolução de problemas de baixo nível. Por exemplo, padrões e memórias.

Essa comunicação entre neurônios, a computação tenta implementar em modelos ou sistemas através de neurônios artificias. A primeira tentativa foi o neurônio linear, o qual a ativação do estado atual não depende de estados anteriores. Outra tentativa foi o neurônio de McCulloch-Pitts proposto em 1943 por Warren McCulloch e Walter Pitts (MCCULLOCH, PITTS, 1943), que é bem semelhante ao neurônio biológico, já que possui apenas dois estados: excitação, valor um; e inibição, valor zero.

Também temos a tentativa chamada *Back Propagation* (RUMELHART et al, 1986), que é uma rede neural com retro propagação de erro cuja arquitetura é feedforward, ou aprendizado supervisionado. Logo a seguir é possível analisar a arquitetura genérica desse tipo de rede neural:

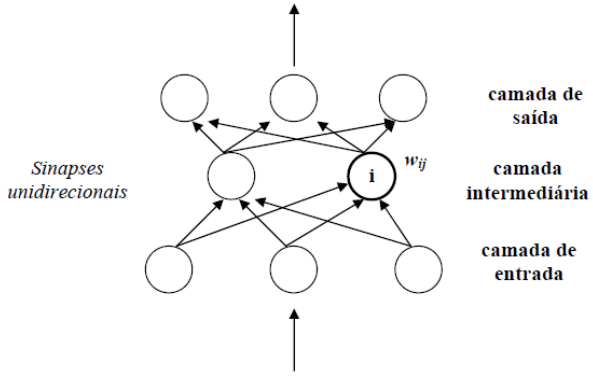


Figura 2.1 Arquitetura genérica de uma rede *Back Propagation*

Fonte: CARVALHO (2001)

Nessa arquitetura, todos os neurônios da rede são interconectados a cada saída de um neurônio é gerada uma nova entrada que é repassada pelos neurônios da camada de entrada para todos os neurônios da camada intermediária, sendo esse ciclo repetido até todos os neurônios do último nível enviem suas respostas a todos os neurônios da camada de saída.

As variáveis como o número de neurônios das camadas e o número de níveis da camada intermediária variam de acordo com o problema a ser solucionado. Sobre a modelagem da camada intermediária, é ela que vai definir se serão resolvidos problemas lineares ou não lineares.

Algoritmo do Back propagation

Dado um conjunto de dados de entrada e suas respectivas saídas, o algoritmo compara as saídas produzidas pela rede com as que são desejadas e ajusta os pesos das conexões sinápticas dos neurônios da rede em função a diferença entre os valores comparados.

2.4 Técnicas de Seleção

Como o próprio nome diz, a técnica de seleção elege um grupo de atributos ou valores atuando no desempenho da imputação, além de poder tornar o resultado mais eficaz. Por exemplo, a complexidade na geração de regras de associação diminui com uma menor quantidade de atributos.

Seu objetivo é identificar qual das 2^n combinações desses atributos deve ser considerada no processo de descoberta de conhecimento desde que a informação original correspondente ao total de atributos seja ao máximo (SOARES, 2007, p.30). Todavia alguns atributos podem ser indiferentes ou praticamente indiferentes ao processo de pré-mineração de dados, como é o caso de chaves primárias e candidatas.

Se a técnica de seleção demorar pouco tempo para executar e se for em seguida aplicado o algoritmo de mineração de dados, então o tempo de processamento pode ser menor do que o tempo de processamento somente do algoritmo de mineração de dados. As técnicas de seleção estudadas nesse trabalho são o *PCA - Análise de Componentes Principais* e o *AG - Algoritmos genéticos*, mas existem muitos outros.

2.4.1 PCA - Análise de Componentes Principais

O *PCA* (SMITH, 2002, SHLENS, 2005) tem por objetivo a análise de um conjunto de dados visando a sua redução, eliminando sobreposições e a escolha das formas mais representativas de dados a partir de combinações lineares das variáreis originais. Podem ser utilizadas em diversas situações: visualização, aquisição de imagens de objetos 2D de acordo com o posicionamento da câmera ou reconhecimento das principais características de medida a serem usadas. É também conhecida por Transformada Discreta de Karhunen-Loève (KLT) ou Transformada Hotelling, em homenagem a Kari Karhunen, Michel Loève e Harold Hotelling.

Trata-se de um método estatístico simples, sendo muito utilizada pela comunidade de reconhecimento de padrões. O *PCA* é uma transformação linear ótima, onde os componentes principais fazem parte do arranjo que representa melhor a distribuição dos dados (vide reta na imagem abaixo) e os componentes secundários são perpendiculares ao principal (vide reta tracejada na imagem abaixo).

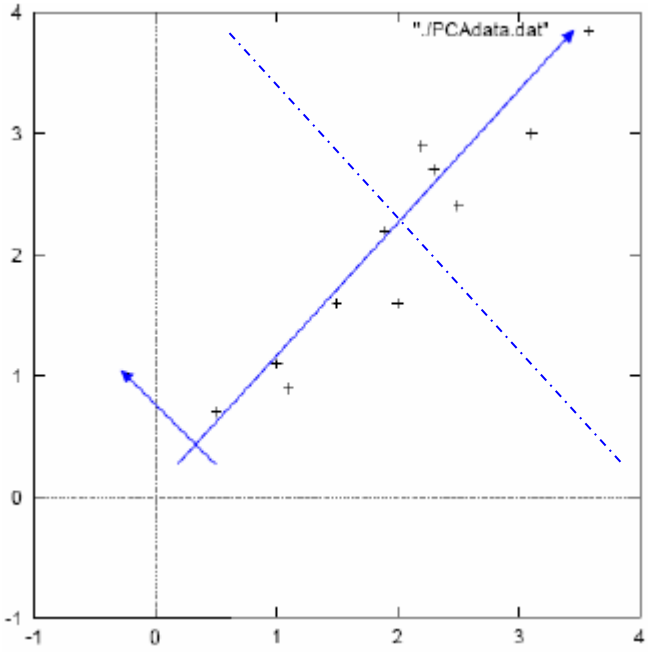


Figura 2.2 Componente principal e componente secundário

Fonte: Adaptado de SOARES (2007)

Algoritmo PCA – Análise de Componentes Principais (Principal Component Analysis)

Dado um conjunto de dados de entrada, o algoritmo calcula primeiro a média de todos os dados e subtrai a média de todos os dados. Depois calcula a matriz (n x n) de covariância usando todas as subtrações sendo esta o resultado da média do produto de cada

subtração por ela mesma. Em seguida calcular os auto valores e auto vetores da matriz covariância, e então arrumar a matriz da Transformada de Hotelling cujas linhas são formadas a partir dos auto vetores da matriz de covariância colocados, de modo que o elemento (0,0) seja o autovetor correspondente ao maior autovalor. Por fim, o componente principal é o autovetor com o maior autovalor associado.

2.4.2 Algoritmos Genéticos

Os *Algoritmos Genéticos* (HOLLAND, 1995) são técnicas probabilísticas e não determinísticas, ou seja, uma mesma população inicial e o mesmo conjunto de parâmetros pode gerar soluções diferentes a cada execução. Isso acontece porque os Algoritmos Genéticos (AGs) são um tipo de algoritmos evolucionários, baseados na teoria da evolução descoberta por Lamarck e Charles Darwin.

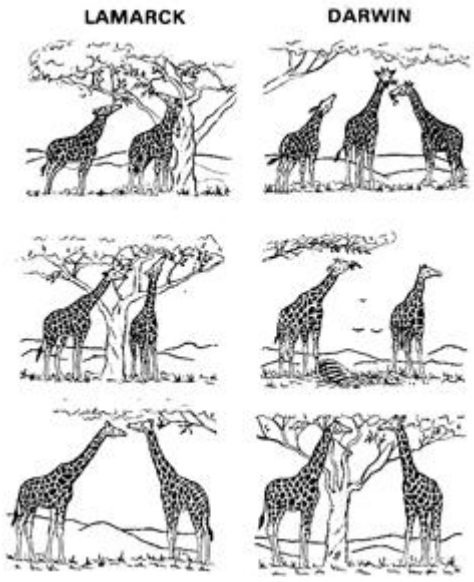


Figura 2.3 As primeiras teorias da evolução
Fonte: Adaptado de DONOSO (2004)

Segundo DONOSO (2004), Lamarck (1744 – 1829) foi um dos primeiros cientistas a defender e propor uma teoria sistemática de evolução. Apesar de atualmente sua teoria ter sido superada, na época era algo muito revolucionário. Ela dizia que o mecanismo evolutivo estava baseado em duas leis: Lei do uso e desuso, e Lei da transmissão dos caracteres adquiridos. Quer dizer, uma diz que o uso frequente de determinadas partes do organismo

leva a hipertrofia, e o desuso prolongado faz com que se atrofiem; e a outra diz que as características conquistadas pelo uso ou perdidas pelo desuso são transmitidas aos descendentes.

Charles Darwin (1809-1882) propôs em 1859, em seu livro a Origem das Espécies, suas ideias sobre os mecanismos da evolução: a Teria da Evolução através da Seleção Natural. Sua teoria dizia que os organismos mais bem adaptados ao meio possuem as maiores chances de sobrevivência, deixando assim um maior número de descendentes e, portanto, esses possuem mais chances de sobrevivência num ambiente de constantes mudanças.

Depois veio outra teoria baseada nas ideias do Darwin: Teoria Sintética da Evolução ou Neodarwinismo. Nesta vemos a adição de novos conhecimentos científicos fundados na Genética e que só foram descobertos pelos trabalhos de Mendel. Além da seleção natural, o Neo darwinismo reconhece como principais fatores evolutivos a mutação genética, a mutação cromossômica, a recombinação genética e o isolamento reprodutivo. Os algoritmos genéticos estão estritamente ligados a esta última teoria e veremos claramente o paralelo entre a teoria e o algoritmo baseando-nos em FREITAS (2002).

Cromossomos

Um cromossomo é uma sequência de DNA que contém vários genes e outras sequências de nucleotídeos com funções específicas nas células e dos seres vivos. Nos algoritmos genéticos a forma mais simples de representá-los é através de uma string binária de tamanho fixo. Porém essa representação torna-se problemática quando as variáveis assumem valores contínuos (LINDEN, 2006, p.165), pois que 0 ou 1 não conseguirá representar adequadamente uma enorme variação de valores, tornando não confiável a solução.

Se MICHALEWICZ (1996), citado por LINDEN (2006), sugere usar outro tipo de representação para essa e outras situações que a representação binária não é apropriada, então é preciso pesquisar as representações já criadas ou até criar uma nova representação se nenhuma atender o problema. Pois que, segundo LINDEN (2006), quanto mais adequada ao problema a representação for, maior a qualidade dos resultados obtidos, então é melhor resistir à tentação de adequar o problema à uma representação.

O que é igual em todas as representações, independente do tipo de codificação utilizada, é que em algoritmo genético convencional o cromossomo não conhece as informações que ele carrega. (DHAR, STEIN, 1997 citado por LINDEN, 2006)

Representação Binária

COSTA (2005, p.39-41) definiu algumas possibilidades de representação binária:

a) O valor do atributo, que é o cromossomo, é representado por uma cadeia de bits que assumem o valor 0 ou 1 conforme abaixo:

CROMOSSOMO	REPRESENTAÇÃO	VALOR
A	010010	18
B	101010	42
C	001100	12
D	001010	10

Tabela 2.1 Representação binária por cadeia de bits

b) Outra possibilidade é representar um registro, uma tupla de atributos, colocando 0 se o valor do atributo existir ou 1 se o valor do atributo não existir, ou seja, se estiver ausente. Vejamos esta representação logo abaixo:

CROMOSSOMO	REPRESENTAÇÃO	TUPLA
A	111101	2 55 9 6 null 0
B	011111	null 2 0 66 1 4
C	100111	55 null null 6 3 1
D	111000	0 6 9 null null null

Tabela 2.2 Representação binária por tupla

Representação por Permutação

Essa representação é útil em problemas de ordenação, como por exemplo o problema do caixeiro viajante: o caixeiro viajante precisa viajar para um número X de cidades, e o problema é achar a ordem em que as cidades foram visitadas para que ele percorra todas as cidades utilizando a menor distância possível.

Aqui um cromossomo é uma string de números que representa uma posição qualquer na sequência, como está na tabela abaixo:

CROMOSSOMO	REPRESENTAÇÃO
A	5 6 7 4 2 1 3
B	1 5 9 4 3 7 2
C	7 6 8 2 4 1 0
D	0 9 5 4 3 6 1

Tabela 2.3 Representação por permutação

Representação Real

O cromossomo pode assumir qualquer valor dependendo do contexto do problema, por exemplo:

CROMOSSOMO	REPRESENTAÇÃO
A	ONHGERGJERIOGHERIOG
B	(castanho), (ruivo), (loiro), (preto), (cinza), (branco)
C	2,3654 3,6954 9,1245 7,2542
D	0 9 5 4 3 6 1

Tabela 2.4 Representação real

População

Cada ser humano ou indivíduo possui seus cromossomos, e esses indivíduos formam uma população.

O mesmo termo é usando em algoritmos genéticos quando se define que uma população é um conjunto de cromossomos, ou conjunto de indivíduos. A diferença é que se considera que cada indivíduo possuiu apenas um cromossomo enquanto na biologia o ser humano possui diversos cromossomos.

Assim como ocorre na biologia, a população é usada para gerar uma nova população até chegar à solução final do problema, ou melhor, a população de indivíduos mais adaptados. O paralelo ainda continua ao pensar que o tamanho da população irá influenciar a boa solução, pois que uma população pequena não irá gerar uma população de indivíduos tão bem adaptados quanto uma população grande que possui maior diversidade.

Entretanto, uma grande população demorará a gerar a população melhor adaptada e isso irá acarretar um grande recurso computacional no caso dos algoritmos genéticos podendo se tornar inviável ou muito custoso além de dispendioso.

Operadores Genéticos

O princípio básico dos operadores genéticos é transformar a população através de sucessivas gerações até chegar ao resultado satisfatório, ou seja, mais bem adaptado. Esses operadores fazem com que a população se diversifique e ao mesmo tempo mantenha características de adaptação adquiridas pelas gerações anteriores.

Temos como operadores a recombinação genética (crossover ou cruzamento) e a mutação, mas em cada um deles também temos uma diversidade de possibilidades de operação que podem facilitar mais ou menos as mudanças genéticas.

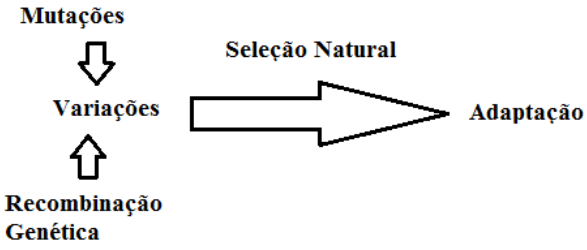


Figura 2.4 Operação genética

Recombinação Genética - Crossover ou Cruzamento

É um mecanismo de reorganização dos genes já existentes nos cromossomos, a principal forma de recombinação genética ocorre durante a reprodução sexuada. Assim como na biologia, crossover, nos algoritmos genéticos, é o operador responsável pela recombinação de características dos pais durante a reprodução, permitindo que as próximas gerações herdem essas características. É considerado o operador genético predominante, por isso é aplicado com uma taxa maior do que a de mutação.

Figura 2.5 O crossover de um-ponto

Os tipos de crossover são: um-ponto, multipontos, uniforme e aritmético.

No caso do crossover de um-ponto, um ponto do cromossomo é escolhido como ponto de cruzamento e o cruzamento das informações genéticas dos pais ocorre a partir desse ponto. Ou seja, as informações genéticas anteriores a este ponto em um pai são “ligadas” às informações genéticas posteriores do outro pai.

Se os cromossomos A e B forem escolhidos para a operação de crossover de um ponto e que o ponto de cruzamento foi escolhido um ponto antes da metade do cromossomo, assim veremos o seguinte resultado:

CROMOSSOMO	REPRESENTAÇÃO PAI	REPRESENTAÇÃO FILHO
A	001|01010	001|11111
B	011|11111	011|01010

Tabela 2.5 Crossover um-ponto

E o crossover multiponto é parecido com o um-ponto sendo que, ao invés de escolher apenas um ponto no cromossomo, são escolhidos vários pontos de cruzamento. Então as informações genéticas dos cromossomos pais serão trocadas a partir desses vários pontos.

Agora os mesmos cromossomos pais farão o crossover multiponto, veremos como ficarão os filhos:

CROMOSSOMO	REPRESENTAÇÃO PAI	REPRESENTAÇÃO FILHO
A	00|1010|10	00|1111|10
B	01|1111|11	01|1010|11

Tabela 2.6 Crossover dois-pontos

Já o crossover uniforme não usa pontos e sim um parâmetro global para determinar a probabilidade de cada variável ser trocada entre os pais. No exemplo abaixo, os cromossomos pais vão trocar os bits 1, 4 e 5.

CROMOSSOMO	REPRESENTAÇÃO PAI	REPRESENTAÇÃO FILHO
A	00101010	00111110
B	01111111	01101011

Tabela 2.7 Crossover uniforme

No crossover aritmético é feita uma operação de bits usando AND ou OR ou XOR. Logo abaixo vemos os filhos gerados através da operação AND.

CROMOSSOMO	REPRESENTAÇÃO PAI	REPRESENTAÇÃO FILHO
A	00101010	00101010
B	01111111	00101010

Tabela 2.8 Crossover aritmético

Mutação

O operador de mutação é um fator importante para a introdução e a manutenção da diversidade genética da população, pois altera arbitrariamente uma ou mais informações genéticas dos cromossomos de um só indivíduo gerando, assim muitas das vezes, cromossomos com informações totalmente novas.

Uma taxa de mutação determina a probabilidade dessa operação ocorrer, grande parte dos problemas atribuem uma taxa bem pequena, por volta de 1%, para não gerar desvios nos resultados e também por essa operação não ser tão importante quanto o crossover. Claro, que essa taxa irá variar de acordo com o problema em questão, pode haver casos que sejam necessários uma taxa não tão baixa, mas certamente essa taxa não será alta.

Na tabela baixo temos o cromossomo A sofrendo a mutação no 4º bit, o cromossomo B sofrendo a mutação de troca de ordem do 3º bit com o 7º bit e o cromossomo C sofrendo a mutação de adição ou subtração de um valor pequeno para um valor real.

CROMOSSOMO	REPRESENTAÇÃO PAI	REPRESENTAÇÃO FILHO
A	00101010	00111010
B	356914278	352914678
C	2,36 9,54 3,98 7,52 1,25	2,36 9,41 3,98 7,52 1,35

Tabela 2.9 Mutação

Função de avaliação ou aptidão (Fitness)

Ao aplicar os operadores genéticos a população inicial irá gerar uma nova população formada pelos seus filhos, mas nem todos os filhos devem fazer parte da nova geração: apenas os indivíduos mais bem adaptados. Assim como ocorre com uma população de animais, só conseguirá reproduzir se sobreviver a predadores e outros obstáculos naturais. Então os predadores e obstáculos atuarão através do que chamamos de *Fitness*, ou Função de avaliação. A *Fitness* determina quão boa é a população gerada para a solução do problema, formando assim uma nova geração que irá sofrer as operações genéticas ou uma solução final do problema.

Se a função objetivo não determinar bem a solução do problema, poderá ser como predadores que irão exterminar os indivíduos que fariam parte de uma população ideal, ou seja, vai eliminar os cromossomos que fariam parte de uma solução ótima que iria atender o problema.

Algoritmo Algoritmos Genéticos

Dada uma população, que é o conjunto de entradas no início, é gerada outra população com n cromossomas, ou indivíduos. É gerada uma função de aptidão para cada cromossoma na população. É criada uma nova população repetindo os passos abaixo até que a nova população esteja completa. Dois outros cromossomas são selecionados de acordo com a aptidão e, de acordo com a taxa de crossover, os pais geram o filho com ou sem a operação de crossover. De acordo com a taxa de mutação, um novo filho é gerado utilizando-se o operador de mutação.

Depois os filhos são introduzidos na população e a nova população está pronta para um próximo ciclo. Os critérios de parada são verificados (na maioria dos casos número de geração e total de indivíduos) e caso não sejam satisfatórios retorna ao passo inicial.

2.5 Técnicas de Agrupamento

Assim como a técnica de seleção seleciona atributos ou valores, a técnica de agrupamento agrupa atributos ou valores que possuam a mesma característica. Mas como em computação normalmente não se conhece quais são essas características, é preciso usar alguma medida de similaridade como base.

O objetivo do agrupamento seria maximizar a similaridade dos atributos ou valores de um grupo e minimizá-la para os atributos ou valores de grupos distintos, sendo assim, entende-se que para que o agrupamento seja bem sucedido depende de uma boa escolha na medida de similaridade (SOARES, 2007, p.43).

2.5.1 K-Means

O K-Means (MCQUEEN, 1967) é um algoritmo muito antigo, mais de 40 anos, mas mesmo assim o mais utilizado até hoje, isso se dá devido a sua simplicidade e alto desempenho.

Segundo FERLIN (2008, p. 217), o K-Means é um algoritmo guloso que tem por objetivo principal definir k centros, ou melhor centróides, um para cada grupo, através de uma função objetivo. Sendo o centróide de um grupo o vetor médio, então o objetivo desse algoritmo é minimizar a função objetivo que calcula a distância total entre os objetos e os centróides do grupo que ele se encontra.

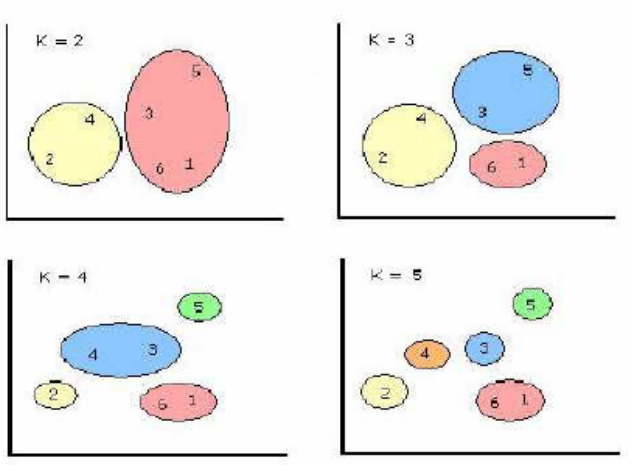


Figura 2.6 Exemplo de K-Means para k=2 a 5

Fonte: FERLIN (2008) adaptado de CHIPMAN *et al* (2003)

Os primeiros centróides podem ser escolhidos de forma aleatória (a forma clássica) ou podem ser calculados pelos primeiro objetos do conjunto original. Duas métricas que podem ser utilizadas são a distância Euclidiana e a Manhattan.

Além do K-Means, temos também o K-Medoids. A diferença é que enquanto no Kmeans o centróide é a média dos objetos do grupo, no K-Medoids o centróide é o objeto da coleção de entrada que está mais próximo do centro.

Algoritmo K-Means

Dada um grupo de objetos iniciais, o algoritmo calcula os primeiros centróides, usando a forma aleatória ou de primeiro registros. Então é associado um centróide a cada objeto e depois calculado um novo centróide para cada grupo em função dos objetos alocados. Isso é feito até que os objetos não mudem de grupo, ou até que o número máximo de iterações tenha sido alcançado.

2.5.2 Redes de Kohonen

Redes de Kohonen (HAYKIN, 1994) é relativamente simples de ser implementada ao comparar com sua capacidade de formar grupos (clusters) com dados complexos. Não é necessário conhecer os padrões de entrada, pois trata-se de um algoritmo auto organizável (SOM - Self Organizing Maps).

Nas redes SOM, todos os neurônios que são as unidades básicas de processamento da rede recebem um estímulo igual e competem para descobrir quem é o vencedor, produzindo assim um aprendizado competitivo. Além disso, as redes SOM são baseadas nas características do córtex cerebral onde funções distintas são realizadas em regiões distintas do cérebro, o que dá uma ideia de agrupamento (FERLIN, 2008, p.223). Assim eles podem ser aplicados se aplicado em diversas áreas de conhecimento.

FERLIN (2008, p.223) define bem esse algoritmo quando diz que “Em resumo, a rede SOM é uma rede competitiva capaz de mapeamentos que preservam a topologia entre os espaços de entrada e de saída e refletem os padrões significativos ou característicos dos dados de entrada”.

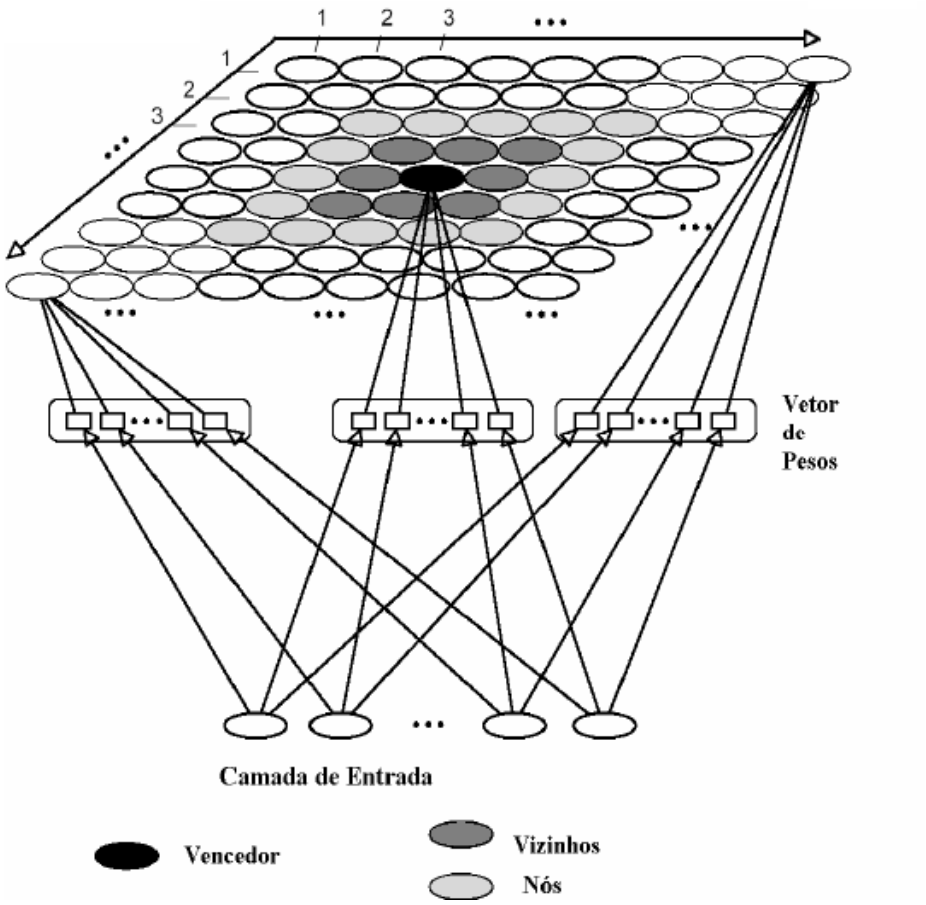


Figura 2.7 Exemplo de uma rede SOM

Fonte: adaptado FERLIN (2008) adaptado de ULTSCH *et al* (1992)

A imagem acima demonstra uma rede SOM, onde cada neurônio de entrada (cada atributo ou registro) está conectado com todos os outros neurônios da camada seguinte por meio de conexões ponderadas. Já a camada acima é conhecida por *Redes de Kohonen*, ou camada de saída, pois que é composta por um conjunto de neurônios em forma de vetor ou matriz e representa no mapa o qual a entrada irá se basear.

Algoritmo Redes de Kohonen

Dada um grupo de objetos iniciais, o algoritmo executa as fases: de competição onde os pesos dos nós são definidos, depois de colaboração onde a vizinhança do nó vencedor é definida e por fim de adaptação onde os pesos do nó vencedor e seus vizinhos são atualizados. Essas fases se repetem até o mapa não mudar ou atingir um número máximo de iterações.

CAPÍTULO 3
TRABALHOS RELACIONADOS

Diversos trabalhos relacionados à complementação de dados ausentes foram amplamente estudados fornecendo um direcionamento a este trabalho. Alguns serviram como orientação por não terem o mesmo foco deste trabalho, por exemplo, os do ramo estatístico (RUBIN, 1988), médico/genético (CELTON et al, 2010), industrial (LAKSHMINARAYAN, HARP, SAMAD, 1999), Engenharia de Software (SONG, SHEPPERD, CARTWRIGHT, 2005), Datawarehouses e ETL (RIBEIRO, 2010), etc. Outros trabalhos se tornaram verdadeiro alicerces como o Pré-processamento em mineração de dados (SOARES, 2007), Seleção de Variáveis na Descoberta de Conhecimento em Base de Dados (COSTA, 2005) e Imputação Multivariada: Uma Abordagem em Cascata (FERLIN, 2008).

3.1 Appraisal

SOARES (2007) desenvolveu um sistema chamado *Appraisal*, em linguagem Java e com flexibilidade para adição de outras técnicas ainda não implementadas já que o código permite novas configurações. Graças a essa flexibilidade RIBEIRO (2008), FERLIN (2008) e MONTEIRO (2008) puderam usar o Appraisal para diferentes formas de complementação de dados ausentes. RIBEIRO (2008) implementando a imputação em ausência multivariada, FERLIN (2008) desenvolvendo a imputação em cascata em ausência multivariada e MONTEIRO (2008) experimentando a imputação composta com dados categóricos.

O *Appraisal* combina técnicas diversas formando as seguintes estratégias: imputação simples, imputação antecedida por seleção, imputação antecedida por agrupamento, imputação antecedida por seleção e agrupamento, imputação antecedida por agrupamento e seleção.

O sistema é formado por quatro módulos:

1- Eraser: Módulo de geração de bases com valores ausentes de acordo com o mecanismo e percentual de ausência escolhidos.
2- Crowner: Módulo de execução das estratégias de complementação de dados ausentes.
3- Committee: Módulo de geração de valores de imputação baseados nos valores sugeridos pelas estratégias e outros atributos da base de dados.

4- Reviewer: Módulo de geração de gráficos que mostram os melhores resultados produzidos pelos módulos Crowner e Committee.

Com exceção do *Committee*, todos os outros módulos são utilizados exatamente na mesma ordem. Sendo que o Cowner e Reviewer foram alterados para se adequarem a proposta desse trabalho como veremos no próximo capítulo.

3.2 Imputação de dados

Métodos para realizar imputação em dados ausentes ou complementação de dados tem sido estudados desde a década de 50 começando com ANDERSON (1946) continuando com PREECE (1971), DEMPSTER et al (1977), RUBIN (1977,1987, 1988), SCHAFER (2000), etc.

Dentre esses métodos de imputação, há um algoritmo que usa regras de associação mas para tal é necessária a intervenção do usuário na montagem das associações e decisão de qual regra utilizar. Esse algoritmo chamado *MVC - Missing Values Completion* foi estudado por CRÉMILLEUX, RAGEL e BOSSON (1999) e RAGEL e CRÉMILLEUX (1999) sendo que este utiliza outro algoritmo proposto anteriormente pelos mesmos autores: *RAR - Robust Association Rules*.

Em LAKSHMINARAYAN, HARP e SAMAD (1999) estudaram dois métodos de imputação em uma base com informações sobre processos industriais: C4.5 (QUINLAN, 1993) e *AutoClasse* (CHEESEMAN, STUTZ, 1996). Os resultados mostram que o algoritmo C4.5 teve melhor qualidade na imputação simples, mas na imputação múltipla os dois algoritmos tiveram bons resultados.

O algoritmo de verossimilhança com informações completas (FIML) já foi comparado com a imputação por média (MYRTEIT, STENSRUD, OLSSON, 2001) e os resultados sugerem o uso de FIML.

HRUSCHKA, HRUSCHKA JR e EBECKEN (2002b) desenvolvem a imputação de dados ausentes com rede bayesiana aplicando o algoritmos *K2* (COOPER, HERSKOVITS, 1992). Outro estudo de métodos bayesianos foi realizado em dados médicos (AUSTIN, ESCOBAR, 2005).

O Algoritmo do *K Vizinhos Mais Próximos* é um método de imputação amplamente estudado, iniciado por TROYANSHYA et al. (2001) com a proposta de um *k-NN Impute*, por

KIM et al. (2004) com a proposta do SKNN (*Sequential k-NN*) e por BRÁS e MENEZES (2007) com a proposta do IKNNimpute (*Iterative k-NN Impute*).

Nos primeiros trabalhos (JÖNSSON, WOHLIN, 2004) o algoritmo dos k vizinhos mais próximos era executado usando somente os dados completos (em dados ausentes), o que causava uma grande perda devido a grande diminuição da base de dados. Mais tarde, mesmos autores sugeriram duas abordagens: casos completos (sem registros com dados ausentes) e casos incompletos (com registros com dados ausentes similares ao registro do atributo a ser imputado, sendo que o próprio atributo a ser imputado não pode estar ausente). Os autores fizeram o uso do mecanismo de ausência completamente aleatório (*MCAR*) e não consideraram os dados já imputados, e conseguiram comprovar que o uso do *k-NN* é viável. Em outro estudo, os mesmos autores obtiveram bons resultados utilizando substituição aleatória, imputação aleatória, imputação por média e por moda, entretanto concluíram que *k-NN* obtém melhor resultado quando a quantidade de atributos é grande.

Outro estudo baseado no *k-NN* foi feito por BASTISTA e MONARD (2001) comparando os resultados com o algoritmo *C5.0*, uma variante do *C4.5* de QUINLAN (1993). Depois os mesmos autores ampliaram esse estudo (BATISTA, MONARD, 2003a) analisando os resultados do algoritmo *k-NN* em comparação com imputação por média ou moda, *C4.5* (QUILAN, 1993) e *CN2* (BOLL, ST. CLAIR, 1995). Complementando, os autores estudam as características dos dados que podem ocasionar uma baixa qualidade na imputação por *k-NN*.

HUANG e LEE (2004) propõem uma modificação no cálculo da medida de similaridade do *k-NN* utilizando a análise relacional de Grey e seus resultados apontam que superou as expectativas em comparação com a média e a imputação múltipla.

Ao comparar a imputação por média com imputação por *k-NN* em dados com ausência *MCAR* e *MAR* (SONG, SHEPPERD, CARTWRIGHT, 2005), concluíram que para problemas de Engenharia de Software a imputação por média obteve mais precisão devido a não importância do mecanismo de ausência nos dados utilizados que podem ser considerados como *MAR*.

Em outro estudo, HRUSCHKA, HRUSCHKA JR e EBECKEN (2003b) desenvolveram *k-NN* usando dois tipos de distância, Euclidiana e *Manhattan*, e compararam os resultados da imputação com *k-NN* e com média. Seus resultados novamente comprovou que a qualidade da imputação com *k-NN* é maior e que a distância *Manhattan* obteve melhores resultados em bases normalizadas.

Algumas características levam o *k-NN* a ser muito utilizado: serve tanto para atributos numéricos quanto nominais, o banco de dados pode ter vários valores ausentes (a qualidade da imputação depende mais da quantidade de dados completos), não é necessário construir um modelo para cada atributo com valor ausente e pode considerar a estrutura de correlação dos dados.

A imputação *hot-deck* (FORD, 1983) foi estudada por MAGNANI e MONTESI (2004) ao aplicarem um algoritmo de agrupamento na base de dados sem retirar nenhum registro com dados ausentes.

Também temos estudos baseados na combinação de métodos. TSENG, WANG e LEE (2003) desenvolveram a imputação e agrupamento para complementar dados ausentes. HRUSCHKA, HRUSCHKA JR e EBECKEN (2003a) propõem a imputação com *k-NN* precedida de agrupamento com Algoritmos Genéticos por eles chamado CGA – *Clustering Genetic Algorithm* para comparar com a imputação com média simples também precedida por CGA concluindo que qualidade da imputação foi melhor usando *k-NN*. CARTWRIGHT, SHEPPERD e SONG (2003) desenvolvem a imputação com média amostral e a imputação com *k-NN* comparando suas aplicações em bases com remoção completa dos casos e concluindo mais uma vez que o *k-NN* obteve menos erros.

Analisando a aplicação de métodos de seleção e agrupamento antecedendo os métodos de imputação (SOARES, 2007), conclui-se que houve melhora na qualidade da complementação tanto com o uso de seleção quanto de agrupamento, dependendo das características das bases de dados. Além disso, SOARES (2007) aplicou comitês de complementação para realizar imputação múltipla com conceito de meta-aprendizado. Para tal, o autor desenvolveu um sistema chamado *Appraisal* que permite a parametrização de novos métodos e a execução dos mesmos nas estratégias já implementadas.

Complementado SOARES (2007), RIBEIRO (2008) estudou e implementou no *Appraisal* a imputação sequencial objetivando demonstrar se a opiniões negativas sobre a qualidade dessa implementação são válidas ou não. E o autor comprova não serem válidas, pois que a reutilização dos dados já imputados serviu para aprimorar a precisão da imputação de outros dados ausentes na maioria dos experimentos por ele realizado.

Continuando o trabalho de SOARES (2007) e unindo o trabalho de RIBEIRO (2008), uma imputação em cascata em base de dados com ausência multivariada foi proposta e implementada no *Appraisal* por FERLIN (2008) desejando avaliar o impacto não só da reutilização dos dados já imputados como também da ordem no processo de imputação. Os

resultados apontaram melhora na qualidade da imputação em cascata em comparação com a imputação sequencial.

Um estudo visando tratar valores ausentes de um *Data Warehouse* (DW) durante o processo de ETL (RIBEIRO, 2010) utilizou a imputação com *k-NN* e construiu um sistema chamado *ComplETL* para realizar os testes, avaliar os resultados e concluir que o *k-NN* não é ideal em todas as bases de dados de um *DW*, apenas nas bases pequenas por possível verificar menor taxa de erro na complementação.

SILVA (2010) propõe e analisa a aplicação de um método de agrupamento para complementação de dados ausentes se baseando em um algoritmo evolutivo utilizando seis bases de dados com ausência *MCAR* e *MAR*. Ao comparar com seis outros algoritmos de imputação (imputação pela média ou moda, *k-NN Impute*, SKNN, *k-NN* Interativo, KML) existentes na literatura o desempenho do algoritmo proposto pelo autor foi parecido.

A proposta de BURGETTE e REITER (2010) de implementar um método de imputação multivariada através de uma cadeia de equações baseadas em árvores de regressão sequencial demonstra que a qualidade de complementação dos dados ausentes é melhor do que os métodos de imputação sequencial baseados em simples regressão.

Para complementar dados ausentes genes de *microarrays* já tinha sido demostrada a aplicação da imputação com *k-NN*, porém em CELTON et al (2010) mais 12 métodos foram aplicados e comparados ao efetuar o agrupamento dos genes na expectativa que as associações entre os genes fosse completamente restaurada. Os autores concluíram que agrupamento com *K-Means* foi mais eficiente para preservar as associações dos genes, mas nenhum métodos de imputação foi capaz de restaurar de forma completas as associações corretas dos genes. Com base nos teste de cinco bases de dados biológicos, o método *EM_array* foi o mais eficiente na imputação.

Em ZENG (2011) dois métodos de imputação foram escolhidos para implementação e testes, algoritmo *EM – Expectation Maximization* e *MI – Multiple Imputation* usando *MCMC - Markov Chain Monte Carlo*, e ambos forneceram bons resultados sendo que um terceiro método baseado em *BART – Árvore de Regressão Aditiva Bayesiana* obteve os melhor resultados.

WHITE, REITER, PETRIN (2011) apresentam duas estratégias para tratar dados ausentes na imputação multivariada, uma imputando valores via sequencia de classificação e árvores de regressão e outra avaliando essas imputações baseadas em posteriores verificações. Os resultados sugeriram que as duas estratégias tiveram grandes diferenças na qualidade.

3.3 Pré-processamento em mineração de dados

A seguir SOARES (2007) é abordado com o objetivo de detalhar quais as bases de dados e as técnicas de imputação, seleção e agrupamento foram utilizadas, juntamente com suas respectivas análises referentes aos resultados.

Três técnicas de imputação para complementar dados ausentes foram implementadas e os combinadas com uma técnica de seleção e uma técnica de agrupamento montando 14 diferentes estratégias para avaliar quais trariam melhores resultados, ou seja, menor taxa de erro absoluto.

Como técnicas de imputação fez-se uso da *Média*, *k-NN* e *Back Propagation*. Já como técnica de seleção usou-se *PCA - Análise de Componentes Principais* e como agrupamento *K-Means*. As configurações de cada algoritmo foram definidas através de testes e este trabalho não irá simplesmente utilizar essas mesmas configurações, mas sim as configurações que deram melhores resultados.

Bases de Dados

Todas as estratégias foram aplicadas em três bases de dados todas provenientes do repositório de bases de dados da Universidade da Califórnia, Irvine (NEWMAN et al, 1998):

- *Iris Plants*, com 150 registros, quatro atributos numéricos e um classificador, que registra 4 tipos de plantas.

- *Pima Indians Diabetes*, com 392 registros, oito atributos numéricos e uma classe, que registra dados médicos de pacientes categorizando-os como diabéticos ou não.

- *Wisconsin Breast Cancer*, com 682 registros, nove atributos numéricos e uma classe, que registra características clínicas de pacientes com câncer de mama ou não.

Além disso, todas essas bases de dados utilizadas estão originalmente completas, sendo que os valores ausentes foram produzidos de forma artificial e controlada, através do mecanismo de ausência completamente aleatório (*MCAR - Missing Completely At Random*) da seguinte forma:

1. Padrão de ausência univariado: Apenas um atributo com valores ausentes por base com dados ausentes.
2. Dados nulos são gerados de forma aleatória.
3. Cinco percentuais de ausência diferentes: 10%, 20%, 30%, 40% e 50%.

Seguindo esses três pontos, este trabalho gerou um total de 105 bases de dados com valores ausentes com a ajuda do módulo *Eraser* do sistema *Appraisal* ao aplicar os cinco percentuais de ausência para cada atributo das três bases de dados escolhidas (menos o atributo classificador): *Iris Plants, Pima Indians Diabetes e Wisconsin Breast Cancer.*

Cada base de dados com valores ausentes foi identificada da seguinte forma após sua geração: nome da base de dados + mecanismo de ausência de dados + nome do atributo com dados ausentes + percentual de ausência. Assim, por ter 4 atributos, o Eraser gerou 20 bases de dados ausentes a partir da base de dados *Iris Plants* completa. *Pima Indians Diabetes* possui 8 atributos e ao passar pelo Eraser deu origem a 40 bases de dados com valores ausentes. E *Wisconsin Breast Cancer* que contém 9 atributos, foi utilizada pelo Eraser para gerar 45 bases de dados com valores ausentes.

É importante destacar o que motivou utilizar percentuais de ausência diferentes com intervalos de 10% tendo como máximo o 50%, deseja-se assim observar como as estratégias irão se comportar com o aumento de dados ausentes mas ao mesmo tempo entende-se que o percentual deve estar de acordo com casos reais e por isso não foi feito o uso de percentuais acima de 50%.

Imputação com Média

Dois tipos de médias foram adotadas: determinística e Estocástica e conclui-se que a média determinística venceu na maior parte das vezes sem sofrer diferenças entre as bases de dados ou estratégias, ou seja, a perturbação no cálculo da média não trouxe vantagens significativas. De qualquer forma, essa conclusão não pode ser generalizada para outros tipos de bases de dado.

	Determinística	Estocástica
Iris Plants	**75** (93,75%)	**5** (6,25%)
Pima Indians Diabetes	**153** (95,63%)	**7** (4,38%)
Wisconsin Breast Cancer	**151** (83,89%)	**29** (16,11%)

Tabela 3.1 Avaliação da *Média* sem e com perturbação

Fonte: SOARES (2007)

Imputação com k-NN

Para imputação com *k-NN*, foram definidos dois tipos de cálculos de distâncias: Euclidiana e Manhattan, mas não obteve conclusões precisas sobre qual distância foi a melhor

já que os resultados mostraram que a Manhattan obteve taxas maiores em duas bases de dados e a Euclidiana em uma base de dados.

	Euclidiana	Manhattan
Iris Plants	**59** (59,00%)	**41** (41,00%)
Pima Indians Diabetes	**62** (31,00%)	**138** (69,00%)
Wisconsin Breast Cancer	**70** (31,11%)	**155** (68,89%)

Tabela 3.2 Avaliação das distâncias usadas no *k-NN*

Fonte: SOARES (2007)

Conclui-se também que existem situações que a heurística proposta por JÖNSSON e WOHLIN (2004), onde melhor valor do k seria a raiz quadrada dos casos válidos, não se aplica e vemos isso na tabela abaixo.

	Menor k	Maior k
Iris Plants	1	37
Pima Indians Diabetes	1	140
Wisconsin Breast Cancer	1	27

Tabela 3.3 Avaliação do tamanho do K usado no K-NN

Fonte: SOARES (2007)

Imputação com Back Propagation

Sobre as configurações utilizadas no Back Propagation, foi escolhida a sugestão contida na documentação do *JOONE* – Java Object Oriented Neural Engine e por não ter criado outras opções de configuração não foi possível analisar e concluir qual seria a melhor. As configurações adotadas por SOARES (2007) a partir do *JOONE* são:

- learningRate = 0.5, momentum = 0.7 e cycles = 3000.

Seleção com PCA

A configuração utilizada no PCA foi a quantidade de colunas usadas na imputação, sendo que esta poderia variar de um até o total de atributos numéricos da base de dados. Entretanto nada consta nas observações dos resultados sobre qual seria a melhor quantidade de colunas para o PCA.

Agrupamento com K-Means

Utilizando sempre os K primeiros elementos da tabela como centroides iniciais visava-se garantir que a cada rodada os grupos gerados fossem os mesmos. O número de grupos gerados para a estratégia vencedora define dessa forma os melhores K centroides.

	Euclidiana	Manhattan
Iris Plants	**92** (51,11%)	**88** (48,89%)
Pima Indians Diabetes	**160** (44,44%)	**200** (55,56%)
Wisconsin Breast Cancer	**208** (51,36%)	**197** (48,64%)

Tabela 3.4 Avaliação da distância usada no *K-Means*

Fonte: SOARES (2007)

	Menor k	Maior k
Iris Plants	2	60
Pima Indians Diabetes	2	44
Wisconsin Breast Cancer	1	50

Tabela 3.5 Avaliação do k utilizado no *K-Means*

Fonte: SOARES (2007)

3.4 Algoritmos Genéticos - Seleção de Variáveis na Descoberta de Conhecimento em Base de Dados

Neste capítulo COSTA (2005) é referenciado para mostrar a técnica de seleção por ele utilizada: *AG - Algoritmos Genéticos* e que foi usada como base neste trabalho.

As bases de dados foram diversas, total de 8, e diferentes das utilizadas em SOARES (2007). Diversos testes e treinos foram realizados seguindo uma metodologia própria para encontrar as melhores resultados ao processar a seleção com *Algoritmos Genéticos* seguido de imputação com *C4.5*, *Back Propagation* e *Bayes Net*.

Para tal, foram utilizadas duas configurações diferentes na seleção com *Algoritmos Genéticos*:

- Configuração 1:

Tamanho da população – 100

Quantidade de gerações – 50

Total de indivíduos – 5.000

Taxa de Mutação – 0.8%

Taxa de Crossover – 65%

Tipo Crossover – 2 pontos

Técnica AG – steady-state e elitismo juntas

Normalização linear (valor máximo) – não utilizado

Normalização linear (valor mínimo) – não utilizado

Steady-State (gap) – 2

Elitismo – 2

- Configuração 2:

Tamanho da população – 100

Quantidade de gerações – 50

Total de indivíduos – 5.000

Taxa de Mutação – 0.8%

Taxa de Crossover – 65%

Tipo Crossover – uniforme

Técnica AG – steady-state e elitismo juntas com normalização linear

Normalização linear (valor máximo) – 1

Normalização linear (valor mínimo) – 100

Steady-State (gap) – 2

Elitismo - 2

Entretanto não temos informação sobre a melhor configuração, a que obteve melhores resultados.

3.5 Redes de Kohonen - Imputação Multivariada: Uma Abordagem em Cascata

Para apresentar a técnica de agrupamento com *Redes de Kohonen* que foi usada como base neste trabalho, referenciamos FERLIN (2008) neste capítulo.

Visando complementar dados ausentes em mais de um atributo na mesma base de dados, a estratégia utilizada foi processar a técnica de agrupamento com *Redes de Kohonen*,

rede *SOM* (*self-organizing map*), seguido de imputação com *k-NN* de forma sequencial com realimentação de valores.

A escolha do *k-NN* para imputação foi por causa da sua grande utilização e eficiência comprovada. Para este algoritmo as configurações podiam ser ajustadas nos testes durante o desenvolvimento do mesmo e dessa forma se decidiu quais configurações utilizar juntamente com o conhecimento exposto em trabalhos relacionados ao assunto:

- k: testes tendem para k=1 e os demais valores escolhidos para o k, 3, 5 e 10 são justificados pela literatura.
- Distância: Euclidiana, conhecida universalmente e Mahalanobis. Ao escolher a distância de Mahalanobis deseja-se verificar as correlações entre as características dos objetos.
- Possíveis vizinhos: incorporar como casos candidatos também os que têm atributos ausentes que não interfiram na imputação da coluna alvo, algo específico para imputação multivariada.
- Valor estimado: média dos valores dos atributos preenchidos dos vizinhos mais próximos.

Já com relação à configuração do agrupamento com *Redes de Kohonen*, a escolha foi feita por já ter sido utilizada na literatura e por existir um framework bastante difundido (*JOONE - Java Object Oriented Neural Engine*) desenvolvido em linguagem de programação Java para aplicações de Inteligência Artificial baseadas em redes neurais:

- Número de ciclos: 1000, 5.000, 10.000 e 15.000.
- Tipo de aprendizado: aprendizado em lote
- Vizinhança: distribuição Gaussiana
- Inicialização: aleatória
- Taxa de aprendizado: 0.7

Infelizmente FERLIN (2008) não relatou quais configurações deram melhores resultados.

CAPÍTULO 4
TRABALHO PROPOSTO

A proposta desse trabalho é utilizar seleção com *Algoritmos Genéticos* e agrupamento com *Redes de Kohonen* na imputação de dados ausentes. Para tal foi implementado no *Appraisal*, sistema produzido por SOARES (2007), essas duas técnicas que foram adaptadas a partir do COSTA (2005) e FERLIN (2008). Aqui são detalhadas as configurações utilizadas em ambas técnicas mostrando o motivo de tal ou qual escolha.

4.1 Novo Appraisal

Dos quatro módulos do *Appraisal*, apenas três foram utilizados sendo que dois foram alterados para realizar a proposta. No módulo *Crowner* foram implementadas duas novas técnicas e o módulo *Reviewer* foi adaptado para gerar os gráficos com essas duas novas técnicas.

Primeiro foi executado o módulo *Eraser* para gerar as bases de dados com valores ausentes. Em seguida, para cada base de dados gerada no *Eraser* foi executado o módulo *Crowner* para processar as estratégias. E por fim o módulo *Reviewer* foi executado gerando os diversos gráficos para indicar a estratégia vencedora, a que teve menor erro absoluto.

4.1.1 Implementação de seleção com Algoritmos Genéticos

A representação binária é a mais simples e a mais usada nos *Algoritmos Genéticos*, porém LINDEN (2006) ressalta que se esta não for adequada poderá ser descartada. Então, para não optar pela representação binária de forma impensada, foi feita uma avaliação de como as bases de dados irão se comportar usando essa representação.

O primeiro problema que encontramos ao utilizar a representação binária é que a mudança dos bits mais significativos geram uma grande variação entre o valor do pai e o valor do filho. Ou seja, se temos o cromossomo 1000 que representa o valor 8 e este cromossomo sofrer mudança no bit de alta ordem gerando o cromossomo filho 0000 que representa o valor 0, teremos uma grande variação de 8 para 0 no valor. Ao mesmo tempo que se o bit mudado for de baixa ordem gerando o cromossomo 1001 que representa o valor 9, teremos uma pequena variação de 8 para 9 no valor. As grandes variações é um problema pois

pode afetar os resultados, um modo de evitar isso é fazer com que a probabilidade de mutação aumente com a diminuição da ordem para que assim as mutações ocorram somente nos bits que não exercem grande influencia na variação do valor.

Outro problema é o Abismo de Hamming que ocorre quando em alguns cromossomos é preciso mudar todos os bits para efetuar uma mudança de apenas 1 valor unitário diferentemente de outros cromossomos que podem mudar apenas 1 bit para mudar um valor unitário. Como solução desse problema pode-se usar esquemas com muitos símbolos. Os esquemas são como templates que descrevem um subconjunto dentre o conjunto de todos os valores possíveis. Assim ao invés do cromossomo conter todos os bits do valor em questão, ele conterá símbolos nos lugares de alguns bits mantendo apenas os bits que são idênticos a outros valores. Ou seja, um esquema representa mais de um cromossomo, mais de um valor. Entretanto, se é necessário esquemas com muitos símbolos a qualidade da solução irá diminuir.

Mesmo com esses dois pontos negativos com relação a representação binária, foi verificado que para o objetivo do trabalho, que é ranquear os atributos das bases de dados para complementação dos dados ausentes, esses problemas não existiriam. O diferencial seria, ao invés de representar os valores, o cromossomo binário representaria os atributos. Ou seja, cada bit do cromossomo faz referência a um atributo. Para entender melhor, a base de dados *Iris Plants* está sendo exemplificada logo a seguir:

Atributos	**Cromossomos**	**Selecionados**
sepallength,sepalwidth,petallength,petalwidth	1 0 1 0	sepallength,petallength
sepallength,sepalwidth,petallength,petalwidth	0 0 1 1	petallength,petalwidth

Tabela 4.1 Representação binária por atributo

Essa representação foi utilizada pelo trabalho do COSTA (2005) e foi a fonte inspiradora para a avaliação da proposta desse trabalho. Nela a função aptidão faz a ligação dos cromossomos-atributos aos valores da seguinte forma: Para cada cromossomo, a partir dos atributos de bit 1, ou seja, selecionados, é feita uma busca na base de dados agrupando os valores desses atributos. Para cada grupo de atributos será calculado quantos registros há em cada grupo diminuindo pelo número de registros da maior classe. Todos os valores calculados para cada grupo são somados, e esse somatório é dividido pela quantidade de registros total da base de dados. O valor da aptidão é 1 menos valor encontrado na divisão, e o cromossomo que tiver a maior aptidão é selecionado para a próxima geração.

Então, com essa função aptidão, o cromossomo que representa atributos selecionados que possuem menos valores repetidos, ou seja, menos duplicidade é considerado o melhor resultado. Para entender melhor o cálculo da função aptidão, a base *Iris Plants* continua servindo de exemplo:

sepallength	sepalwidth	petallength	petalwidth	Classe
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3,0	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
5,0	3,6	1,4	0,2	Iris-setosa
5,4	3,9	1,7	0,4	Iris-setosa

Tabela 4.2 Seis registros iniciais da base *Iris Plants*

Cromossomo 1 0 1 0:

sepallength	petallength	classe	count(*)
5,1	1,4	Iris-setosa	1
4,9	1,4	Iris-setosa	1
4,7	1,3	Iris-setosa	1
4,6	1,5	Iris-setosa	1
5,0	1,4	Iris-setosa	1
5,4	1,7	Iris-setosa	1

Tabela 4.3 Registros relacionados ao cromossomo 1010

6 grupos possuem o mesmo valor (0 – 0) então:

Aptidão (cromossomo 1 0 1 0) = 1 – (0 / 6) = 1

Cromossomo 0 0 1 1:

petallength	Petalwidth	classe	count(*)
1,4	0,2	Iris-setosa	3
1,3	0,2	Iris-setosa	1
1,5	0,2	Iris-setosa	1
1,7	0,4	Iris-setosa	1

Tabela 4.4 Registros relacionados ao cromossomo 0011

Dos 4 grupos, 3 possuem valor (0 – 0) e 1 possui valor (3 – 1):

Aptidão (cromossomo 0 0 1 1) = 1 – (2 / 6) = 0,66

4.1.2 Implementação de agrupamento com Redes de Kohonen

FERLIN (2008) é uma continuação de SOARES (2007), a qual foi feita uma imputação em cascata com *Redes de Kohonen*, porém utilizando imputação multivariada então a ideia dessa tese foi adaptada para a imputação simples. Além dessa adaptação, necessário se fez adicionar configurações que o agrupamento com *K-Means* de SOARES (2007) utiliza para se adequar ao *Appraisal*.

4.2 Configurações

Todas as configurações das técnicas de imputação, seleção e agrupamento foram definidas com base nos melhores testes e melhores resultados de SOARES (2007), COSTA (2005) e FERLIN (2008).

Durante a execução do módulo *Crowner* do sistema *Appraisal*, os planos de cada técnica acessam o arquivo *"nome da técnica".properties* para obter as informações das configurações que serão utilizadas. Essas configurações são definas com base nos trabalhos relacionados e em experimentos.

Imputação com Média

Já que, de acordo com o capítulo 3.1.1, a perturbação trouxe poucas vantagens, esta não foi utilizada nesse trabalho e a configuração do algoritmo da média ficou quase igual ao do *Appraisal* original:

Variável	Descrição da Variável	Valor
avg.type	Tipo de média aplicada ("arithmetic" - Média aritmética simples)	arithmetic
avg.disturb	Habilita um fator de perturbação para a média ("true" e/ou "false")	false

Tabela 4.5 Configuração da imputação com *Média*

Algoritmo de *Média* ou *Moda*: O mesmo implementado por SOARES (2007).

Imputação com k-NN

Como a melhor distância numa base de dados não foi a mesma distância nas outras bases de dados, foi decidido usar as duas distâncias nesse trabalho: euclidiana e manhattan. Já

com relação ao k, mesmo com intervalos diferentes para cada base de dados, iremos usar os melhores intervalos conforme mostra a configuração adotada para o algoritmo *k-NN*:

Variável	Descrição da Variável	Valor
knn.k	Quantidade de vizinhos ("all" - Valor especial, seleciona de 1 a N-1 vizinhos, onde N é o número total de tuplas)	Iris Plants =1;37;1 Pima Indians = 1;140;1 Breast Cancer =1;27;1
knn.distance	Distância entre os vizinhos ("euclidian" e/ou "manhattan")	euclidian,manhattan
knn.strategy	Estratégia que consolida o atributo a ser regredido ("avg" - Média aritmética)	avg

Tabela 4.6 Configuração da imputação com *k-NN*

Algoritmo dos *K Vizinhos Mais Próximos*: O mesmo implementado por SOARES (2007).

Imputação com Back Propagation

E como não foi definida nenhuma configuração com melhores resultados para o algoritmo *Back Propagation*, foi adotada a mesma configuração do SOARES (2007):

Variável	Descrição da Variável	Valor
bkprop.cycles	Número de vezes que os dados são apresentados a rede	3000
bkprop.learningRate	Taxa de treinamento da rede	0.5
bkprop.momentum	Momentum da rede	0.7

Tabela 4.7 Configuração da imputação com *Back Propagation*

Algoritmo do *Back Propagation*: O mesmo implementado por SOARES (2007).

Seleção com PCA - Análise de Componentes Principais

SOARES (2007) realizou testes com o número de atributos a serem usados na seleção variando entre um e o total de atributos numéricos da tabela. Como não foi informado se com uma quantidade específica de colunas os resultados foram ou não melhores, a

configuração do algoritmo *PCA - Análise de Componentes Principais* utilizará todos os atributos:

Variável	Descrição da Variável	Valor
pca.columns	Quantidade de colunas a serem selecionadas ("all" - Valor especial, seleciona N-1 colunas, onde N é o número total de colunas)	all

Tabela 4.8 Configuração da seleção com PCA

Algoritmo do *PCA*: O mesmo implementado por SOARES (2007).

Agrupamento K-Means

Semelhantemente ao que ocorreu no *k-NN*, as distâncias geraram melhores resultados em bases de dados diferentes usando *K-Means* então o melhor é fazer nesse trabalho as duas distâncias. Já com relação ao k, serão adotados os melhores valores assim como também no *k-NN*:

Variável	Descrição da Variável	Valor
kmeans.centroids	Quantidade de centróides ("all" - Valor especial, seleciona de 1 a N-1 vizinhos, onde N é o número total de registros)	Iris Plants = 1;60;1 Pima Indians =1;44;1 Breast Cancer =1;50;1
kmeans.initial	Diz como configurar os primeiros centróides ("firstTuples", usa o primeiro registro ou "random", usa valores totalmente aleatórios, dentro dos limites da base)	firstTuples
kmeans.iterations	Numero máximo de iterações (opcional)	1000
kmeans.distance	Escolha do algoritmo que calcula a distância ("euclidian" e/ou "manhattan")	euclidian,manhattan

Tabela 4.9 Configuração do agrupamento com *K-Means*

Algoritmo do *k-Means*: O mesmo implementado por SOARES (2007).

Seleção com Algoritmos Genéticos

As configurações do algoritmo *Algoritmos genéticos* seguiram a escolha do trabalho de COSTA (2005), tendo apenas uma modificação: foi escolhido utilizar a técnica de *Algoritmos genéticos* padrão deixando assim como uma das possibilidades de ampliação desse trabalho a implementação das outras técnicas *Algoritmos genéticos* e suas combinações.

Há também duas informações de configuração, marcadas em negrito logo abaixo, que foram definidas exclusivamente para esse trabalho: número de repetições e número de colunas. Número de repetições serve para que as x gerações configuradas sejam repetidas e dessa forma os resultados fiquem ainda mais bem apurados. Para entender melhor, a primeira repetição vai retornar como resultado um cromossomo de máxima aptidão e este possui alguns bits 1 que representam atributos selecionados, então na repetição seguinte somente são criados cromossomos referentes a esses atributos selecionados no cromossomo resultado da geração anterior. Quer dizer, se o primeiro resultado retornar o cromossomo 1 0 1 0 então na repetição serão criados apenas os cromossomos 1 0, 0 1 e 11 e assim sucessivamente. Nos testes iniciais, foi constatado que a quantidade ideal de repetições é cinco e por isso adotamos esse valor.

A outra nova configuração é a quantidade de colunas, ou atributos, que possibilita fazer o uso de um ou mais atributos no processo de seleção e utilizar a quantidade de atributos que forneceu melhor resultado. Como a seleção com *PCA - Análise de Componentes Principais* está fazendo uso dessa configuração, o mesmo valor foi adotado para a seleção com *Algoritmos genéticos*.

Variável	Definição da Variável	Valor
ag.tamPop	Tamanho da população (qualquer valor inteiro de 1 a infinito)	100
ag.numGers	Quantidade de gerações (qualquer valor inteiro de 1 a infinito)	50
ag.totInd	Total de indivíduos (qualquer valor inteiro de 1 a infinito)	5000
ag.taxaMutacao	Taxa de Mutação (valores reais de 0% a 100%)	0,8
ag.taxaCrossover	Taxa de Crossover (valores reais de 0% a 100%)	65
ag.tipoCrossover	Tipo Crossover (1. Um ponto, 2. Dois pontos, 3. uniforme)	2
ag.tecnicaAG	Técnica AG (8 possibilidades) (*)	1
ag.normalizacao	Normalização linear (true / false)	false
ag.norMin	Valor Máximo da Normalização linear	0

	(qualquer valor inteiro de 1 a 100)	
ag.norMax	Valor Mínimo da Normalização linear (qualquer valor inteiro de 1 a 100)	0
ag.qtdeGap	Steady-State (qualquer valor inteiro de 1 a infinito)	2
ag.qtdeElite	Elitismo (qualquer valor inteiro de 1 a infinito)	2
ag.numRepet	**Número de repetições do AG** **(Qualquer valor inteiro de 1 a infinito)**	**5**
ag.columns	**Quantidade de colunas a serem selecionadas** **("all" - Valor especial, seleciona N-1 colunas, onde N é o número total de colunas)**	**all**

Tabela 4.10 Configuração da seleção com *AG*

(*) *1. padrão, 2. padrão com normalização linear, 3. steady-state, 4. steady-state com normalização linear, 5. elitismo, 6. elitismo com normalização linear, 7. steady-state e elitismo juntas, 8. steady-state e elitismo juntas com normalização linear.*

Algoritmo do *Algoritmos Genéticos*: Adaptado do que foi implementando por FERLIN (2008).

Enquanto a quantidade de execuções for menor ou igual ao número de repetições estabelecido na configuração:

1. Enquanto a quantidade de gerações for menor ou igual ao número de gerações estabelecido na configuração
   - 1.1 Se geração igual a 1
     - 1.1.1 Enquanto a quantidade de população for menor ou igual ao número de indivíduos da população estabelecido na configuração
       - 1.1.1.1 Buscar colunas selecionadas na execução anterior, se for a primeira execução então busca todas as colunas de acordo com a quantidade de colunas na configuração
       - 1.1.1.2 Gerar um indivíduo ou cromossoma:

         Para cada coluna selecionada atribuir 0 ou 1 randomicamente gerando uma string de bits ou genes de tamanho igual a quantidade de colunas selecionadas
       - 1.1.1.3 Calcular a aptidão do indivíduo gerado:

1 – (somatório de x / quantidade de registros total), onde, agrupando os registros por colunas do indivíduo com valor 1 e a coluna classe, x = somatório de grupos com registros iguais – máxima quantidade de registros iguais de todos os grupos

1.2 Se geração maior do que 1

1.2.1 Calcular a probabilidade da população da geração anterior:

Somatório da (aptidão da população da geração anterior / somatório das aptidões de todas as populações)

1.2.2 Escolher randomicamente dois indivíduos da população da geração anterior

Para escolher os dois indivíduos se busca duas vezes o indivíduo gerado há mais tempo e que possui probabilidade maior ou igual a um valor entre 0 e 1 gerado randomicamente

1.2.3 Se a taxa de crossover na configuração for maior ou igual a um percentual gerado randomicamente então efetua o crossover

1.2.3.1 Efetuar o crossover entre os cromossomas dos dois indivíduos escolhidos, pode ser crossover de um ponto ou dois pontos ou uniforme conforme configuração:

Um ponto: gerar um valor randomicamente para ser o ponto. O individuo 1 fica com o início do seu próprio cromossoma até o ponto e com o restante do cromossoma do individuo 2 a partir do ponto. O inverso ocorre com o indivíduo 2

Dois pontos: gerar dois valores randomicamente para serem os pontos 1 e 2. O individuo 1 fica com o início do seu próprio cromossoma até o ponto 1, com o cromossoma do individuo 2 a partir do ponto 1 até o ponto 2 e com o restante do seu próprio cromossoma a partir do ponto 2. O inverso ocorre como indivíduo 2

Uniforme: gerar um valor randomicamente para ser a quantidade de pontos que serão gerados. Se a quantidade de pontos for 1 então será igual ao crossover de um ponto, se for igual a dois será igual ao crossover de dois pontos e se for três ou mais

pontos a lógica será a mesma do crossover de dois pontos só que dividindo os cromossomas até terminar os pontos.

1.2.4 Se a taxa de mutação na configuração for maior ou igual a um percentual gerado randomicamente, então efetue a mutação.

1.2.4.1 Efetuar mutação no cromossoma do indivíduo 1

Gerar um valor randomicamente para ser o ponto de mutação, depois alterar de 0 para 1 ou de 1 para 0 neste ponto

1.2.5 Se a taxa de mutação na configuração for maior ou igual a um outro percentual gerado randomicamente, então efetue a mutação.

1.2.5.1 Efetuar mutação no cromossoma do indivíduo 2

Gerar um valor randomicamente para ser o ponto de mutação, depois alterar de 0 para 1 ou de 1 para 0 neste ponto

1.2.6 Calcular a aptidão dos dois novos indivíduos da geração atual:

1 – (somatório de x / quantidade de registros total), onde, agrupando os registros por colunas do cromossoma com valor 1 e a coluna classe, x = somatório de grupos com registros iguais – máxima quantidade de registros iguais de todos os grupos

2. Buscar melhor indivíduo da execução atual:
   Indivíduo que possui maior aptidão entre toda a população da última geração
3. Buscar as colunas selecionadas na execução anterior, se for a primeira execução então busca todas as colunas de acordo com a quantidade de colunas na configuração
4. Gerar colunas selecionadas nesta execução:
   Apenas as colunas selecionadas anteriormente que possuem valor 1 relacionado ao melhor indivíduo da execução atual

Agrupamento com Redes de Kohonen

As configurações escolhidas para *Redes de Kohonen* são as mesmas de FERLIN (2008), sendo que está marcado em negrito as novas configurações necessárias: a quantidade de centróides, os primeiros centróides e a distância. Todas essas configurações são utilizadas

por SOARES (2007) no algoritmo de agrupamento com *K-Means*, então optamos os mesmos valores que sabemos que deram bons resultados.

Variável	Descrição da Variável	Valor
kohonen.learningRate	Taxa de aprendizado	0.7
kohonen.outputWidth	Largura da saída	2
kohonen.outputHeight	Altura da saída	2
kohonen.epochs	Número de ciclos	1000,5000,10000,15000
kohonen.synapse	Sinapse (wta e/ou gauss)	wta,gauss
kohonen.centroids	**Quantidade de centróides ("all" - Valor especial, seleciona de 2 a N centroides, onde N é o número total de tuplas)**	**Iris Plants = 2;60;1 Pima Indians = 2;44;1 Breast Cancer = 1;50;1**
kohonen.initial	**Diz como configurar os primeiros centróides ("firstTuples", primeira tupla, e/ou "random", valores totalmente aleatórios, dentro dos limites da base)**	**firstTuples**
kohonen.distance	**Escolha do algoritmo que calcula a distância (euclidian e/ou manhattan)**	**euclidian,manhattan**

Tabela 4.11 Configuração do agrupamento com *Redes de Kohonen*

Algoritmo de *Redes de Kohonen:* Adaptado do que foi implementando por FERLIN (2008).

1. Configurar Rede
    - 1.1 Gerar uma matriz dos dados completos, onde valor nulo terá zero.
    - 1.2 Buscar configurações de tipo de sinapse, altura e largura
    - 1.3 Buscar quantidade de neurônios que é igual a quantidade de colunas da base de dados
    - 1.4 Inicializar NeuralNet (biblioteca org.joone.net) com as informações dos itens 1.1, 1.2 e 1.3
2. Executar NeuralNet para gerar sinapses
3. Gerar resultado

3.1 Para cada registro x da base de dados

3.1.1 Inicializar a **melhor sinapse** com o valor 0

3.1.2 Para cada coluna y do registro x da base de dados

3.1.2.1 Se sinapse (x,y) for maior do que a **melhor sinapse**

3.1.2.1.1 **Melhor sinapse** passa a ser a sinapse (x,y)

3.1.2.1.2 **Posição do neurônio da melhor sinapse** passa a ser a posição y da sinapse (x,y)

3.1.3 Calcular altura do registro

*Altura (x) = **Posição do neurônio da melhor sinapse** dividido pela largura configurada*

3.1.4 Calcular largura do registro

*Largura (x) = **Posição do neurônio da melhor sinapse** mod largura configurada*

4. Calcular os centroides

4.1 Buscar quantidade de centróides (k) na configuração

4.2 Inicializar os k centróides

Usando os valores da base de dados gerar uma matrix com valores invertidos, ou seja, os valores de um registro se tornam uma só coluna de vários registros. Isso pode ser feito com FirstTuples ou Random

FirstTuples: Usa os k primeiros registros da base de dados

Random: Usa k registros da base de dados buscados aleatoriamente

4.3 Associar k centroides aos registros mais próximos

4.3.1 Buscar o tipo de distância (Euclidiana ou Manhattan) na configuração

4.3.2 Para cada registro x da base de dados

4.3.2.1 Inicializar **distância mais próxima até o centroide** e **centroide mais próximo** com nulo

4.3.2.2 Para cada centroide z

4.3.2.2.1 Para cada coluna y do centroide z e do registro x

4.3.2.2.1.1 Calcular a **distância (z,y)** entre o valor do centróide z da coluna y e o valor registro x da coluna y da base de dados

*Euclidiana: somatório do [(valor do registro x da coluna y da base de dados – valor do centroide z da coluna w) * 2]*

Manhattan: somatório do [valor do registro x da coluna y da base de dados – valor do centroide z da coluna w], sendo que se o valor da diminuição der negativo multiplica-se por -1

4.3.2.2.2 Se **distância (z,y)** calculada for maior do que **distância mais próxima entre o centroide**

4.3.2.2.2.1 **Distância mais próxima até o centroide** recebe valor da **distância (z,y)** calculada

4.3.2.2.2.2 **Centroide mais próximo** recebe o centroide usado para calcular a **distância (z,y)**

4.3.2.3 Associar o registro x ao **centroide mais próximo**

Bases de Dados

Assim como SOARES (2007), foram utilizadas 105 bases de dados com dados ausentes a partir das bases completas. O módulo *Eraser* do sistema *Appraisal* foi executado 105 vezes para gerar percentuais de ausência de 10%, 20%, 30%, 40% ou 50% para cada por cada atributo de cada base de dados.

- *Iris Plants*

Quantidade de atributos: 4

Total de bases com dados ausentes utilizadas: 20

id	sepallength	sepalwidth	petallength	petalwidth	class
1	5,1	3,5	1,4	0,2	Iris-setosa
2		3	1,4	0,2	Iris-setosa
3	4,7	3,2	1,3	0,2	Iris-setosa
4		3,1	1,5	0,2	Iris-setosa
5	5	3,6	1,4	0,2	Iris-setosa
6	5,4	3,9	1,7	0,4	Iris-setosa
7	4,6	3,4	1,4	0,3	Iris-setosa
8	5	3,4	1,5	0,2	Iris-setosa
9	4,4	2,9	1,4	0,2	Iris-setosa
10	4,9	3,1	1,5	0,1	Iris-setosa

Figura 4.1 Parte da base *iris_mcar_sepallength_10*

id	sepallength	sepalwidth	petallength	petalwidth	class
1		3,5	1,4	0,2	Iris-setosa
2	4,9	3	1,4	0,2	Iris-setosa
3	4,7	3,2	1,3	0,2	Iris-setosa
4		3,1	1,5	0,2	Iris-setosa
5		3,6	1,4	0,2	Iris-setosa
6		3,9	1,7	0,4	Iris-setosa
7		3,4	1,4	0,3	Iris-setosa
8	5	3,4	1,5	0,2	Iris-setosa
9	4,4	2,9	1,4	0,2	Iris-setosa
10		3,1	1,5	0,1	Iris-setosa

Figura 4.2 Parte da base *iris_mcar_sepallength_50*

- *Pima Indians Diabetes*

Quantidade de atributos: 8

Total de bases com dados ausentes utilizadas: 40

id	pregnancy_times	glucose_concentration	blood_pressure	skin_fold_thickness	serum_insulin	body_mass	pedigree_function	age	class
1	6	148	72	35	0	33,6	0,627	50	1
2	1	85	66	29	0	26,6	0,351	31	0
3	8	183	64	0	0	23,3	0,672	32	1
4	1	89	66	23	94	28,1	0,167	21	0
5	0	137	40	35	168	43,1	2,288	33	1
6	5	116	74	0	0	25,6	0,201	30	0
7	3	78	50	32	88	31	0,248	26	1
8	10	115	0	0	0	35,3	0,134	29	0
9	2	197	70	45	543	30,5	0,158	53	1
10	8	125	96	0	0	0		54	1

Figura 4.3 Parte da base *pima_mcar_pedigree_function_10*

id	pregnancy_times	glucose_concentration	blood_pressure	skin_fold_thickness	serum_insulin	body_mass	pedigree_function	age	class
1	6	148	72	35	0	33,6	0,627	50	1
2	1	85	66	29	0	26,6	0,351	31	0
3	8	183	64	0	0	23,3		32	1
4	1	89	66	23	94	28,1	0,167	21	0
5	0	137	40	35	168	43,1		33	1
6	5	116	74	0	0	25,6		30	0
7	3	78	50	32	88	31	0,248	26	1
8	10	115	0	0	0	35,3	0,134	29	0
9	2	197	70	45	543	30,5		53	1
10	8	125	96	0	0	0	0,232	54	1

Figura 4.4 Parte da base *pima_mcar_pedigree_function_50*

- *Wisconsin Breast Cancer*

Quantidade de atributos: 9

Total de bases com dados ausentes utilizadas: 45

id	code_number	Clump_Thickness	Uniformity_of_Cell_Size	Uniformity_of_Cell_Shape	Marginal_Adhesion	Single_Epithelial_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses	Class
1	1000025	5	1	1	1	2	1		1	1	2
2	1002945	5	4	4	5	7	10	3	2	1	2
3	1015425	3	1	1	1	2	2	3	1	1	2
4	1016277	6	8	8	1	3	4	3	7	1	2
5	1017023	4	1	1	3	2	1	3	1	1	2
6	1017122	8	10	10	8	7	10	9	7	1	4
7	1018099	1	1	1	1	2	10	3	1	1	2
8	1018561	2	1	2	1	2	1	3	1	1	2
9	1033078	2	1	1	1	2	1	1	1	5	2
10	1033078	4	2	1	1	2	1	2	1	1	2

Figura 4.5 Parte da base *breast_mcar_Bland_Chromatin_10*

id	code_number	Clump_Thickness	Uniformity_of_Cell_Size	Uniformity_of_Cell_Shape	Marginal_Adhesion	Single_Epithelial_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses	Class
1	1000025	5	1	1	1	2	1	3	1	1	2
2	1002945	5	4	4	5	7	10	3	2	1	2
3	1015425	3	1	1	1	2	2		1	1	2
4	1016277	6	8	8	1	3	4		7	1	2
5	1017023	4	1	1	3	2	1	3	1	1	2
6	1017122	8	10	10	8	7	10	9	7	1	4
7	1018099	1	1	1	1	2	10		1	1	2
8	1018561	2	1	2	1	2	1	3	1	1	2
9	1033078	2	1	1	1	2	1	1	1	5	2
10	1033078	4	2	1	1	2	1	2	1	1	2

Figura 4.6 Parte da base *breast_mcar_Bland_Chromatin_50*

CAPÍTULO 5
ANÁLISE DOS RESULTADOS

Seguindo a proposta descrita no capítulo anterior, novos algoritmos de seleção e agrupamento foram desenvolvidos dentro do sistema Appraisal (SOARES, 2007) para que os resultados fossem gerados com a mesma plataforma, utilizando as mesmas bases de dados e as mesmas estratégias para que a análise e comparação dos resultados fossem justos. O módulo Eraser também foi utilizado, com vistas a gerar os valores ausentes nos originais das bases utilizadas, seguindo a metodologia original daquele trabalho.

5.1 Bases de dados

Conforme já abordado, tendo em vista o caráter continuísta deste trabalho a partir do que foi realiza SOARES (2007) e por isso as mesmas bases de dados foram utilizadas. As três bases de dados existem no repositório da Universidade da Califórnia, Irvine (NEWMAN, 1998): *Iris Plants, Pima Indians Diabetes* e *Wisconsin Breast Cancer*.

Todas essas bases de dados escolhidas são muito utilizadas nos trabalhos de complementação de dados ausentes devido à relação existente entre os atributos e por serem uma representação de dados reais. Este é um dos motivos da escolha dessas bases de dados por SOARES (2007) além de também possuírem somente dados numéricos, e terem um atributo classificador dos registros, informação que não é considerada na imputação mas que é usada no processo de validação para saber se a complementação de dados manteve a relação existente entre os atributos.

As bases de dados escolhidas também possuem um atributo chave, um identificador único, e este não é considerado nos testes por serem irrelevantes ao processo de complementação de dados ausentes.

Iris Plants

A *Iris Plants* é uma base de dados com poucos atributos e bem comportada que foi construída por R. A. Fischer por volta dos anos 30 (WITTEN, FRANK, 2005). Nenhum dos

seus registros possui valor ausente e os 150 registros são distribuídos igualmente entre cada classe, tendo assim 50 representantes por classe.

O atributo que representa a classe contém o tipo de planta que pode ser *Virginica* ou *Versicolor* ou *Setosa*. Os demais atributos são: *sepallength* (comprimento do caule), *sepalwidth* (largura do caule), *petallength* (comprimento das pétalas) e *petalwidth* (largura das pétalas). Seus valores máximo e mínimos nos mostram que se trata de uma base com um domínio baixo de valores para os atributos.

	Tipo	**Unidade**	**Valor Mínimo**	**Valor Máximo**
sepallength	Real	cm	4,3	7,9
sepalwidth	Real	cm	2,0	4,4
petallength	Real	cm	1,0	6,9
petalwidth	Real	cm	0,1	2,5

Tabela 5.1 Valores mínimos e máximos da base *Iris Plants*

Podemos dizer que suas principais características são:

1) Pouca diversidade em cada atributo;
2) Registros de uma classe, a *Setosa*, são independentes dos demais registros das outras classes tendo só atributo que as diferencia das outras, o *petalwidth*. Quando o registro tem o valor igual ou menor que 0,6 no atributo *petalwidth* significa que a classe é *Setosa* independente dos demais atributos.
3) Forte correlação de seus atributos, conforme indicado abaixo:

	sepallength	***sepalwidth***	***petallength***	***petalwidth***
sepallength	1,00	- 0,11	0,82	0,87
sepalwidth	- 0,11	1,00	- 0,36	- 0,42
petallength	0,82	- 0,36	1,00	0,96
petalwidth	0,87	- 0,42	0.96	1,00

Tabela 5.2 Matriz correlação da base *Iris Plants*

A baixa correlação entre o atributo *sepalwidth* e os outros mostram que seus valores não estão relacionados aos valores do outros atributos.

Pima Indians Diabetes

A base de dados *Pima Indians Diabetes* contém informações sobre a saúde de mulheres com no mínimo 21 anos e que são da tribo Pima do Arizona, EUA. Os atributos são:

1: *pedigree_function* (características de diabetes)

2: *glucose_concentration* (glicose no sangue)

3: *body_mass* (peso/altura)

4: *skin_fold_thickness* (espessura da pele do tríceps)

5: *blood_pressure* (pressão)

6: *age* (idade)

7: *serum_insulin* (insulina no sangue)

8: *pregnancy_times* (quantidade de gestação)

Há um total de 768 registros nessa base, sendo que 500 são referentes a pacientes que não possuem diabetes e os 268 restantes possuem. Em sua descrição, a base está sem valores ausentes, entretanto encontra-se vários valores zerados em atributos que deveriam possuir valor diferente de zero. SOARES (2007) optou por retirar todos os registros com valor zero em qualquer atributo com exceção do atributo *pregnancy_times* e assim gerar uma base de dados com apenas 392 registros, mas aqui optamos por utilizar a base completa.

	Valor Mínimo	**Valor Máximo**
Age	21	81
blood_pressure	0	122
body_mass	0	67,1
glucose_concentration	0	199
pedigree_function	0,078	2,42
pregnancy_times	0	17
serum_insulin	0	846
skin_fold_thickness	0	99

Tabela 5.3 Valores mínimos e máximos base *Pima Indians Diabetes*

E vê-se logo abaixo sua matriz de correlação indicada pelo número do atributo:

	1	2	3	4	5	6	7	8
1	1.00	0.14	0.14	0.18	0.04	0.03	0.19	-0.03
2	0.14	1.00	0.22	0.06	0.15	0.26	0.33	0.13
3	0.14	0.22	1.00	0.39	0.28	0.04	0.20	0.02
4	0.18	0.06	0.39	1.00	0.21	-0.11	0.44	-0.08
5	0.04	0.15	0.28	0.21	1.00	0.24	0.09	0.14
6	0.03	0.26	0.04	-0.11	0.24	1.00	-0.04	0.54
7	0.19	0.33	0.20	0.44	0.09	-0.04	1.00	-0.07
8	-0.03	0.13	0.02	-0.08	0.14	0.54	-0.07	1.00

Tabela 5.4 Matriz correlação da base *Iris Plants*

Diferentemente da base *Iris Plants*, a correlação entre atributos é baixa, ou seja, eles são muito independentes. Apenas alguns poucos atributos chegam quase a 55% de correlação, o que pode tornar mais difícil o processo de complementação de dados com ausência completamente aleatória.

Wisconsin Breast Cancer

Uma base de dados que armazena informações de pacientes de câncer de mama do hospital da Universidade de Wisconsin foi doada para o repositório UCI, se tornando assim a *Wisconsin Breast Cancer* (MANGASARIAN, WOLBERG, 1990). Seus atributos são:

1: *Uniformity_of_Cell_Size*

2: *Clump_Thickness*

3: *Bland_Chromatin*

4: *Uniformity_of_Cell_Shape*

5: *Marginal_Adhesion*

6: *Mitoses*

7: *Bare_Nuclei*

8: *Normal_Nucleoli*

9: *Single_Epithelial_Cell_Size*

Tendo um total de 699 registros nesta base, onde 17 já possuem dados ausentes, optou-se por usar a base completa sem retirar esses dados ausentes iniciais.

	Valor Mínimo	Valor Máximo
Bare_Nuclei	0	10
Bland_Chromatin	1	10
Clump_Thickness	1	10
Marginal_Adhesion	1	10
Mitoses	1	10
Normal_Nucleoli	1	10
Single_Epithelial_Cell_Size	1	10
Uniformity_of_Cell_Shape	1	10
Uniformity_of_Cell_Size	1	10

Tabela 5.5 Valores mínimos e máximos da base *Wisconsin Breast Cancer*

E vê-se logo abaixo sua matriz de correlação indicada pelo número do atributo:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	1.00	0.64	0.76	0.91	0.71	0.46	0.68	0.72	0.75
2	0.64	1.00	0.56	0.65	0.49	0.35	0.59	0.54	0.52
3	0.76	0.56	1.00	0.74	0.67	0.34	0.67	0.67	0.62
4	0.91	0.65	0.74	1.00	0.68	0.44	0.70	0.72	0.72
5	0.71	0.49	0.67	0.68	1.00	0.42	0.67	0.60	0.60
6	0.46	0.35	0.34	0.44	0.42	1.00	0.34	0.43	0.48
7	0.68	0.59	0.67	0.70	0.67	0.34	1.00	0.57	0.58
8	0.72	0.54	0.67	0.72	0.60	0.43	0.57	1.00	0.63
9	0.75	0.52	0.62	0.72	0.60	0.48	0.58	0.63	1.00

Tabela 5.6 Matriz correlação da base *Wisconsin Breast Cancer*

Assim como a base *Iris Plants*, a correlação entre atributos da base *Wisconsin Breast Cancer* é forte, pois que oito dos nove atributos possuem valores de correlação maior do que 50%. O atributo com mais baixa correlação é o *Mitoses* (menor do que 0.5) enquanto os atributos com mais alta correlação são os *Uniformity_of_Cell_Size* e *Uniformity_of_Cell_Shape* (0.91).

5.2 Resultados por base de dados

Após utilizar o módulo *Eraser* do sistema *Appraisal* para gerar 105 bases de dados com percentuais diferentes de ausência para cada atributo separadamente, foi então executado o módulo *Crowner* utilizando cinco estratégias de imputação composta de 13 técnicas diferentes, quer dizer 37 combinações de técnicas, para complementar os dados ausentes em todas as 105 bases. Chamamos essas combinações de técnicas da imputação composta listadas a seguir de *Estratégia Proposta mais Soares*:

1. Imputação com *Média*
2. Agrupamento com *K-Means* seguido de imputação com *Média*
3. Agrupamento com *Redes Kohonen* seguido de imputação com *Média*
4. Seleção com *PCA* seguido de agrupamento com *K-Means* seguida de imputação com *Média*
5. Seleção com *PCA* seguido de agrupamento com *K-Means* seguida de imputação com *Média*
6. Seleção com *AG* seguido de agrupamento com *Redes Kohonen* seguida de imputação com *Média*
7. Seleção com *AG* seguido de agrupamento com *Redes Kohonen* seguida de imputação com *Média*
8. Agrupamento com *K-Means* seguido de seleção com *PCA* seguida de imputação com *Média*
9. Agrupamento com *K-Means* seguido de seleção com *AG* seguida de imputação com *Média*
10. Agrupamento com *Redes Kohonen* seguido de seleção com *PCA* seguida de imputação com *Média*
11. Agrupamento com *Redes Kohonen* seguido de seleção com *AG* seguida de imputação com *Média*
12. Imputação com *k-NN*
13. Seleção com *PCA* seguida de imputação com *k-NN*
14. Seleção com *AG* seguida de imputação com *k-NN*
15. Agrupamento com *K-Means* seguido de imputação com *k-NN*
16. Agrupamento com Redes Kohonen seguido de imputação com *k-NN*

17. Seleção com *PCA* seguido de agrupamento com *K-Means* seguida de imputação com *k-NN*
18. Seleção com *PCA* seguido de agrupamento com *K-Means* seguida de imputação com *k-NN*
19. Seleção com *AG* seguido de agrupamento com *Redes Kohonen* seguida de imputação com *k-NN*
20. Seleção com *AG* seguido de agrupamento com *Redes Kohonen* seguida de imputação com *k-NN*
21. Agrupamento com *K-Means* seguido de seleção com *PCA* seguida de imputação com *k-NN*
22. Agrupamento com *K-Means* seguido de seleção com *AG* seguida de imputação com *k-NN*
23. Agrupamento com *Redes Kohonen* seguido de seleção com *PCA* seguida de imputação com *k-NN*
24. Agrupamento com *Redes Kohonen* seguido de seleção com *AG* seguida de imputação com *k-NN*
25. Imputação com *Back Propagation*
26. Seleção com *PCA* seguida de imputação com *Back Propagation*
27. Seleção com *AG* seguida de imputação com *Back Propagation*
28. Agrupamento com *K-Means* seguido de imputação com *Back Propagation*
29. Agrupamento com *Redes de Kohonen* seguido de imputação com *Back Propagation*
30. Seleção com *PCA* seguido de agrupamento com *K-Means* seguida de imputação com *Back Propagation*
31. Seleção com *PCA* seguido de agrupamento com *K-Means* seguida de imputação com *Back Propagation*
32. Seleção com *AG* seguido de agrupamento com *Redes de Kohonen* seguida de imputação com *Back Propagation*
33. Seleção com *AG* seguido de agrupamento com *Redes de Kohonen* seguida de imputação com Back Propagation
34. Agrupamento com *K-Means* seguido de seleção com *PCA* seguida de imputação com *Back Propagation*

35. Agrupamento com *K-Means* seguido de seleção com *AG* seguida de imputação com *Back Propagation*
36. Agrupamento com *Redes de Kohonen* seguido de seleção com *PCA* seguida de imputação com *Back Propagation*
37. Agrupamento com *Redes de Kohonen* seguido de seleção com *AG* seguida de imputação com *Back Propagation*

Por fim, o módulo *Reviewer* do *Appraisal* foi usado para gerar gráficos que indicam quais estratégias foram vencedoras (com menor taxa de erro) seja por atributo, percentual de ausência ou base de dados. Mas já que o uso de Comitê não foi adotado nesse trabalho, então o módulo *Committee* não foi executado e por isso sempre aparecerá zerado nos gráficos.

A análise dos resultados a partir dos gráficos está desenvolvida de três maneiras para cada base de dados: primeiro avaliamos de uma forma geral quantas vezes cada combinação de técnica venceu, em seguida analisamos quantas vezes cada combinação de técnica venceu por percentual de ausência e por fim analisamos, dentre a combinação de técnica vencedora, qual combinação de técnicas venceu analisando gráficos por atributo e percentual de ausência.

5.2.1 Gráficos: Técnicas por base de dados

Com a análise de técnicas ou combinação de técnicas por base de dados teremos uma visão geral de qual combinação de técnica foi a melhor para preencher os dados ausentes, se trata de uma contagem total de combinações vencedoras de cada base de dados com percentuais de ausência diferentes.

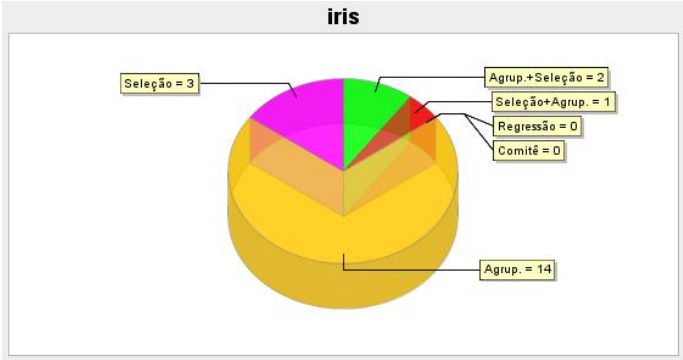
Assim podemos perceber que a base *Iris Plants* teve a técnica de agrupamento como melhor resultado em 70% dos casos, e isso ocorreu possivelmente por causa da alta correlação entre três dos quatro atributos desta base.

Similarmente, a base *Pima Indians Diabetes* obteve um percentual muito próximo a base *Iris Plants* na técnica de agrupamento, 67,5%. Diferentemente de SOARES (2007) em que a técnica de agrupamento teve sucesso em apenas 32,5% na base *Pima Indians Diabetes* e que a técnica de seleção combinada com agrupamento obteve bom resultado em 52% dos casos. Mais adiante iremos avaliar a importância da técnica de agrupamento Redes de Kohonen no aumento do percentual de eficácia para a técnica de agrupamento em comparação com SOARES (2007).

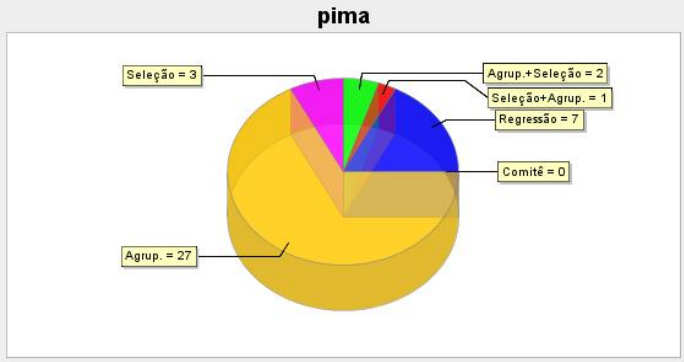
Na base *Wisconsin Breast Cancer* o agrupamento também obteve melhores resultados só que em mais de 90% dos casos, entretanto a taxa de erro foi alto nos melhores casos, o que veremos nos gráficos posteriores. Enquanto em SOARES (2007) a técnica de agrupamento empatou com a técnica de agrupamento seguindo de seleção.

Nesta simples análise vemos que a redução dos registros de dados através do agrupamento antes da imputação de dados é de grande importância na qualidade do processo de complementação de dados ausentes.

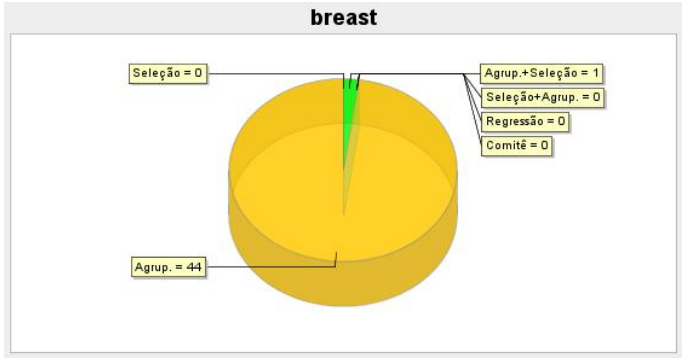
Iris Plants



Pima Indians Diabetes



Wisconsin Breast Cancer



5.2.2 Gráficos: Técnicas por percentual de ausência

Com a análise de técnicas ou combinações de técnicas por percentual de ausência teremos uma visão de qual combinação foi a melhor para preencher os dados ausentes em cada base de dados com dados ausentes de percentuais diferentes.

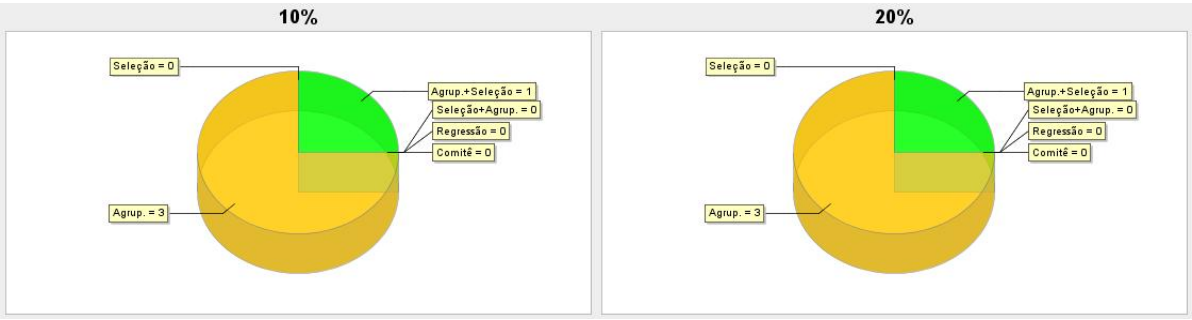
Nessa análise busca-se verificar se a quantidade de dados ausentes por coluna de cada base de dados impacta nos resultados das combinações de técnicas. Pois quanto maior o percentual de dados ausentes, menos informações poderão ser utilizadas pelas técnicas para complementação dos dados e imaginamos que a imputação com *Back Propagation* e *k-NN* podem sofrer impactos. Além disso, o tamanho dos grupos do agrupamento com *K-Means* e *Redes de Kohonen* podem ser significantemente alterados.

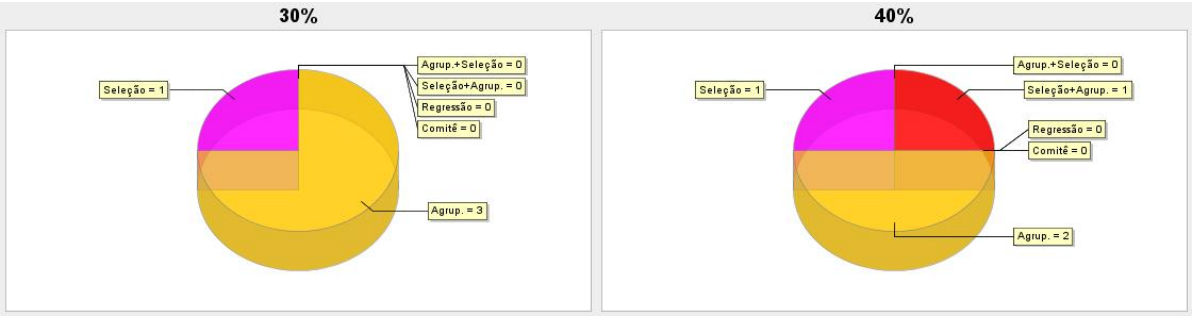
Mas com os gráficos vemos que a base *Iris Plants* não foi impactada, em todos os cinco percentuais de ausência o agrupamento foi considerado a melhor técnica. Possivelmente os resultados constantes se devem as características dessa base de dados: pouca variação entre os valores, alta correlação entre grande parte dos atributos e poucos registros.

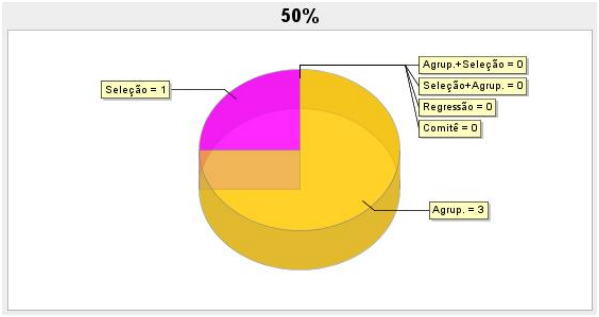
Em todos percentuais o agrupamento foi o vencedor na base *Pima Indians Diabetes*. Em SOARES (2007) a técnica de seleção fez importante papel em todos os percentuais de ausência desta base de dados, entretanto os resultados analisados aqui contam não só com a técnica de agrupamento *k-Means* usado por SOARES (2007) como também Redes de Kohonen, o que certamente fez a diferença nos resultados apresentados.

Em todos os percentuais de ausência da base *Wisconsin Breast Cancer* o agrupamento foi muito melhor do que as outras combinações de técnicas. E novamente ressaltamos esse ótimo desempenho do agrupamento o vendo como uma importante técnica a ser processada antes da complementação de dados ausentes.

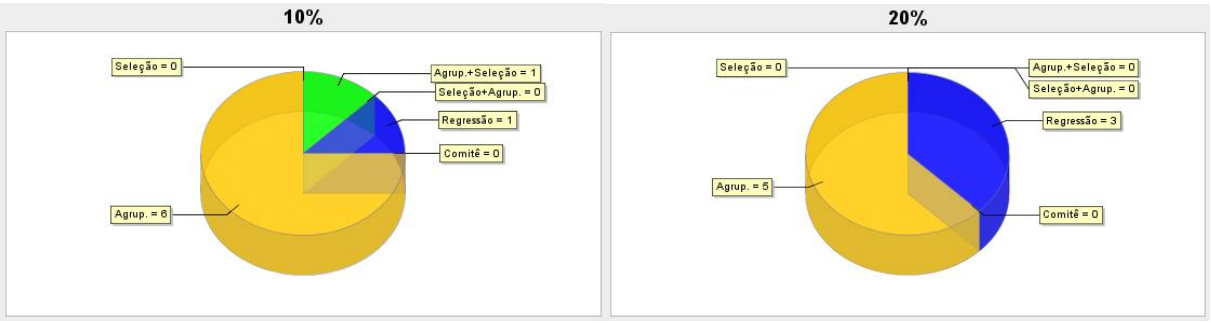
Iris Plants

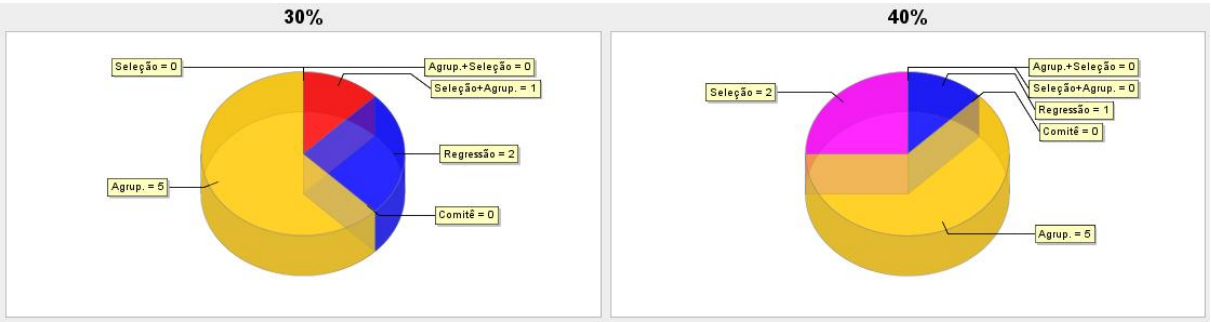

10%
Seleção = 0
Agrup.+Seleção = 1
Seleção+Agrup. = 0
Regressão = 0
Comitê = 0
Agrup. = 3
20%
Seleção = 0
Agrup.+Seleção = 1
Seleção+Agrup. = 0
Regressão = 0
Comitê = 0
Agrup. = 3

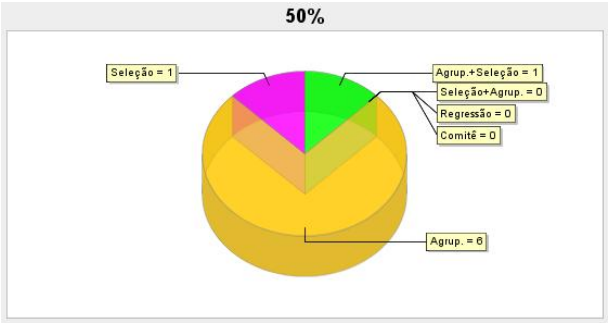

30%
Seleção = 1
Agrup.+Seleção = 0
Seleção+Agrup. = 0
Regressão = 0
Comitê = 0
Agrup. = 3
40%
Seleção = 1
Agrup.+Seleção = 0
Seleção+Agrup. = 1
Regressão = 0
Comitê = 0
Agrup. = 2


50%
Seleção = 1
Agrup.+Seleção = 0
Seleção+Agrup. = 0
Regressão = 0
Comitê = 0
Agrup. = 3

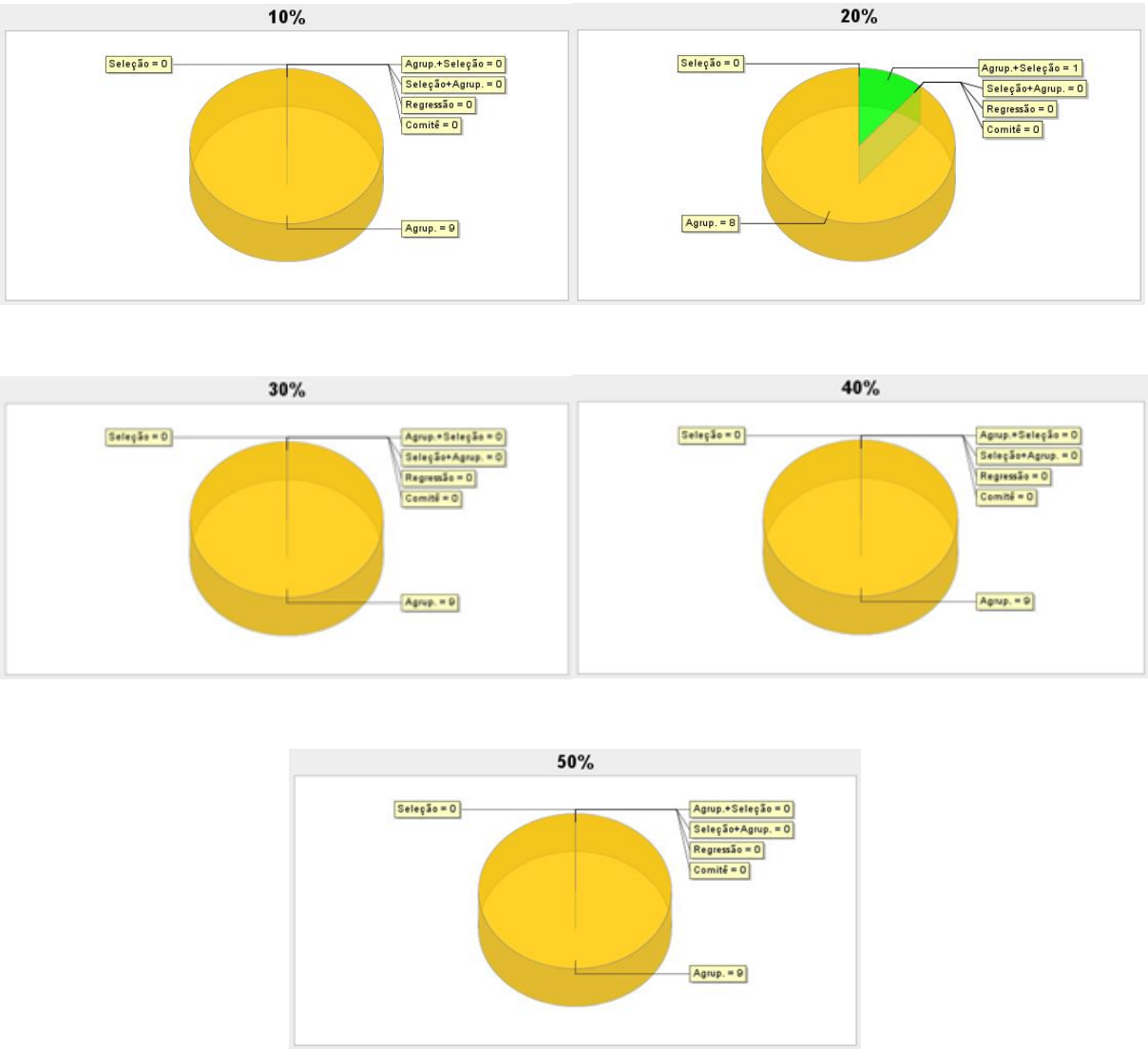
Pima Indians Diabetes







Wisconsin Breast Cancer



5.2.3 Gráficos: Técnicas por ranking

Nessa análise saberemos quais foram as melhores combinações de técnicas dentre as técnicas ou combinações de técnicas utilizadas ao avaliar os gráficos de rankeamento que mostram percentuais de colocação para cada combinação estratégica.

A partir dos gráficos logo vemos que a simples imputação com *Média* foi a última colocada na base *Iris Plants*, apresentando maiores casos de erro. Já a técnica de agrupamento com *K-Means* seguida de imputação com *k-NN* foi a vencedora em disparada, as outras técnicas também tiveram bom desempenho, como *K-Means* seguida de imputação com *Média* e *Kohonen* seguida de imputação com *k-NN.* Algumas combinações de técnicas de seleção

com agrupamento tiveram resultados satisfatórios sendo que *Kohonen* aparece em quatro técnicas das cinco com melhores resultados.

Analisando os gráficos da base *Pima Indians Diabetes* vemos que a *Média* também foi a última colocada porém não foi em disparada pois que as outras técnicas tiveram um desempenho geral regular. Novamente as técnicas envolvendo agrupamento antes da imputação tiveram bons resultados e em todos os melhores casos a imputação com *k-NN* se sobressaiu.

Já a base *Wisconsin Breast Cancer* obteve mais de uma técnica com pior resultado e em nenhum delas a *Média* se encontra. Entretanto novamente ressaltamos a influência das Redes de Kohonen nos melhores resultados, pois essa técnica aparece em quatro das seis técnicas com melhores resultados da base *Wisconsin Breast Cancer*.

Assim como SOARES (2007) apresentou que a imputação com *Média* gera resultados ruins para todas as três bases de dados, vemos que ao utilizar mais duas técnicas, *AG* para seleção e *Kohonen* para agrupamento, essa conclusão permaneceu firme para as bases *Iris Plants* e *Pima Indians Diabetes*.

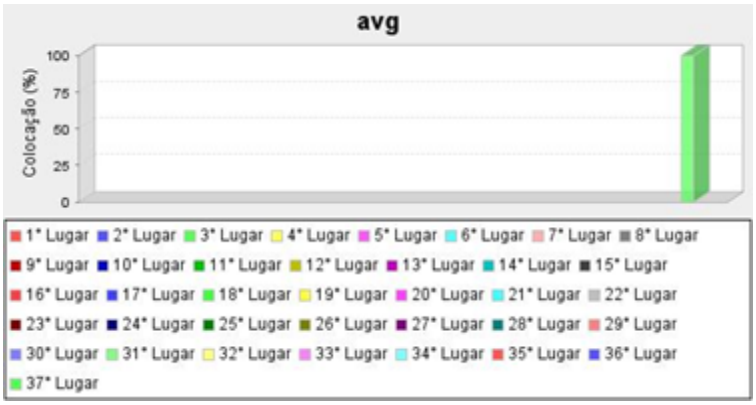
As técnicas de agrupamento seguida de imputação tiveram melhor resultado geral em todas as três bases de dados. E o que se destaca é o agrupamento seguido de imputação com *k-NN* nas bases *Iris Plants* e *Wisconsin Breast Cancer*, isso porque a correlação entre os atributos dessas duas bases é alta. A base de dados *Pima Indians Diabetes* também obteve bons resultados com essa técnica, o que indica que a baixa correlação entre os atributos favorece o uso da imputação simples com *k-NN* nesta base.

E de forma unânime Redes de Kohonen esteve presente na maioria das melhores técnicas sendo de fundamental importância para o aumento do percentual de colocação da técnica de agrupamento seguida de imputação.

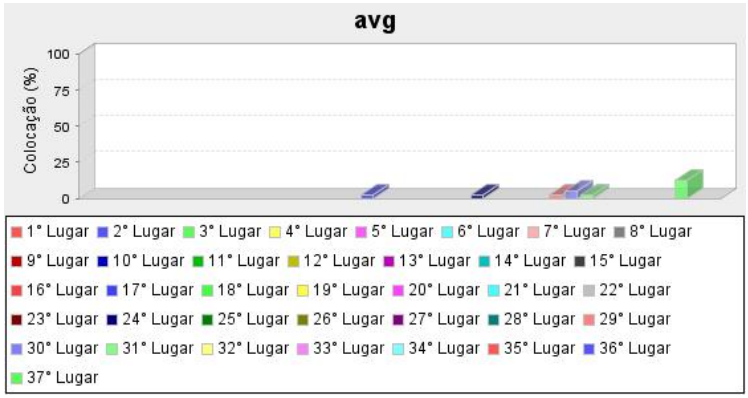
Estratégia 1: Imputação com *Média*

Os resultados não foram satisfatórios para esta estratégia, pois nenhuma base alcançou o 1º lugar ranking, pelo contrário, a base *Iris Plants* ficou com 100% da última colocação. O melhor resultado foi na base *Wisconsin Breast Cancer* com 10% no 4º lugar.

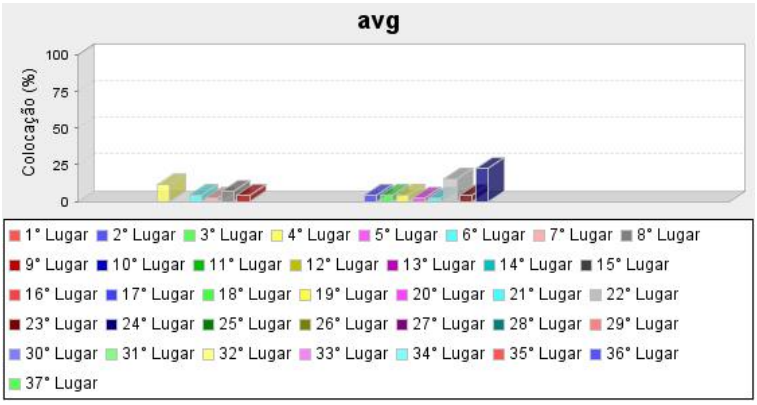
Iris Plants



Pima Indians Diabetes



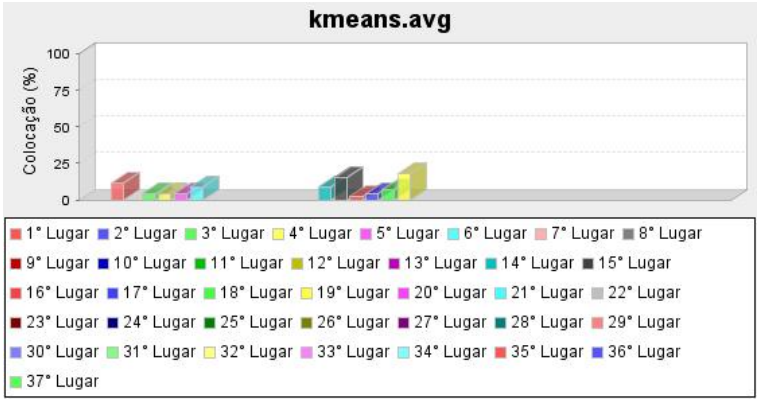
Wisconsin Breast Cancer



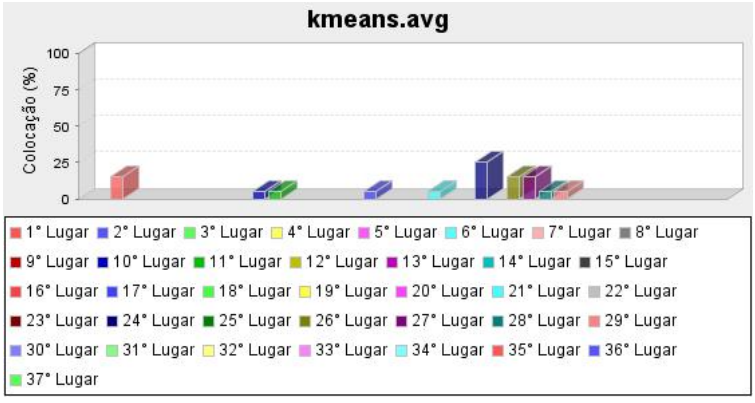
Estratégia 2: Agrupamento com *K-Means* e Imputação com *Média*

Os resultados foram bons para esta estratégia que conseguiu alcançar os primeiros lugares do ranking em quase todas as bases de dados, as bases *Iris Plants* e *Pima Indians Diabetes* ficaram com 10% de colocação no 1º lugar.

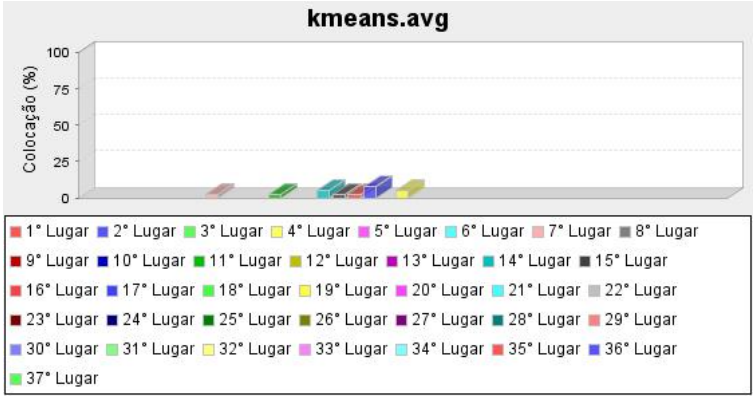
Iris Plants



Pima Indians Diabetes



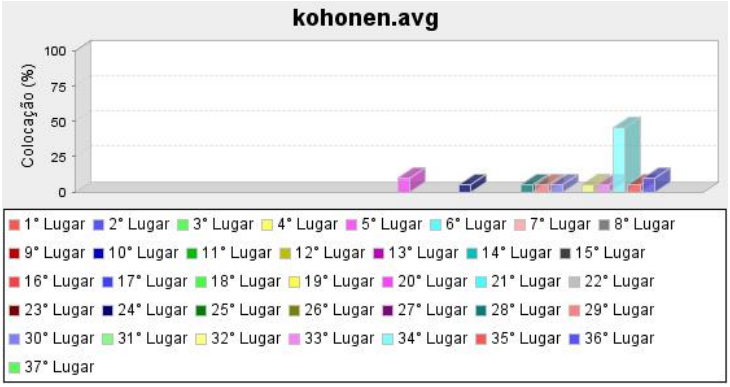
Wisconsin Breast Cancer



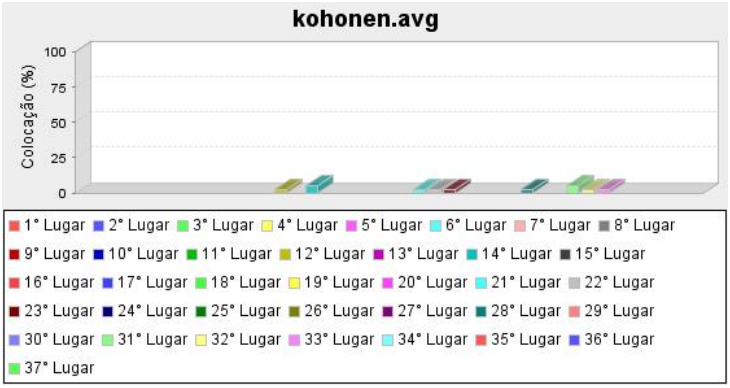
Estratégia 3: Agrupamento com *Redes de Kohonen* e Imputação com *Média*

Os resultados foram muito regulares para esta estratégia que se apresentou próxima aos primeiros lugares do ranking apenas na base *Wisconsin Breast Cancer* alcançando boas colocações, 1º e 2º lugar, entretanto nas bases *Iris Plants* e *Pima Indians Diabetes* ficaram mais próximas dos últimos lugares.

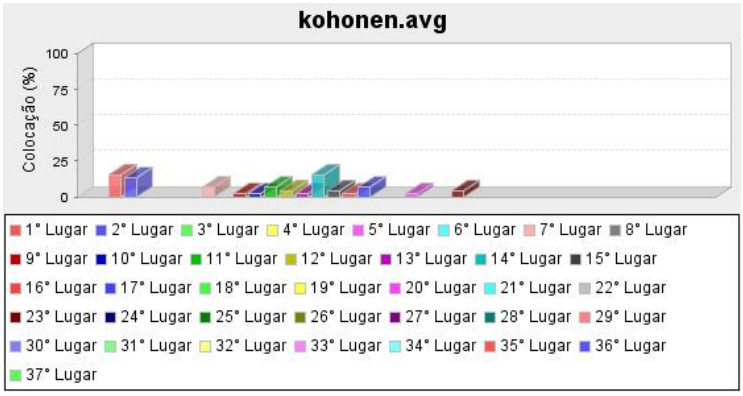
Iris Plants



Pima Indians Diabetes



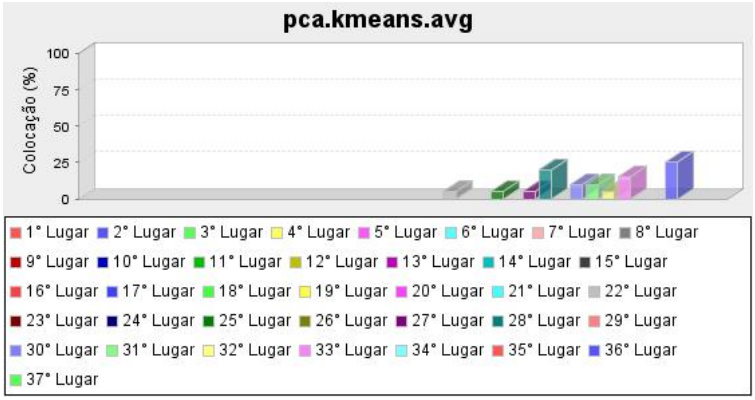
Wisconsin Breast Cancer



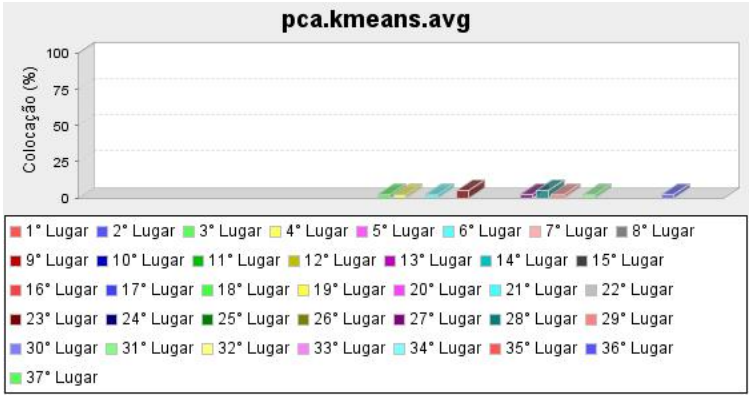
Estratégia 4: Seleção com *PCA*, Agrupamento com *K-Means* e Imputação com *Média*

Os resultados não foram satisfatórios para esta estratégia que se apresentou nos últimos lugares do ranking nas bases *Iris Plants* e *Pima Indians Diabetes*, sendo que o melhor resultado foi na base *Wisconsin Breast Cancer* com o 6ª lugar.

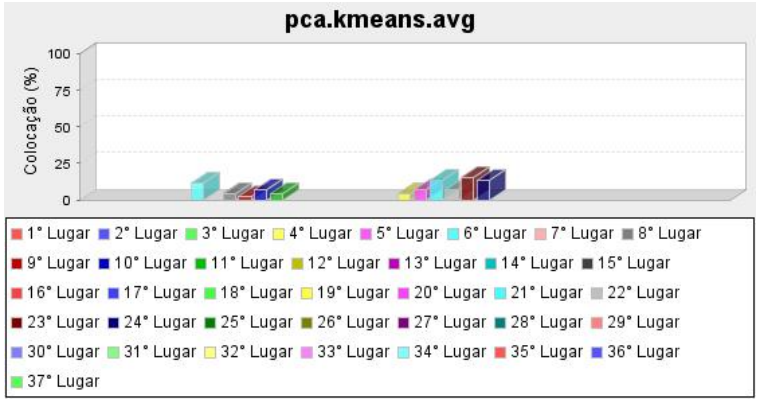
Iris Plants



Pima Indians Diabetes



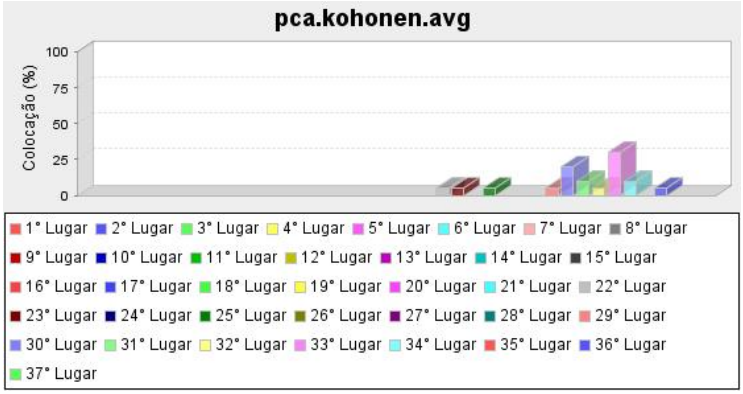
Wisconsin Breast Cancer



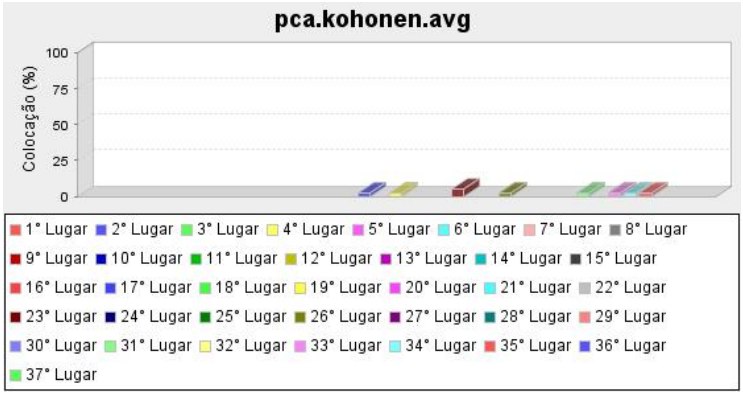
Estratégia 5: Seleção com PCA, Agrupamento com *Redes de Kohonen* e Imputação com *Média*

Os resultados não foram satisfatórios para esta estratégia que se apresentou no máximo no 2º lugar do ranking apenas na base *Wisconsin Breast Cancer* enquanto nas bases *Iris Plants* e *Pima Indians Diabetes* a colocação ficou mais próxima dos últimos lugares.

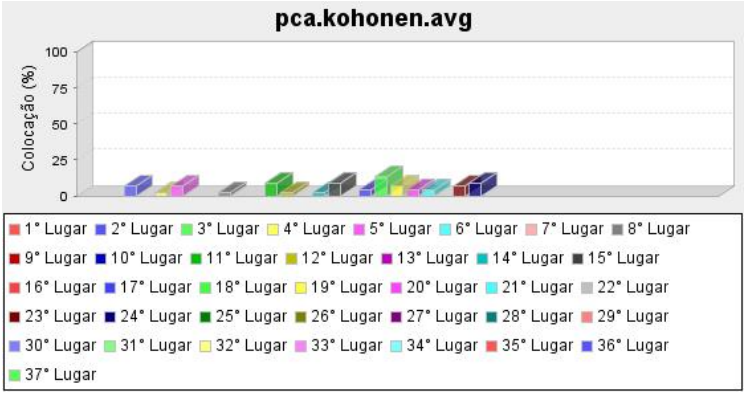
Iris Plants



Pima Indians Diabetes



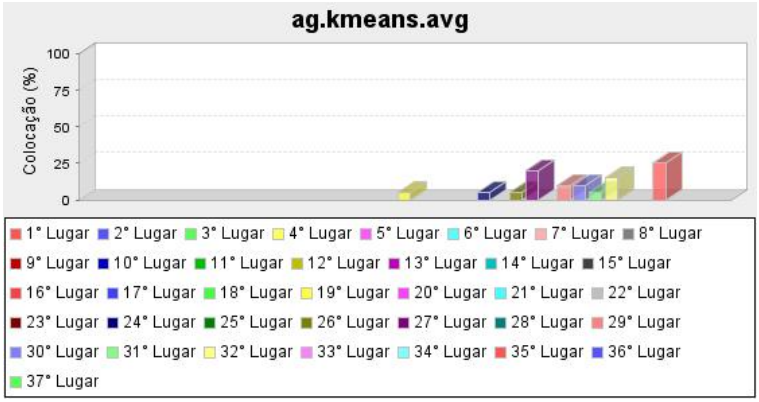
Wisconsin Breast Cancer



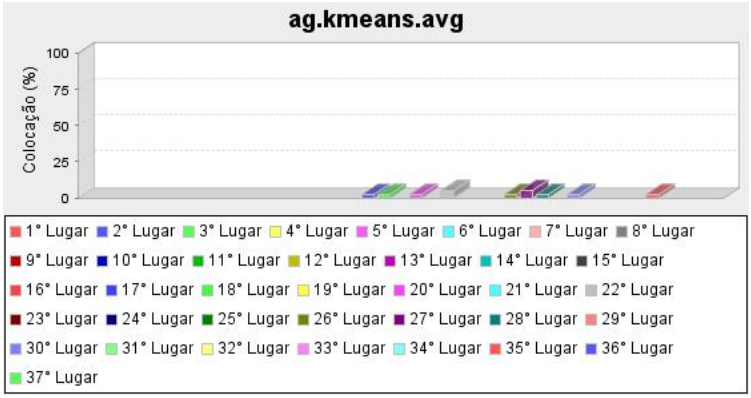
Estratégia 6: Seleção com *AG*, Agrupamento com K-Means e Imputação com *Média*

Os resultados não foram satisfatórios para esta estratégia, pois nenhuma base alcançou o 1º lugar ranking e duas bases ficaram próximas as últimas colocações. O melhor resultado foi na base *Wisconsin Breast Cancer* alcançando apenas o 5º lugar.

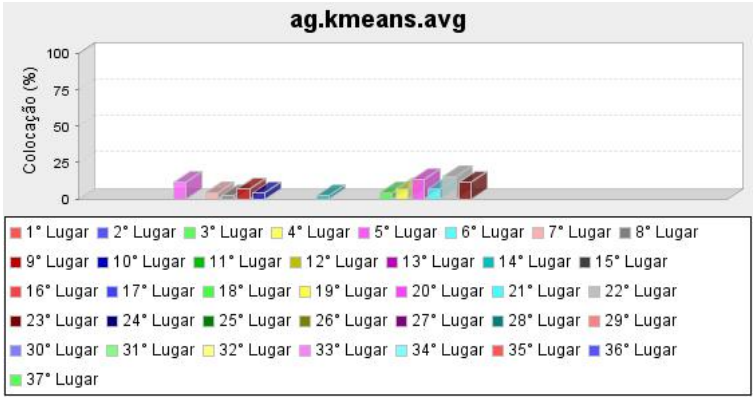
Iris Plants



Pima Indians Diabetes



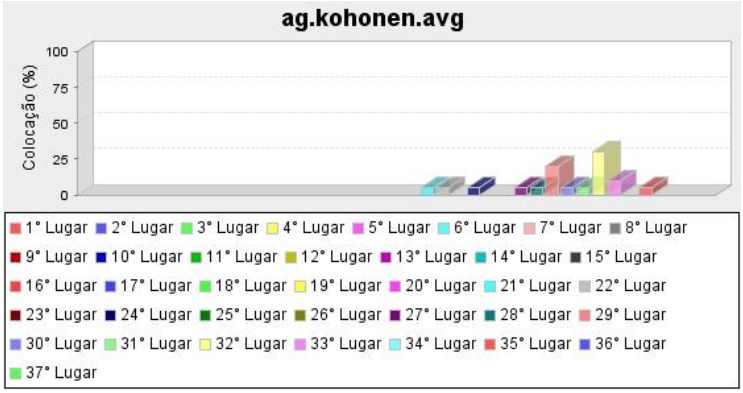
Wisconsin Breast Cancer



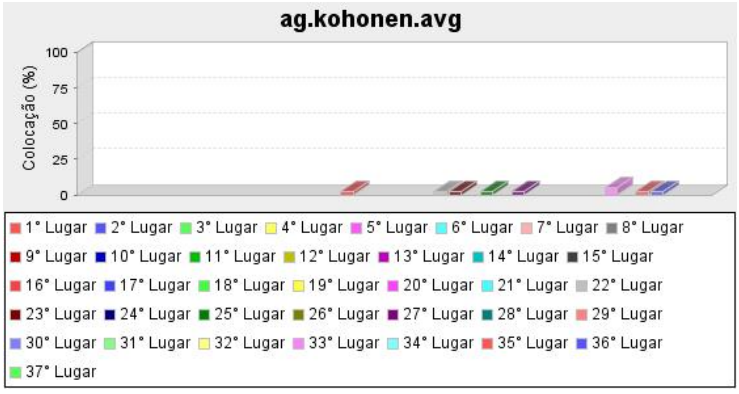
Estratégia 7: Seleção com *AG*, Agrupamento com *Redes de Kohonen* e Imputação com *Média*

Os resultados foi muito regular para esta estratégia que se apresentou no 1º lugar do ranking apenas na base *Wisconsin Breast Cancer* enquanto nas bases *Iris Plants* e *Pima Indians Diabetes* a colocação ficou mais próxima dos últimos lugares.

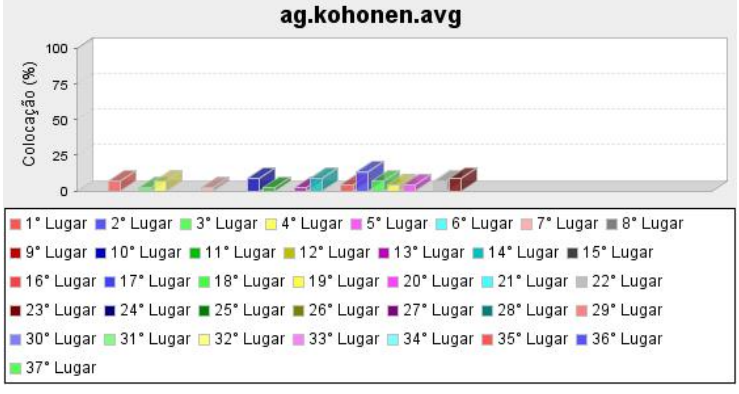
Iris Plants



Pima Indians Diabetes



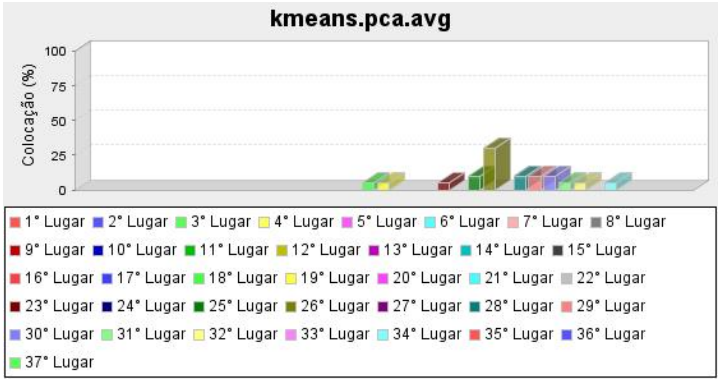
Wisconsin Breast Cancer



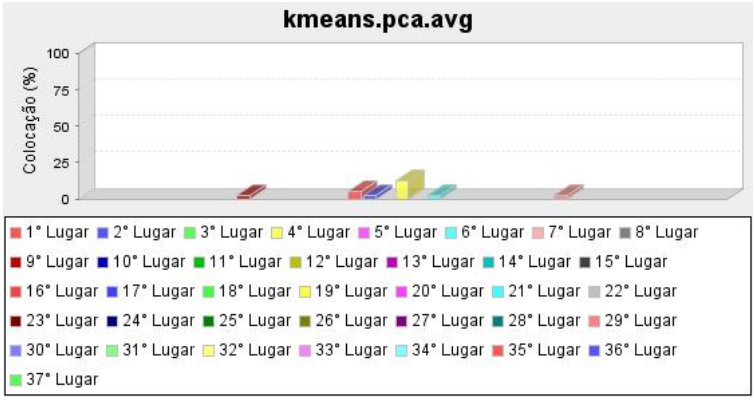
Estratégia 8: Agrupamento com *K-Means*, Seleção com *PCA* e Imputação com *Média*

Os resultados não foram satisfatórios para esta estratégia que se apresentou nos últimos lugares do ranking na bases *Iris Plants*, já as bases *Pima Indians Diabetes* e *Wisconsin Breast Cancer* conseguiram o 9º e 3º lugar respectivamente.

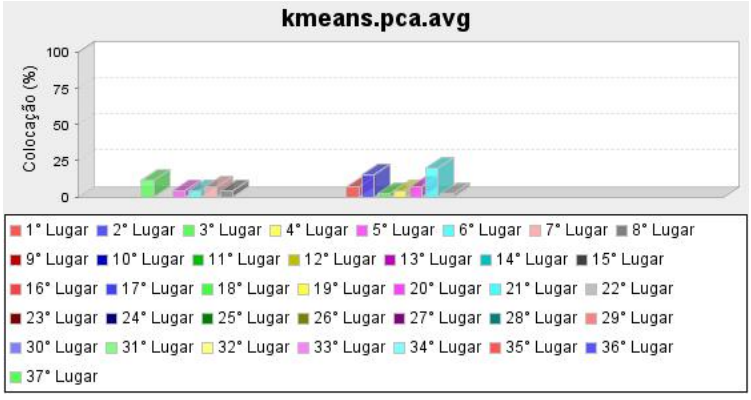
Iris Plants



Pima Indians Diabetes



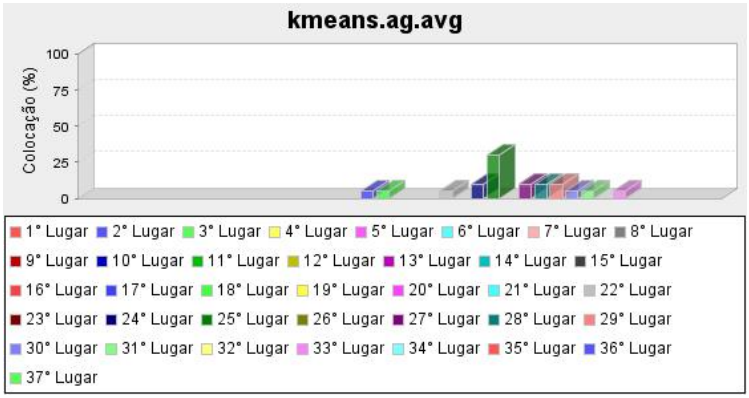
Wisconsin Breast Cancer



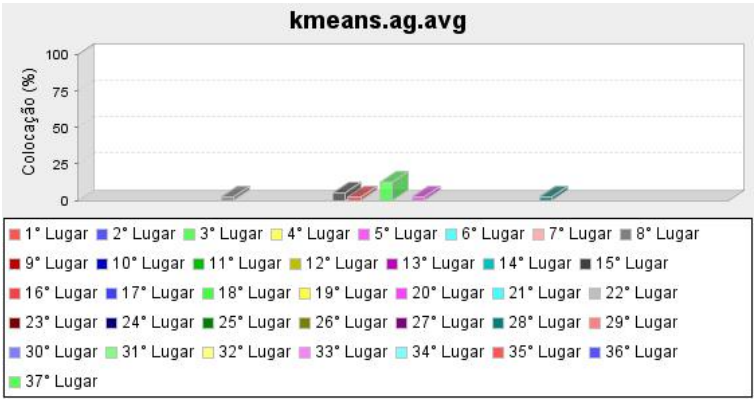
Estratégia 9: Agrupamento com *K-Means*, Seleção com *AG* e Imputação com *Média*

Os resultados não foram satisfatórios para esta estratégia, pois nenhuma base alcançou o 1º lugar ranking. O melhor resultado foi na base *Wisconsin Breast Cancer* com 10% no 2º lugar.

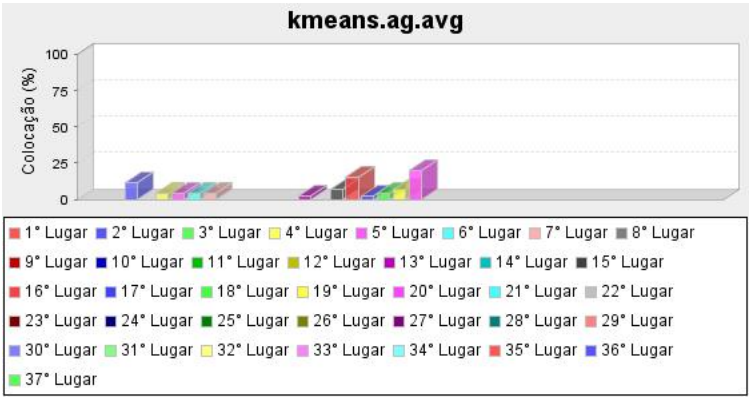
Iris Plants



Pima Indians Diabetes



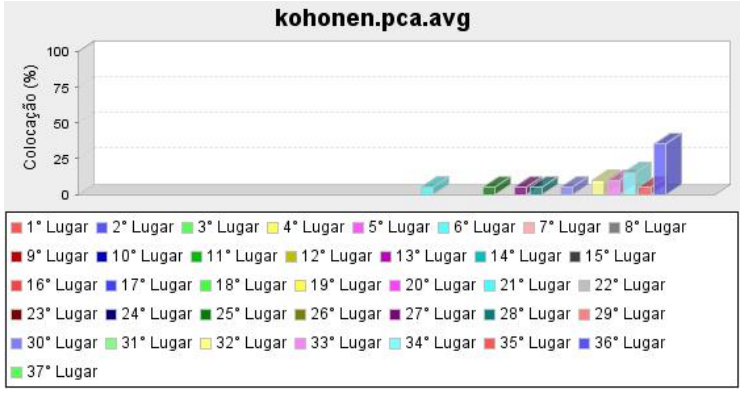
Wisconsin Breast Cancer



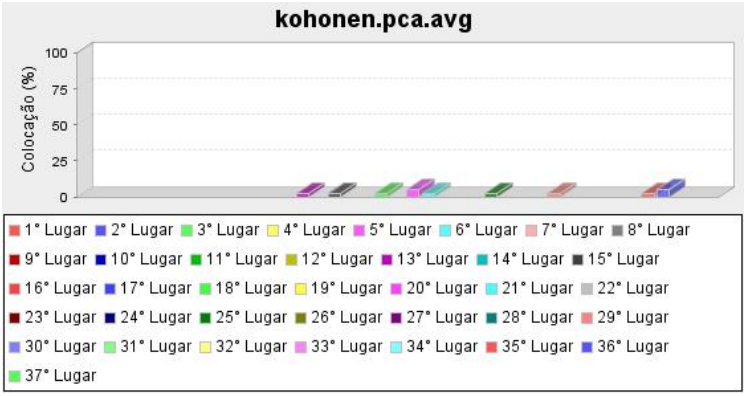
Estratégia 10: Agrupamento com *Redes de Kohonen*, Seleção com *PCA* e Imputação com *Média*

Os resultados não foram satisfatórios para esta estratégia que se apresentou no máximo no 3º lugar do ranking apenas na base *Wisconsin Breast Cancer* enquanto nas bases *Iris Plants* e *Pima Indians Diabetes* a colocação ficou mais próxima dos últimos lugares.

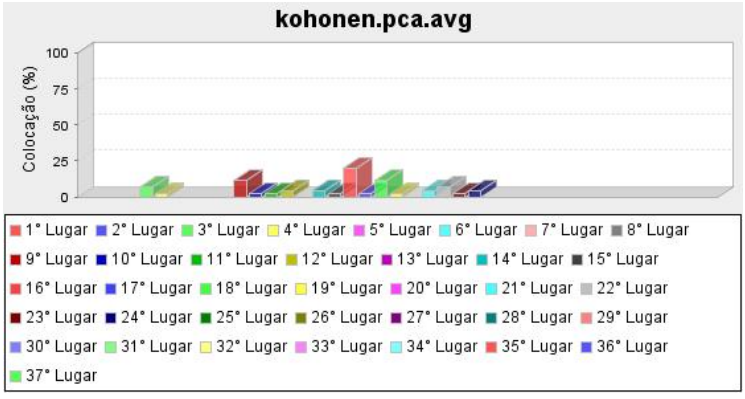
Iris Plants



Pima Indians Diabetes



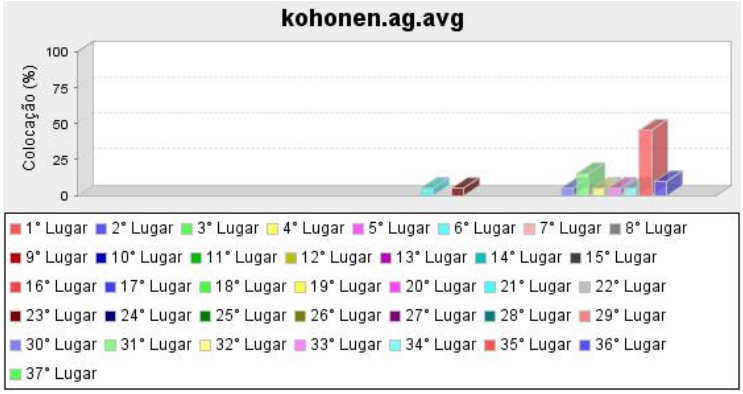
Wisconsin Breast Cancer



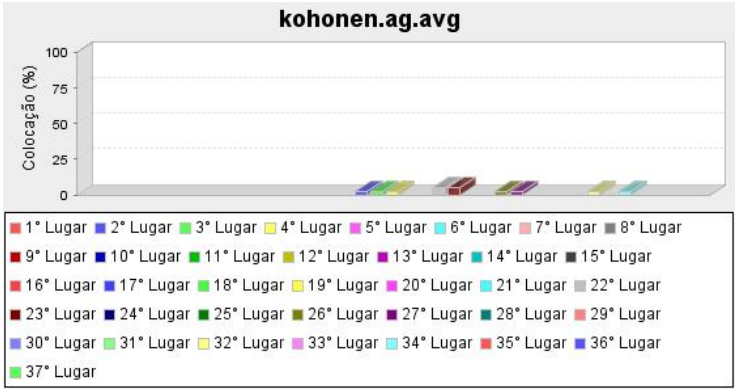
Estratégia 11: Agrupamento com *Redes de Kohonen*, Seleção com *AG* e Imputação com *Média*

Os resultados foi muito regular para esta estratégia que se apresentou no 1º lugar do ranking apenas na base *Wisconsin Breast Cancer* enquanto nas bases *Iris Plants* e *Pima Indians Diabetes* a colocação ficou mais próxima dos últimos lugares.

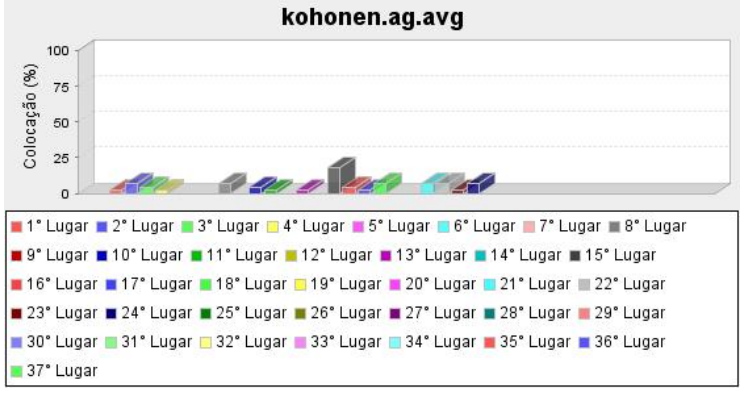
Iris Plants



Pima Indians Diabetes



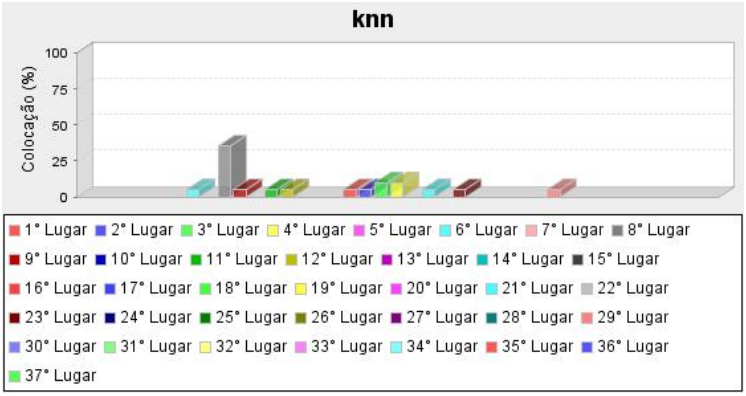
Wisconsin Breast Cancer



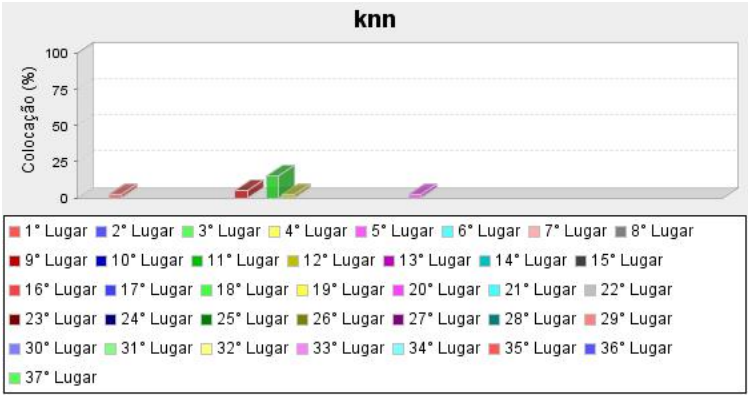
Estratégia 12: Imputação com *k-NN*

Os resultados foram muito regulares para esta estratégia que se apresentou próxima aos primeiros lugares do ranking em todas as bases de dados. Apenas a base *Iris Plants* conseguiu o 1º lugar, mas com um percentual de colocação muito baixo.

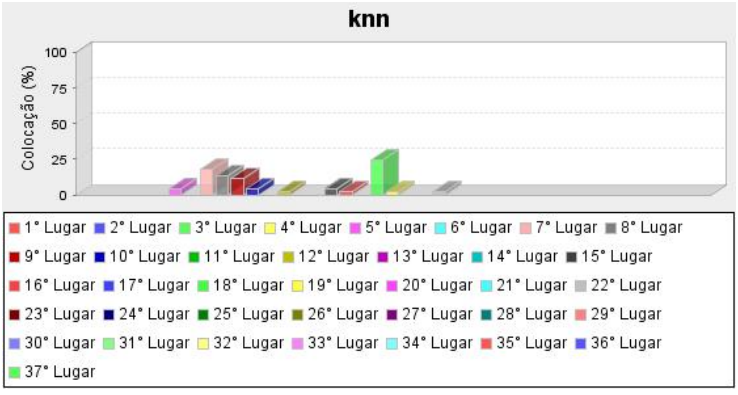
Iris Plants



Pima Indians Diabetes



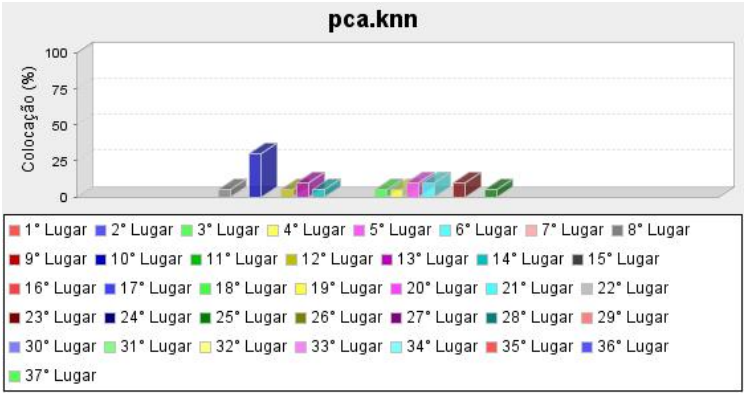
Wisconsin Breast Cancer



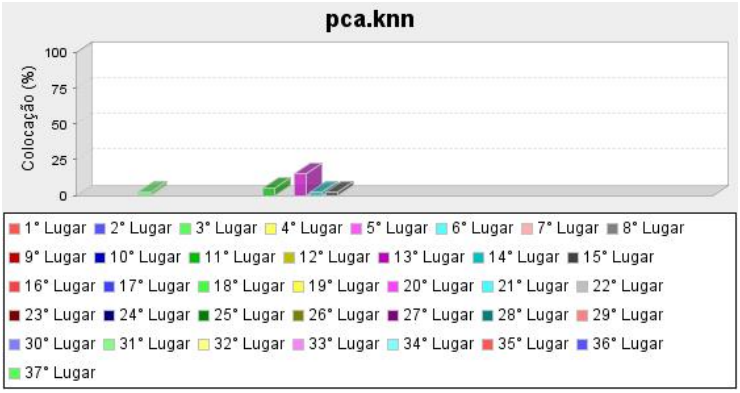
Estratégia 13: Seleção com *PCA* e Imputação com *k-NN*

Os resultados não foram satisfatórios para esta estratégia pois nenhuma base alcançou o 1º lugar ranking. A base *Iris Plants* conseguiu no máximo o 8º lugar, *Pima Indians Diabetes* o 3º lugar e *Wisconsin Breast Cancer* o 11º lugar.

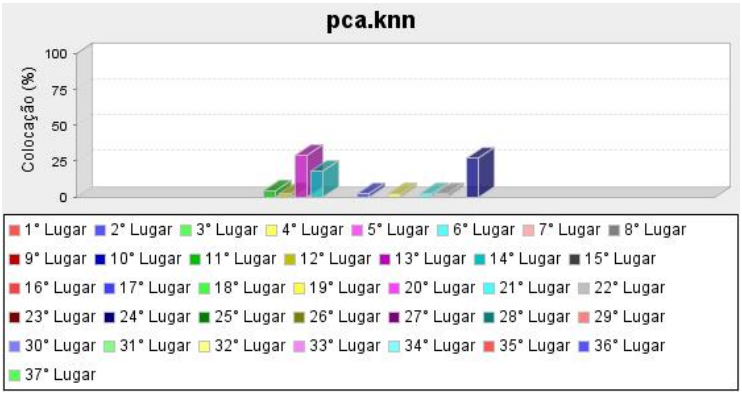
Iris Plants



Pima Indians Diabetes



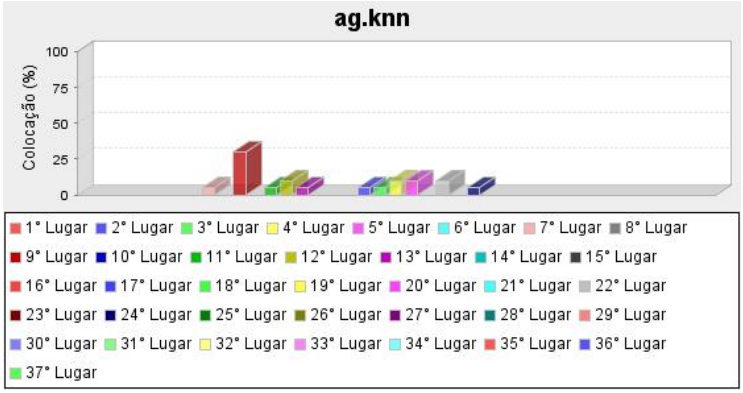
Wisconsin Breast Cancer



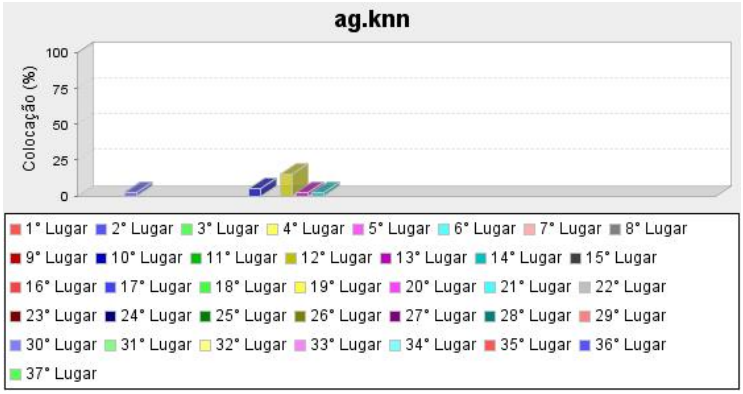
Estratégia 14: Seleção com *AG* e Imputação com *k-NN*

Os resultados não foram satisfatórios para esta estratégia, pois apesar de não ficarem próximos aos últimos lugares, nenhuma base alcançou o 1º lugar ranking. A base *Iris Plants* conseguiu no máximo o 5º lugar, *Pima Indians Diabetes* o 2º lugar e *Wisconsin Breast Cancer* o 3º lugar.

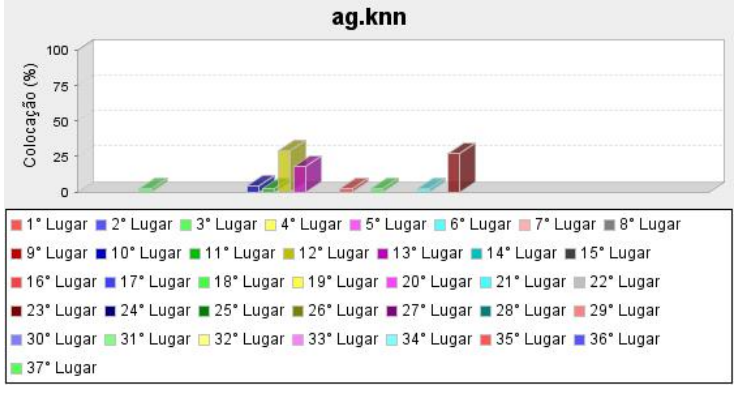
Iris Plants



Pima Indians Diabetes



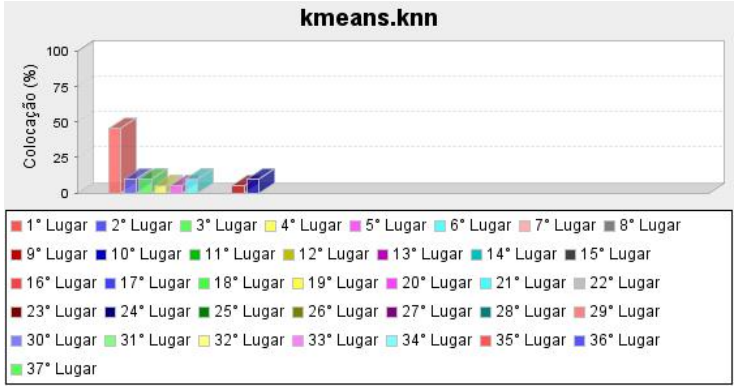
Wisconsin Breast Cancer



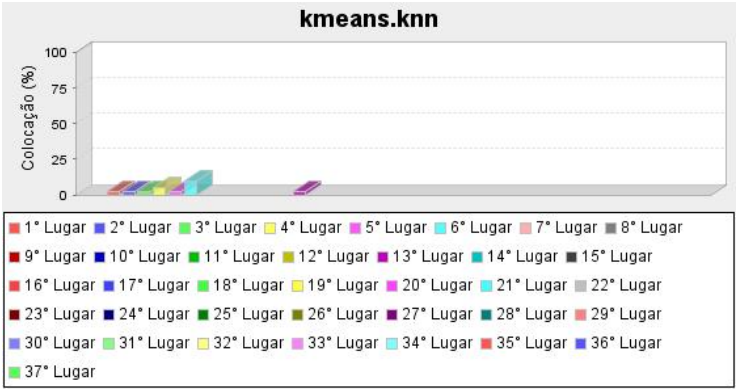
Estratégia 16: Agrupamento com *K-Means* e Imputação com *k-NN*

Os resultados foram bons para esta estratégia que conseguiu alcançar o 1º lugar do ranking em todas as bases de dados. A base *Iris Plants* conseguiu um excelente resultado com 50% de colocação, a base *Pima Indians Diabetes* ficou com 5% de colocação enquanto a base *Wisconsin Breast Cancer* ficou com 25%.

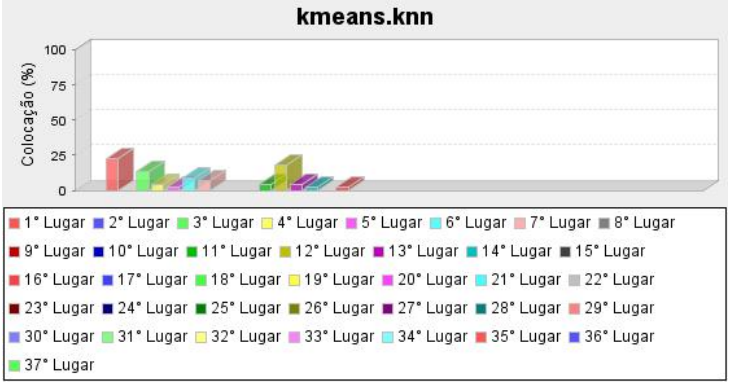
Iris Plants



Pima Indians Diabetes



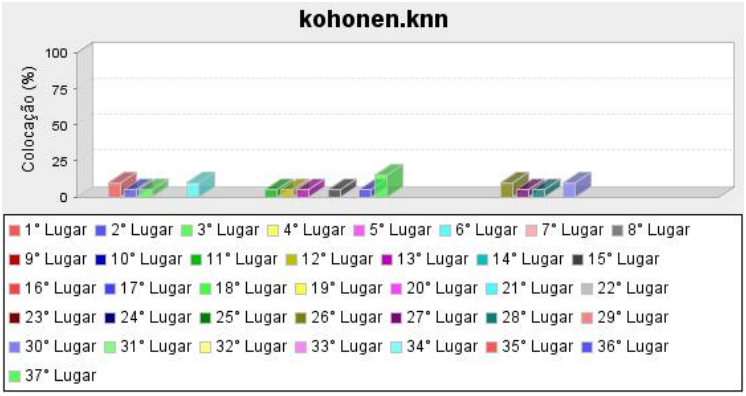
Wisconsin Breast Cancer



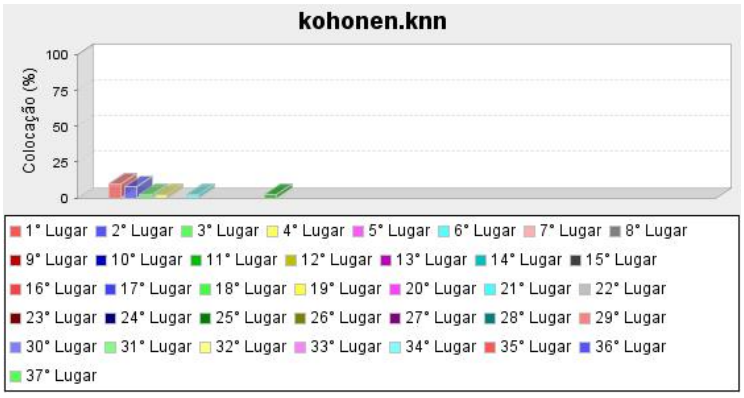
Estratégia 15: Agrupamento com *Redes de Kohonen* e Imputação com *k-NN*

Os resultados foram bons para esta estratégia que conseguiu alcançar os primeiros lugares do ranking em todas as bases de dados, sendo que as bases *Iris Plants* e *Pima Indians Diabetes* ficaram com 10% de colocação enquanto a base *Wisconsin Breast Cancer* ficou com 25%.

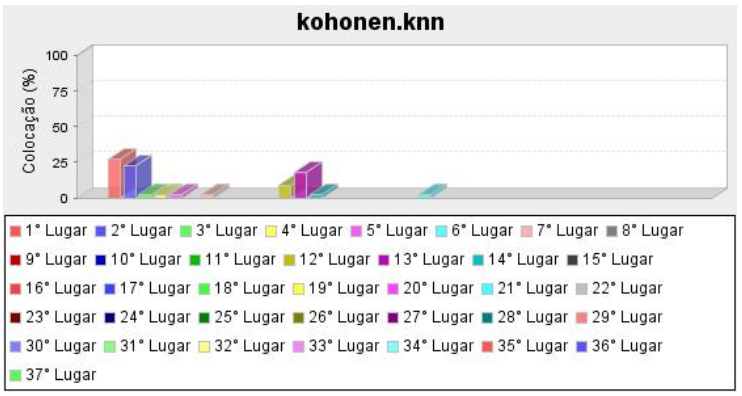
Iris Plants



Pima Indians Diabetes



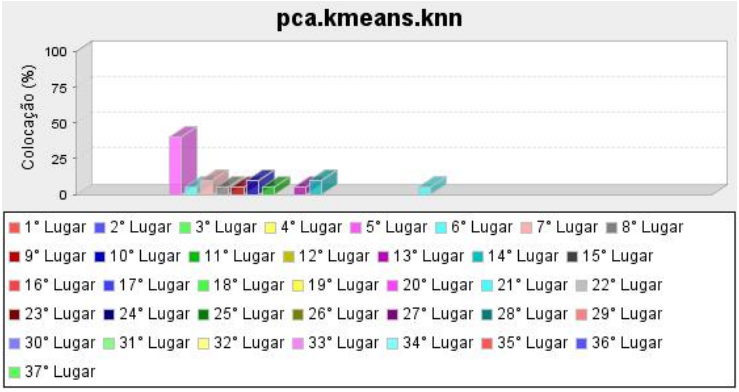
Wisconsin Breast Cancer



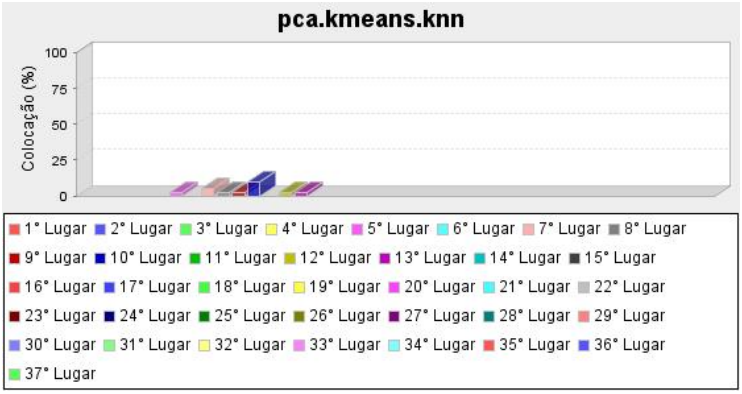
Estratégia 17: Seleção com *PCA*, Agrupamento com *K-Means* e Imputação com *k-NN*

Os resultados não foram satisfatórios para esta estratégia pois nenhuma base alcançou o 1º lugar ranking. As bases *Iris Plants, Pima Indians Diabetes* e *Wisconsin Breast Cancer* conseguiram no máximo o 5º lugar.

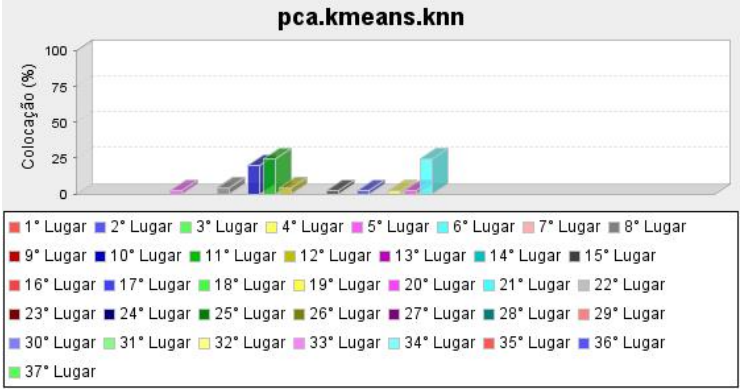
Iris Plants



Pima Indians Diabetes



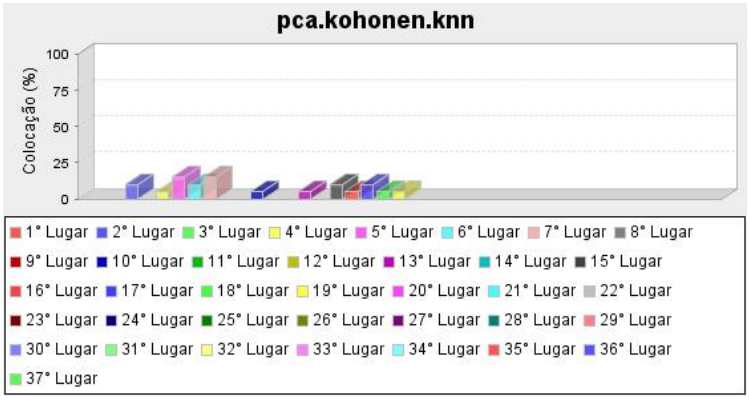
Wisconsin Breast Cancer



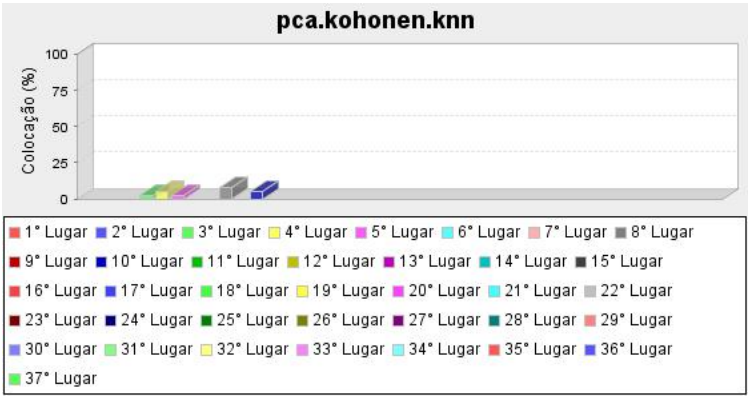
Estratégia 18: Seleção com *PCA*, Agrupamento com *Redes de Kohonen* e Imputação com *k-NN*

Os resultados foram regulares para esta estratégia que se apresentou próxima aos primeiros lugares do ranking em todas as bases de dados, sendo que o melhor resultado não foi o primeiro lugar e sim o 2º lugar na base *Iris Plants*.

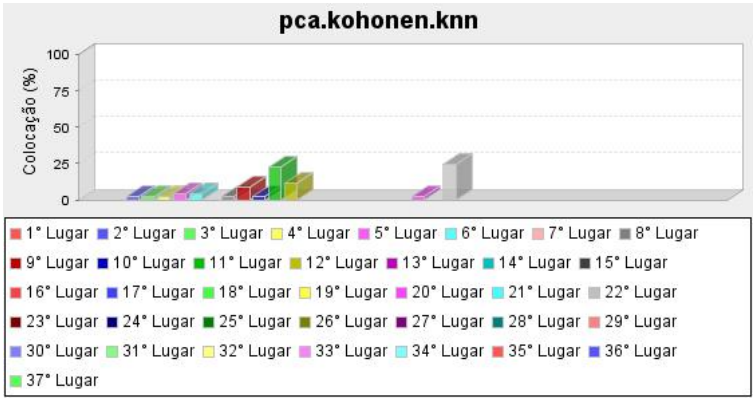
Iris Plants



Pima Indians Diabetes



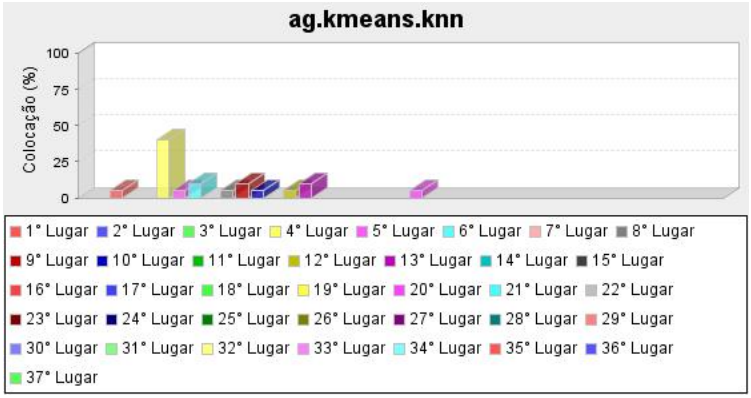
Wisconsin Breast Cancer



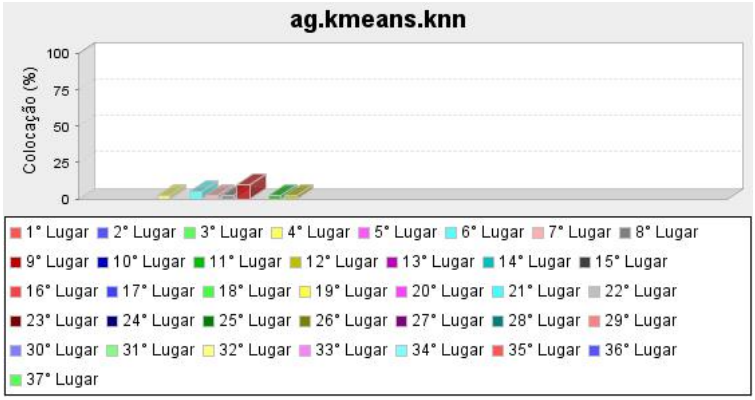
Estratégia 19: Seleção com *AG*, Agrupamento com K-Means e Imputação com *k-NN*

Os resultados foram regulares para esta estratégia que se apresentou próxima aos primeiros lugares do ranking em todas as bases de dados, pois somente na base *Iris Plants* obteve o 1º lugar e com um baixo percentual de colocação.

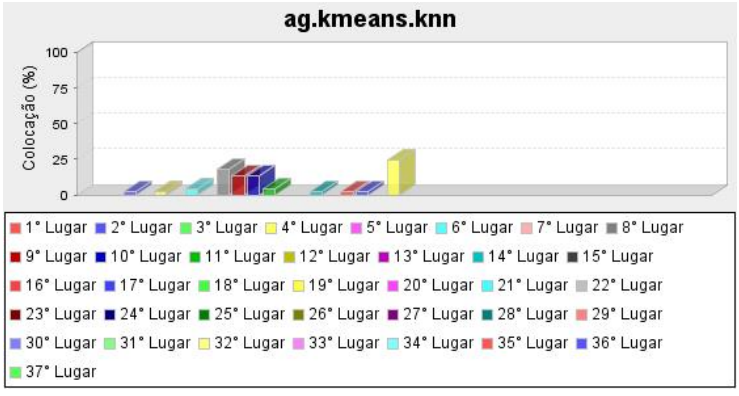
Iris Plants



Pima Indians Diabetes



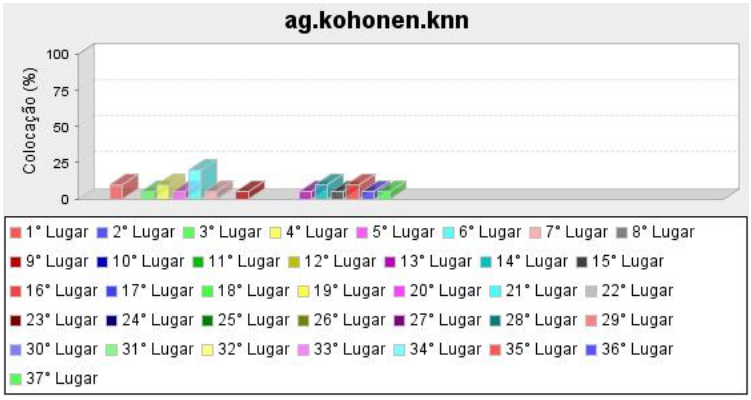
Wisconsin Breast Cancer



Estratégia 20: Seleção com *AG*, Agrupamento com *Redes de Kohonen* e Imputação com *k-NN*

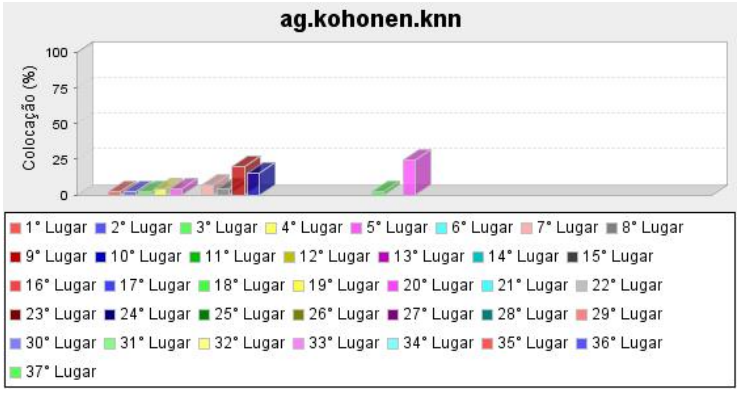
Os resultados foram regulares para esta estratégia que se apresentou próxima aos primeiros lugares do ranking em todas as bases de dados, sendo que a base *Iris Plants* conseguiu 10% de colocação no 1º lugar.

Iris Plants



Pima Indians Diabetes

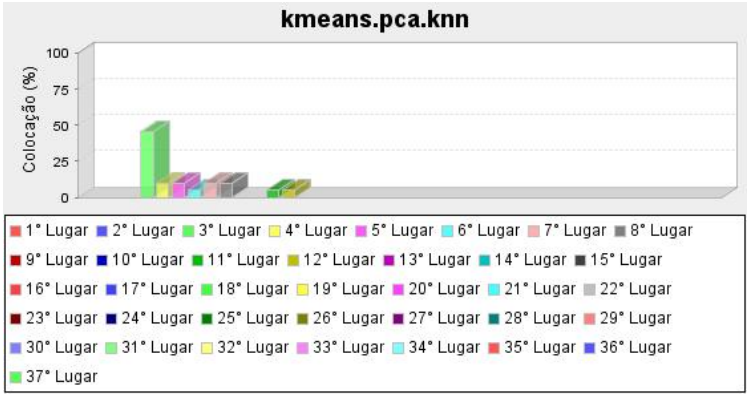
Wisconsin Breast Cancer



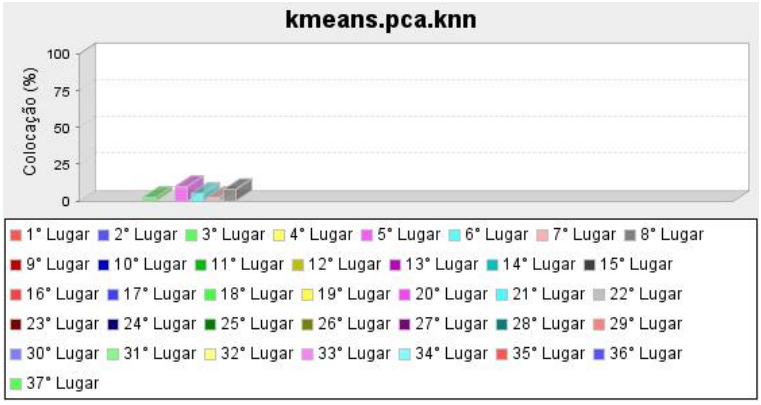
Estratégia 21: Agrupamento com K-Means, Seleção com *PCA* e Imputação com k-NN

Os resultados não foram satisfatórios para esta estratégia pois nenhuma base alcançou o 1º lugar ranking. As bases *Iris Plants, Pima Indians Diabetes* e *Wisconsin Breast Cancer* conseguiram no máximo o 3º lugar.

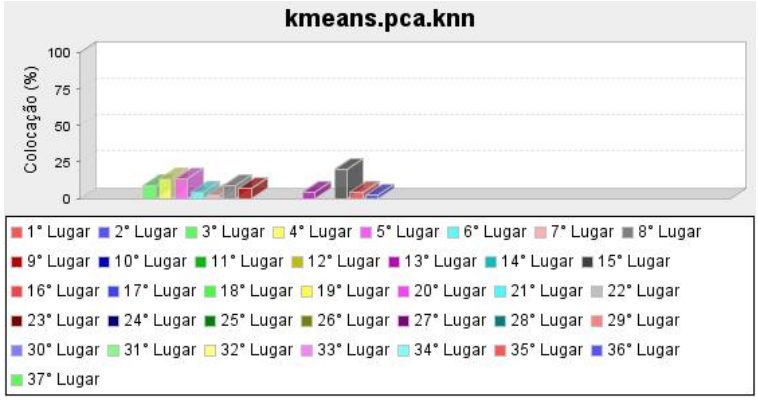
Iris Plants



Pima Indians Diabetes



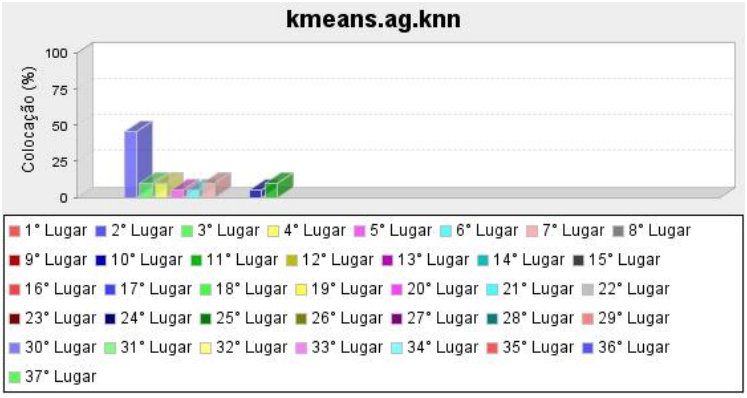
Wisconsin Breast Cancer



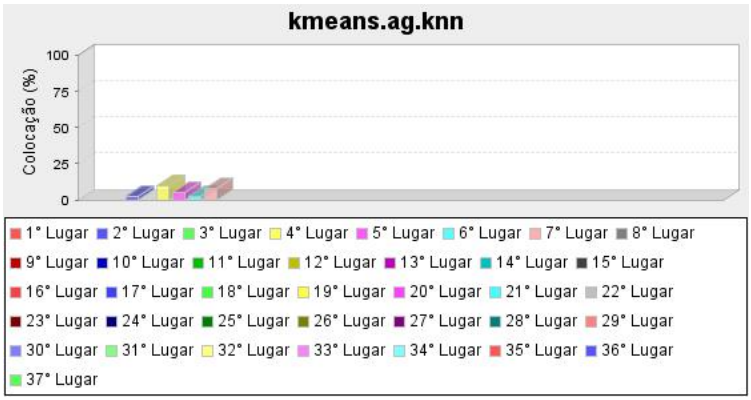
Estratégia 22: Agrupamento com *K-Means*, Seleção com *AG* e Imputação com *k-NN*

Os resultados foram regulares para esta estratégia que se apresentou próxima aos primeiros lugares do ranking em todas as bases de dados. Na *Iris Plants* o 2º lugar se destacou com alto percentual, 50%, e na base *Wisconsin Breast Cancer* o 1º lugar obteve 5% na colocação.

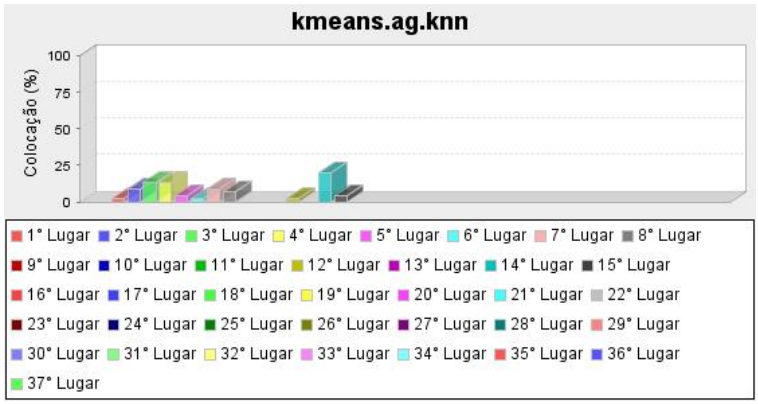
Iris Plants



Pima Indians Diabetes



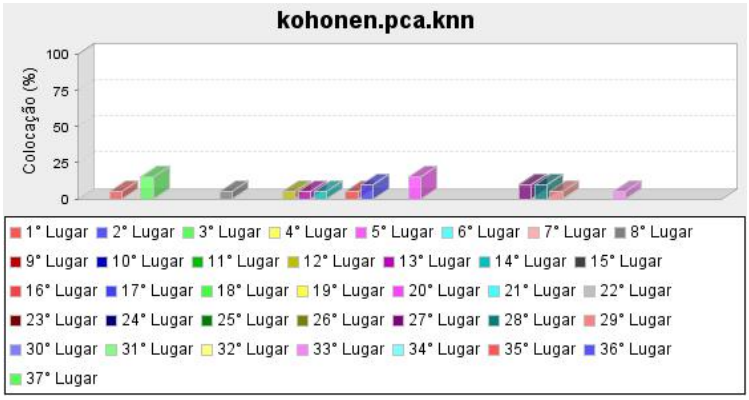
Wisconsin Breast Cancer



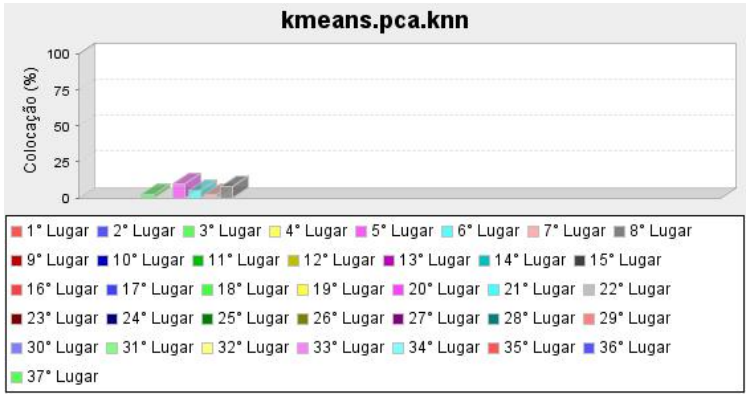
Estratégia 23: Agrupamento com *Redes de Kohonen*, Seleção com *PCA* e Imputação com *k*-NN

Os resultados foram regulares para esta estratégia que se apresentou próxima aos primeiros lugares do ranking em todas as bases de dados, sendo que a base *Iris Plants* conseguiu 5% de colocação no 1º lugar.

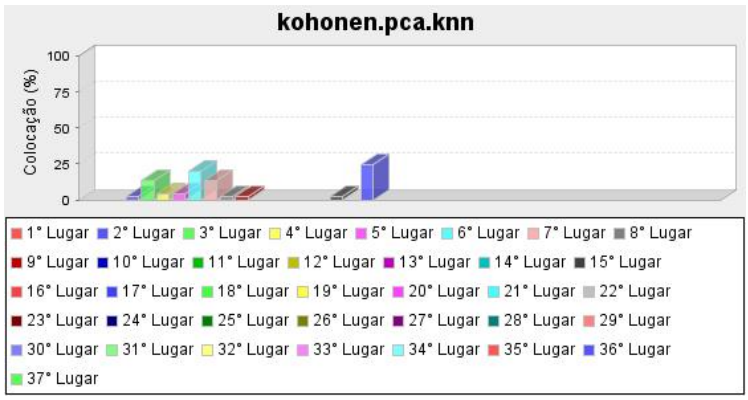
Iris Plants



Pima Indians Diabetes



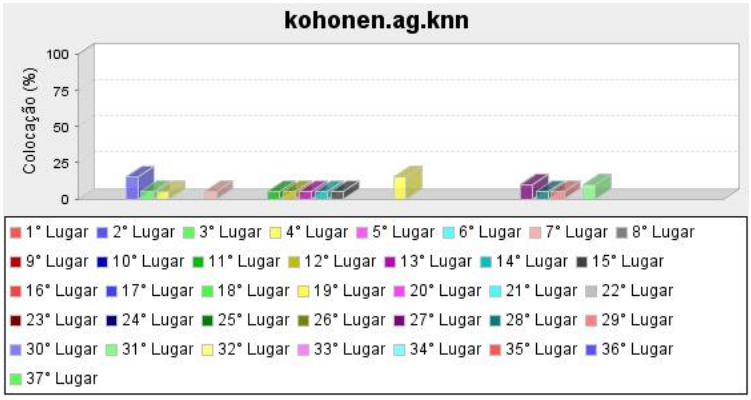
Wisconsin Breast Cancer



Estratégia 24: Agrupamento com *Redes de Kohonen*, Seleção com *AG* e Imputação com *k-NN*

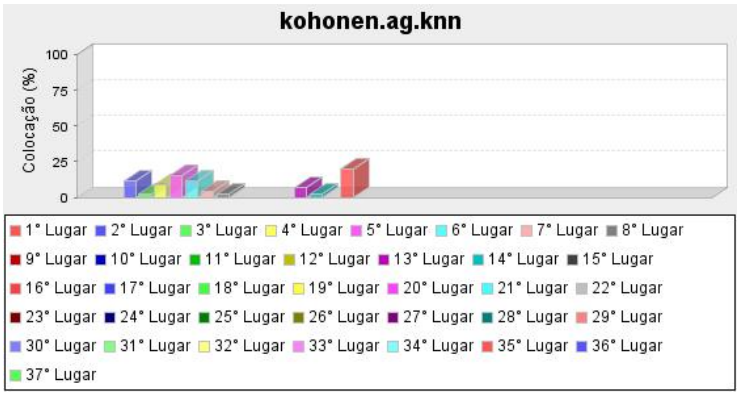
Os resultados foram regulares para esta estratégia que se apresentou próxima aos primeiros lugares do ranking em todas as bases de dados. *Iris Plants* e *Wisconsin Breast Cancer* ficaram com 2º lugar enquanto a base *Pima Indians Diabetes* obteve 10% de colocação no 1º lugar.

Iris Plants



Pima Indians Diabetes

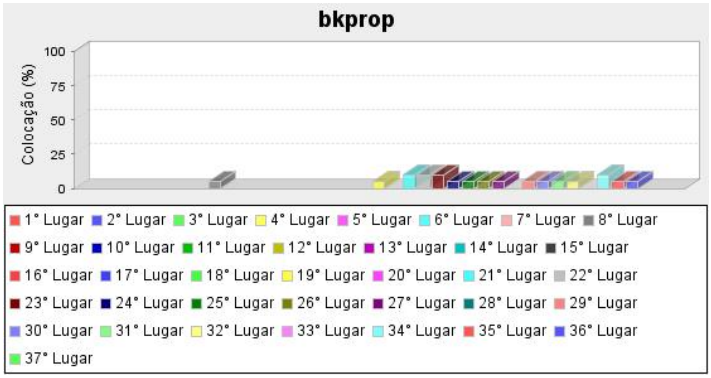
Wisconsin Breast Cancer



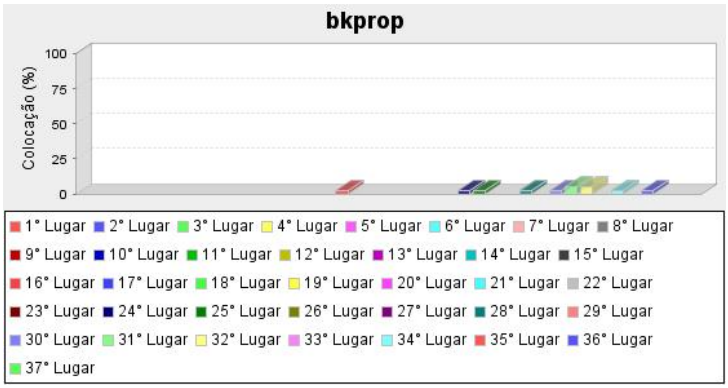
Estratégia 25: Imputação com *Back Propagation*

Os resultados não foram satisfatórios para esta estratégia que se apresentou nos últimos lugares do ranking em todas as bases de dados.

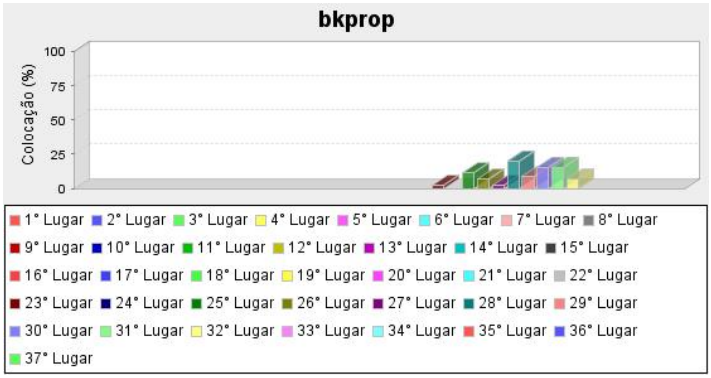
Iris Plants



Pima Indians Diabetes



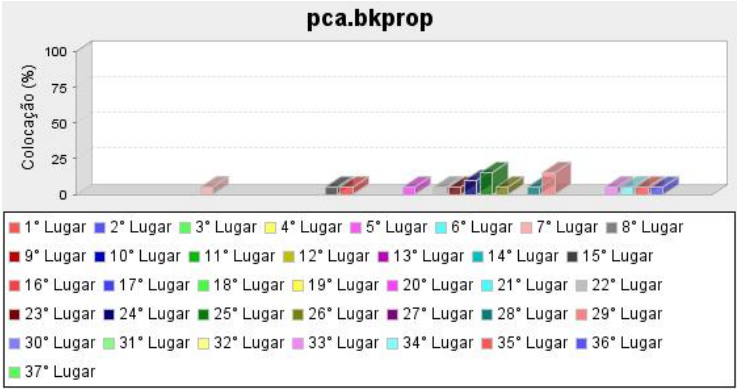
Wisconsin Breast Cancer



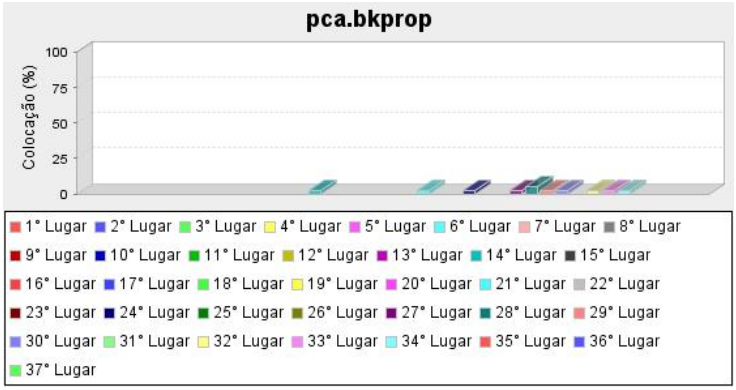
Estratégia 26: Seleção com *PCA* e Imputação com *Back Propagation*

Os resultados não foram satisfatórios para esta estratégia que se apresentou nos últimos lugares do ranking em todas as bases de dados.

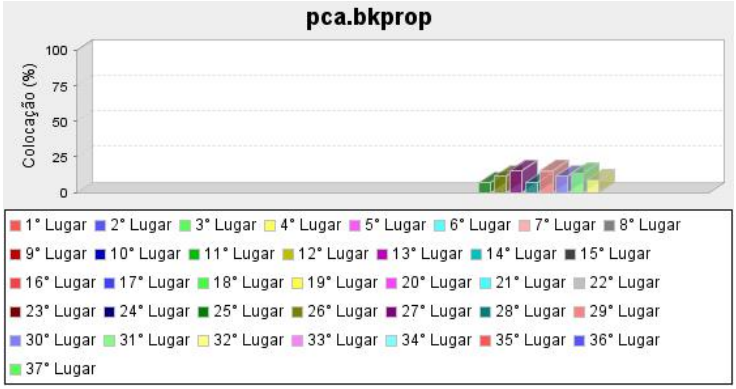
Iris Plants



Pima Indians Diabetes



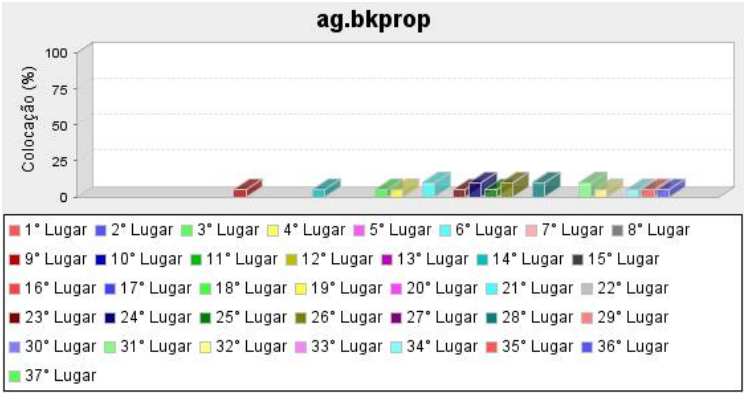
Wisconsin Breast Cancer



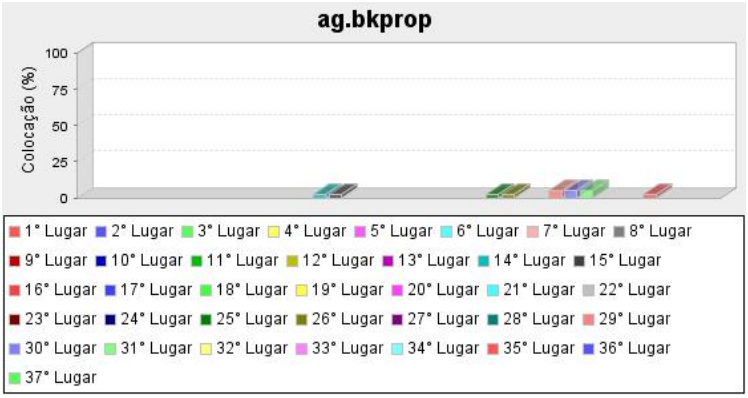
Estratégia 27: Seleção com *AG* e Imputação com *Back Propagation*

Os resultados não foram satisfatórios para esta estratégia que se apresentou nos últimos lugares do ranking nas bases *Pima Indians Diabetes* e *Wisconsin Breast Cancer*, sendo que o melhor resultado está na base *Iris Plants* com o 9ª lugar.

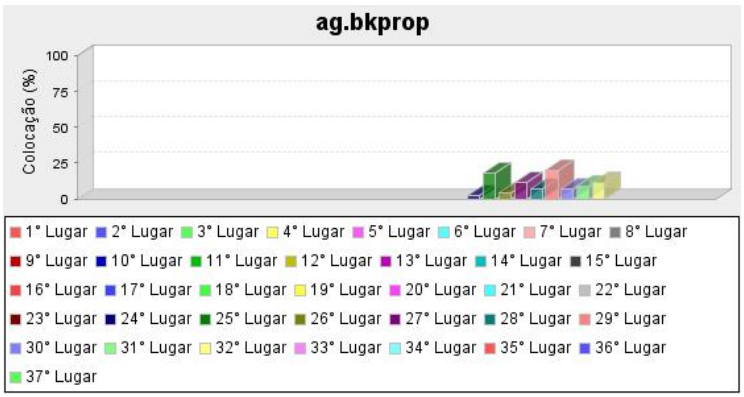
Iris Plants



Pima Indians Diabetes



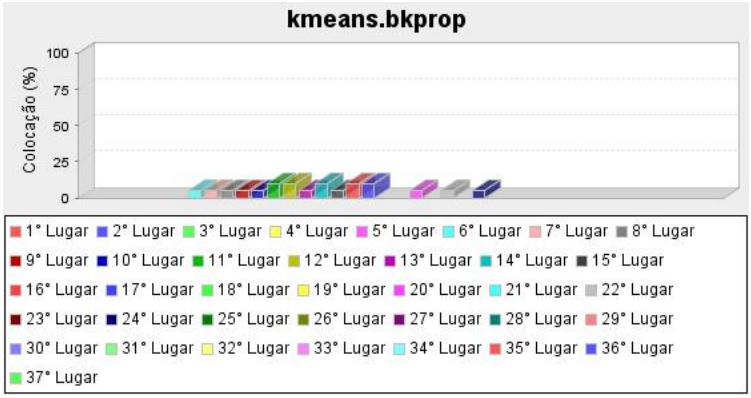
Wisconsin Breast Cancer



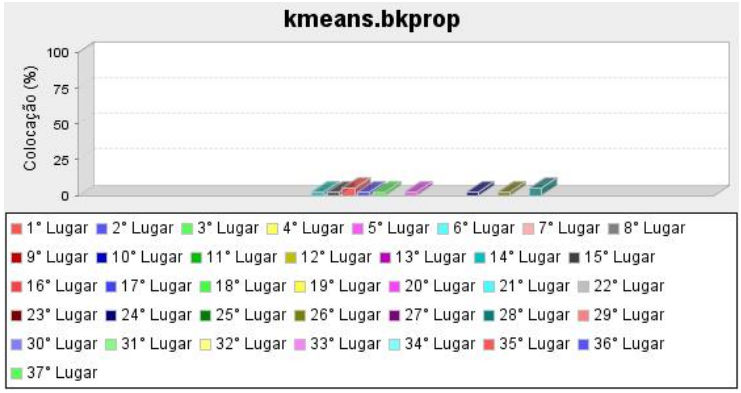
Estratégia 28: Agrupamento com K-Means e Imputação com *Back Propagation*

Os resultados não foram satisfatórios para esta estratégia que se apresentou distante dos primeiros lugares do ranking em todas as bases de dados, o melhor resultado foi o 6º lugar na base *Iris Plants*.

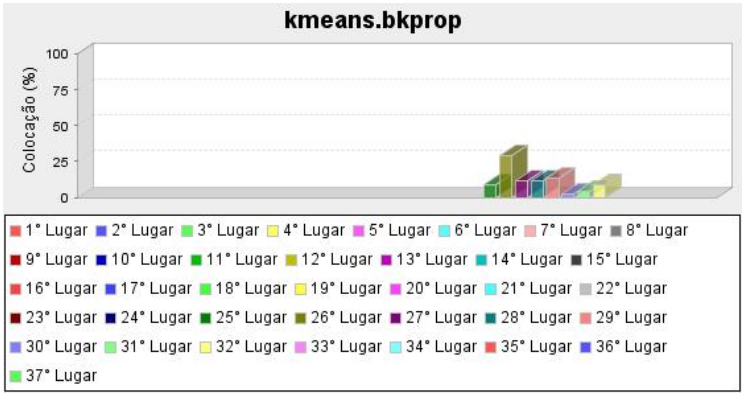
Iris Plants



Pima Indians Diabetes



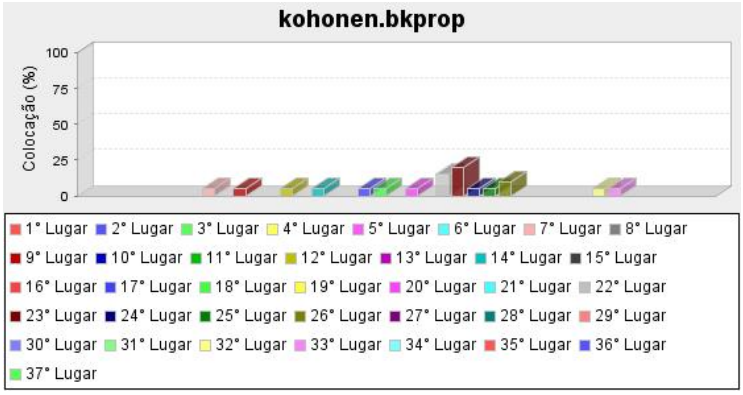
Wisconsin Breast Cancer



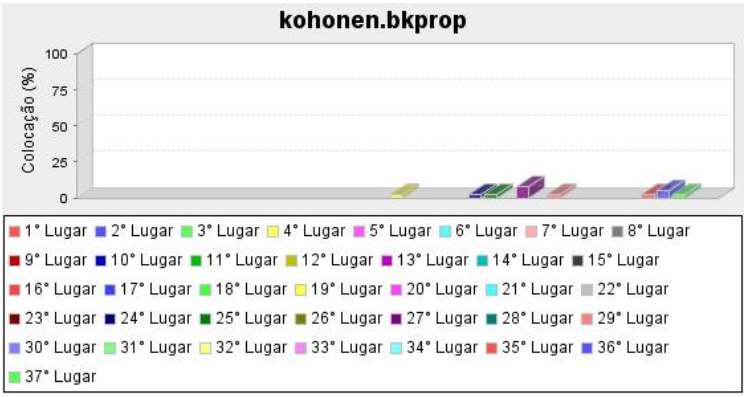
Estratégia 29: Agrupamento com *Redes de Kohonen* e Imputação com Back Propagation

Os resultados não foram satisfatórios para esta estratégia que se apresentou nos últimos lugares do ranking nas bases *Pima Indians Diabetes* e *Wisconsin Breast Cancer*, sendo que o melhor resultado está na base *Iris Plants* com o 7ª lugar.

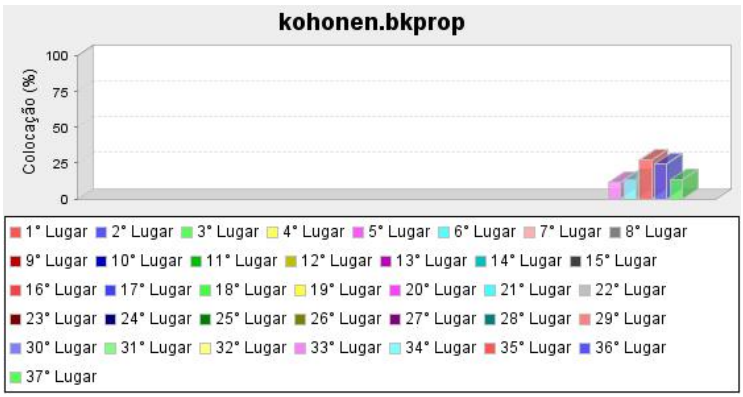
Iris Plants



Pima Indians Diabetes



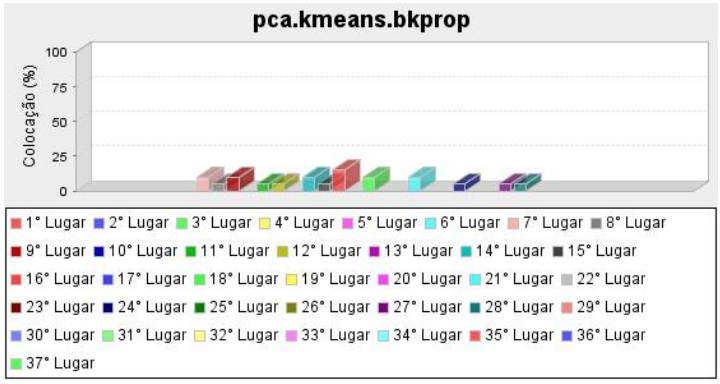
Wisconsin Breast Cancer



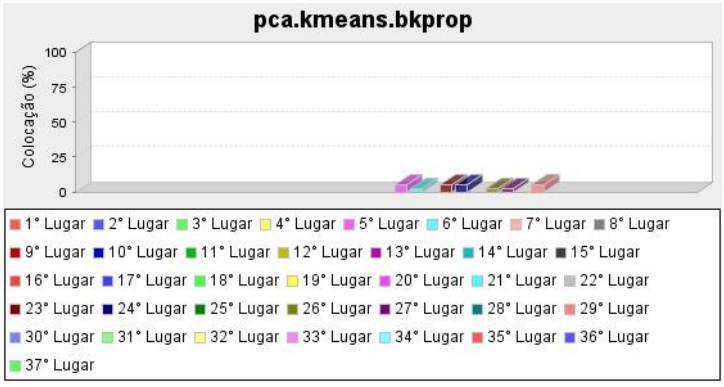
Estratégia 30: Seleção com *PCA*, Agrupamento com *K-Means* e Imputação com *Back Propagation*

Os resultados não foram satisfatórios para esta estratégia que se apresentou nos últimos lugares do ranking em todas as bases de dados, o melhor resultado foi o 7º lugar na base *Iris Plants*.

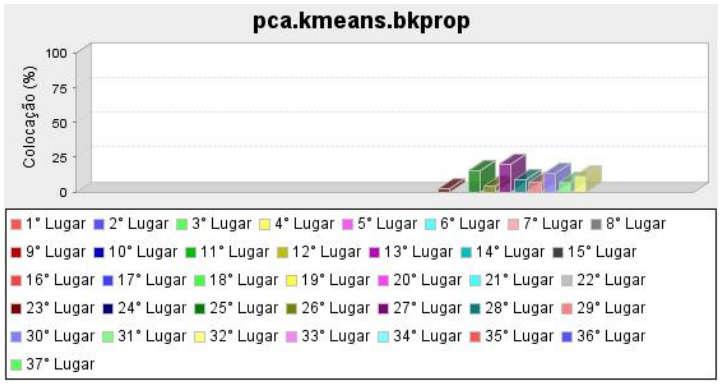
Iris Plants



Pima Indians Diabetes



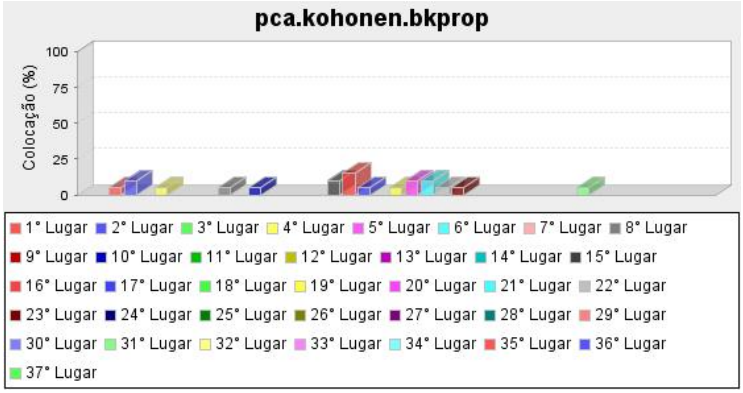
Wisconsin Breast Cancer



Estratégia 31: Seleção com *PCA*, Agrupamento com *Redes de Kohonen* e Imputação com Back Propagation

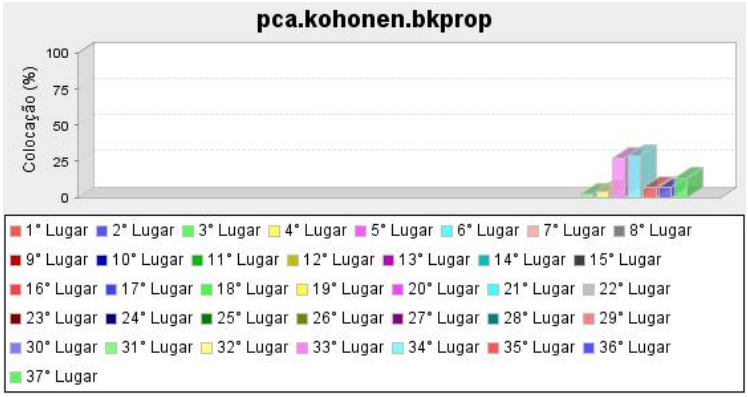
Os resultados não satisfatórios para esta estratégia nas bases *Pima Indians Diabetes* e *Wisconsin Breast Cancer* por ter ocupado os últimos lugares do ranking, entretanto na base *Iris Plants* o 1º lugar até foi alcançado mesmo que com um baixo percentual de colocação.

Iris Plants



Pima Indians Diabetes

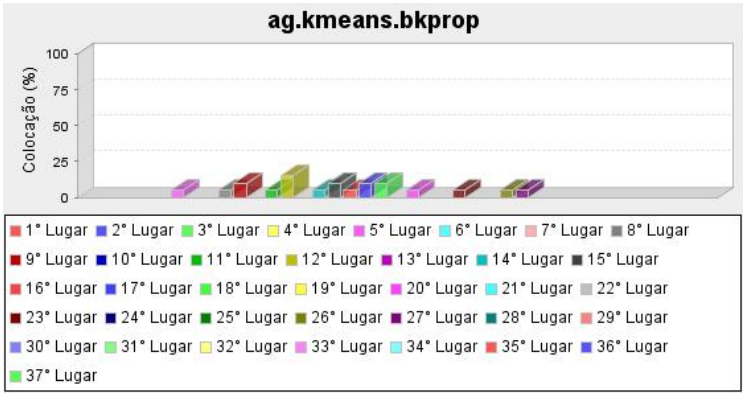
Wisconsin Breast Cancer



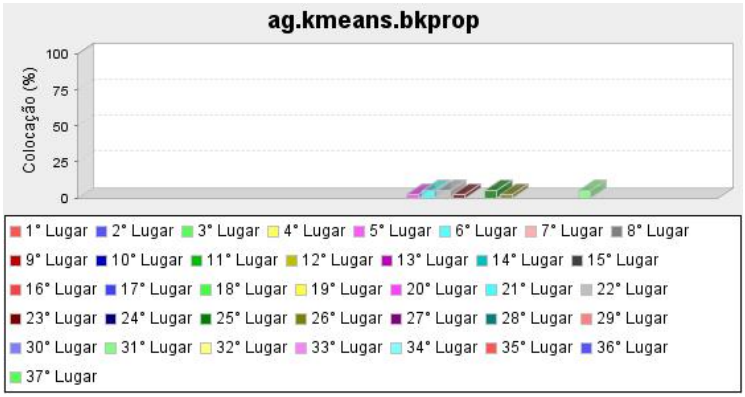
Estratégia 32: Seleção com *AG*, Agrupamento com *K-Means* e Imputação com *Back Propagation*

Os resultados foram não satisfatórios para esta estratégia que obteve os últimos lugares do ranking nas bases *Pima Indians Diabetes* e *Wisconsin Breast Cancer* e alcançou no máximo o 5º lugar na base *Iris Plants*.

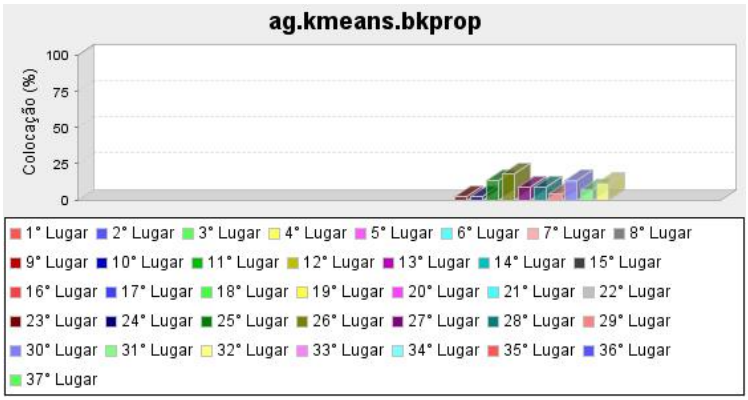
Iris Plants



Pima Indians Diabetes



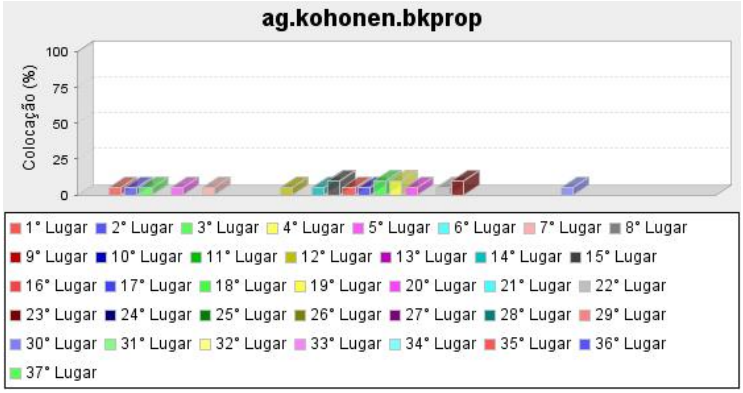
Wisconsin Breast Cancer



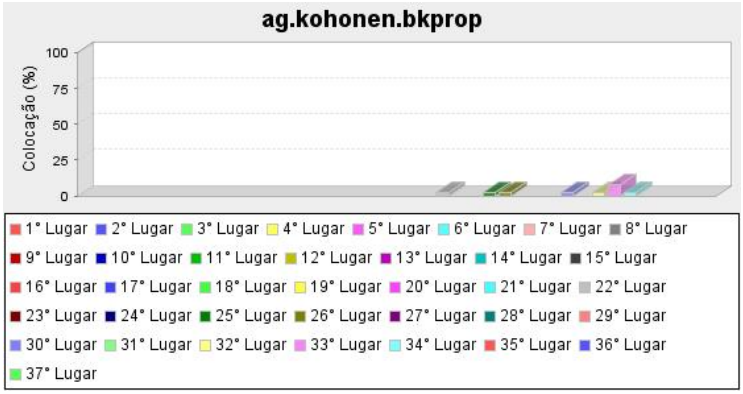
Estratégia 33: Seleção com *AG*, Agrupamento com *Redes de Kohonen* e Imputação com *Back Propagation*

Os resultados não satisfatórios para esta estratégia nas bases *Pima Indians Diabetes* e *Wisconsin Breast Cancer* por ter ocupado os últimos lugares do ranking, entretanto na base *Iris Plants* o 1º lugar até foi alcançado mas com um baixo percentual de colocação.

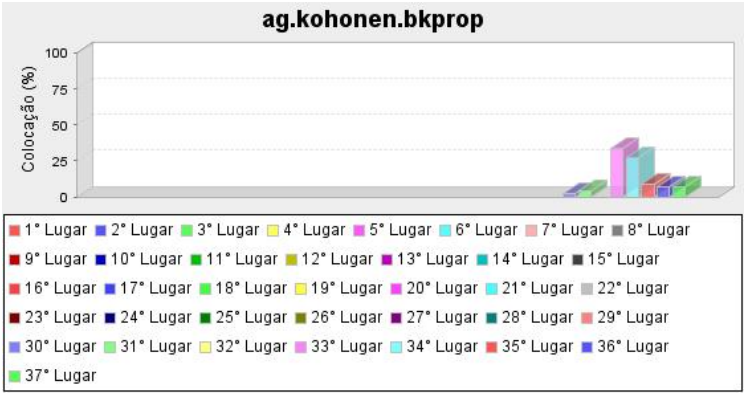
Iris Plants



Pima Indians Diabetes



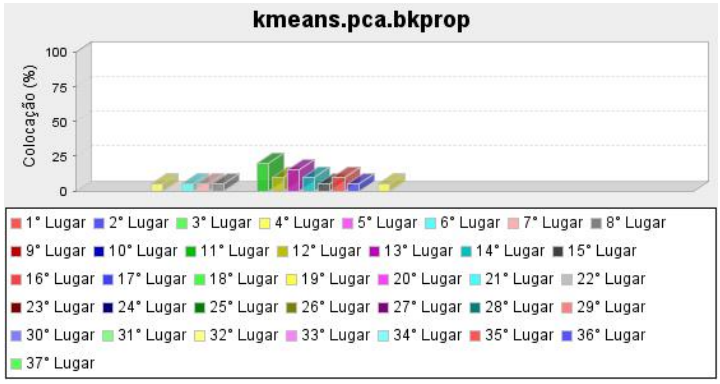
Wisconsin Breast Cancer



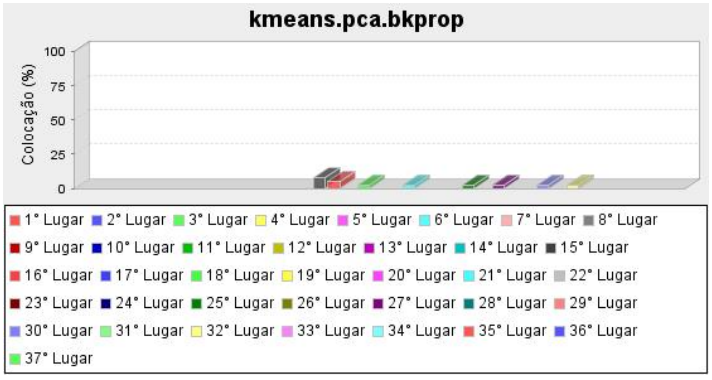
Estratégia 34: Agrupamento com *K-Means*, Seleção com *PCA* e Imputação com *Back Propagation*

Os resultados não foram satisfatórios para esta estratégia que se apresentou nos últimos lugares do ranking em todas as bases de dados, o melhor resultado foi o 4º lugar na base *Iris Plants*.

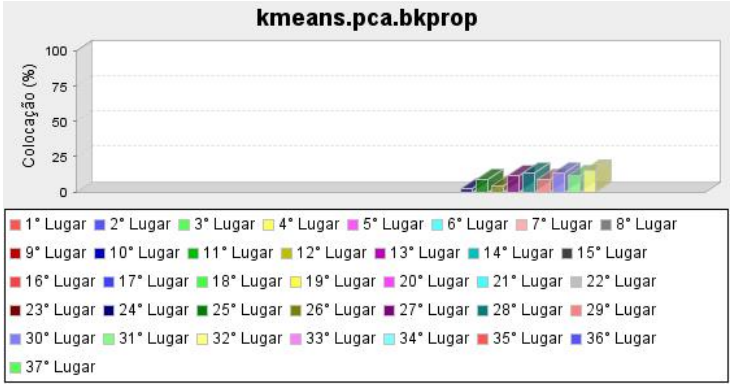
Iris Plants



Pima Indians Diabetes



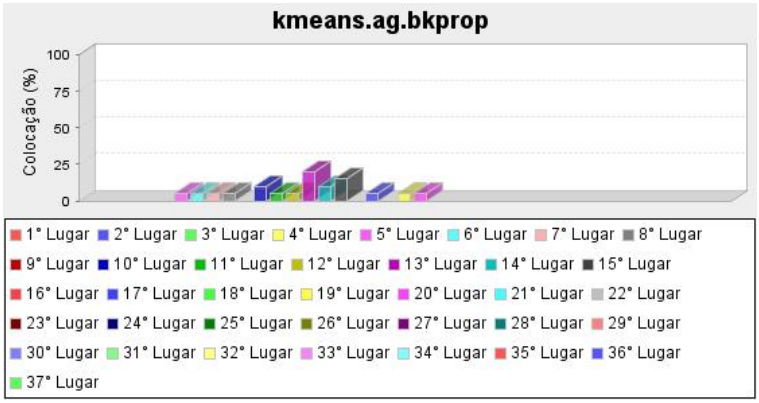
Wisconsin Breast Cancer



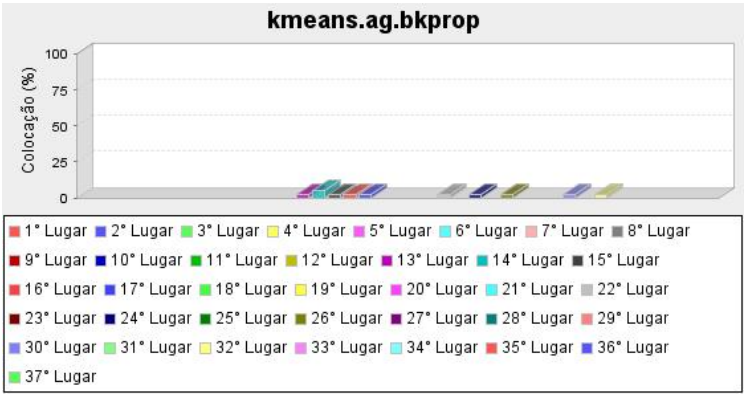
Estratégia 35: Agrupamento com *K-Means*, Seleção com *AG* e Imputação com *Back Propagation*

Os resultados foram não satisfatórios para esta estratégia que se mostrou nos últimos lugares do ranking nas bases *Pima Indians Diabetes* e *Wisconsin Breast Cancer* e no máximo alcançou o 5º lugar na base *Iris Plants*.

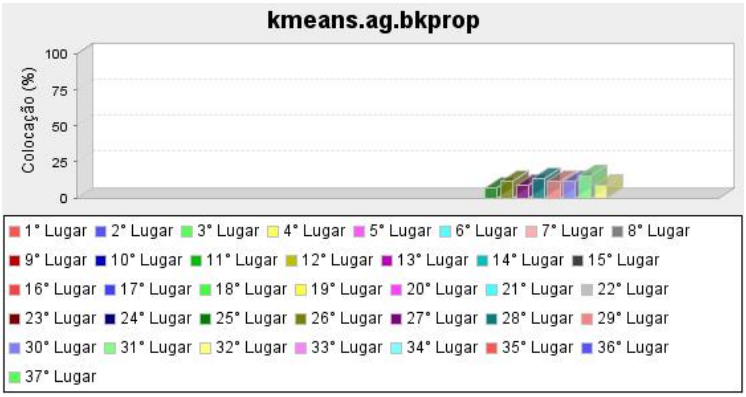
Iris Plants



Pima Indians Diabetes



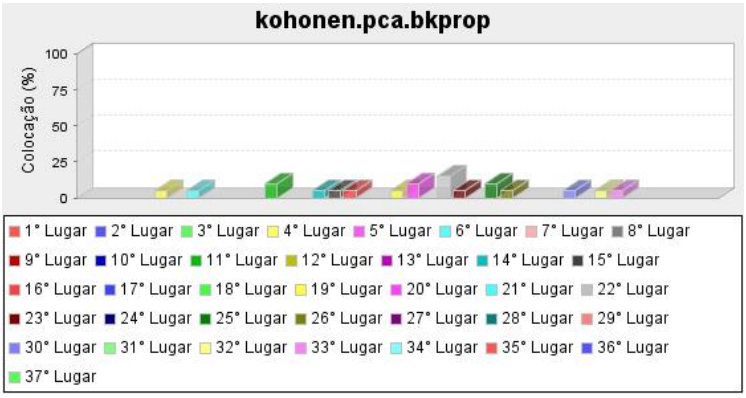
Wisconsin Breast Cancer



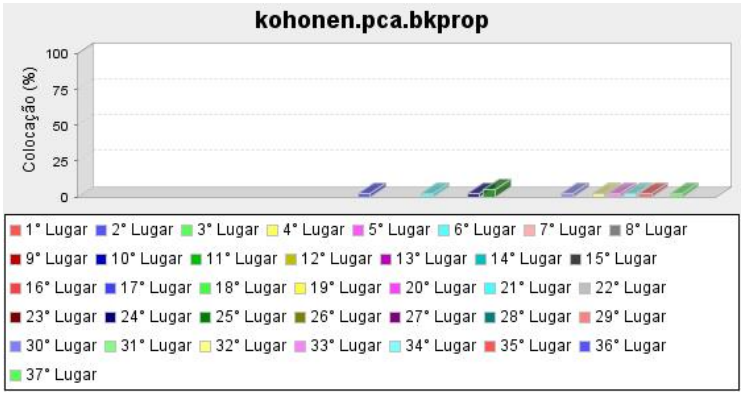
Estratégia 36: Agrupamento com *Redes de Kohonen*, Seleção com *PCA* e Imputação com *Back Propagation*

Os resultados não satisfatórios para esta estratégia que se apresentou nos últimos lugares do ranking nas bases *Pima Indians Diabetes* e *Wisconsin Breast Cancer*, sendo que o melhor resultado está na base *Iris Plants* com o 4ª lugar.

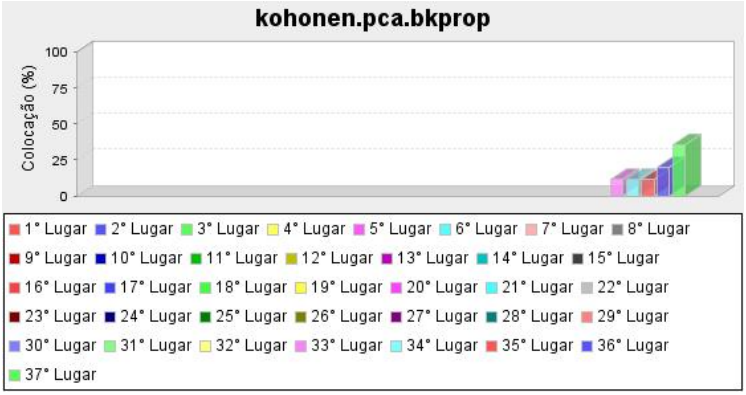
Iris Plants



Pima Indians Diabetes



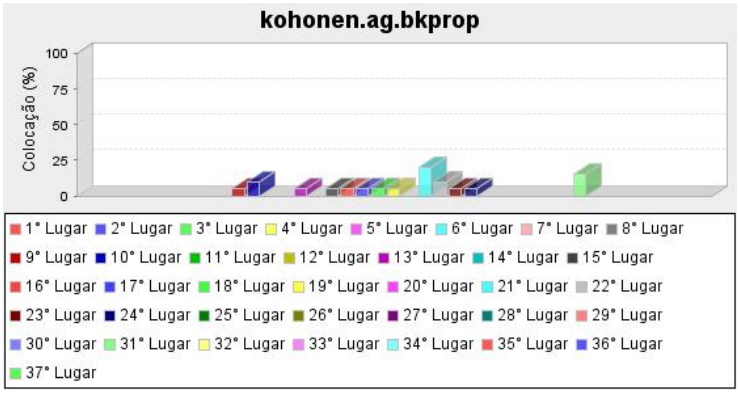
Wisconsin Breast Cancer



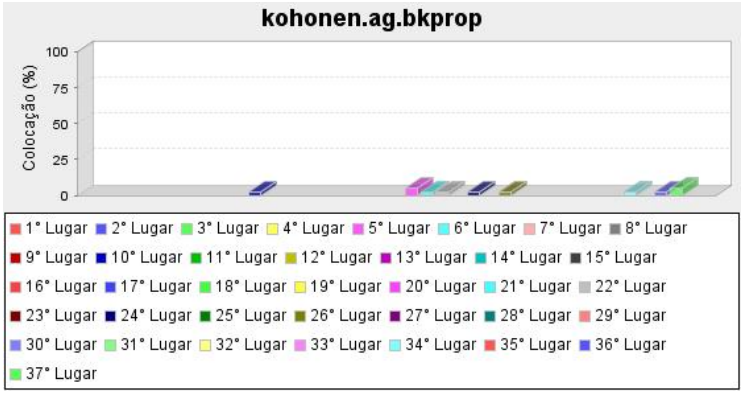
Estratégia 37: Agrupamento com *Redes de Kohonen*, Seleção com *AG* e Imputação com *Back Propagation*

Os resultados foram não satisfatórios para esta estratégia que se mostrou nos últimos lugares do ranking nas bases *Pima Indians Diabetes* e *Wisconsin Breast Cancer* e no máximo alcançando o 9º lugar na base *Iris Plants*.

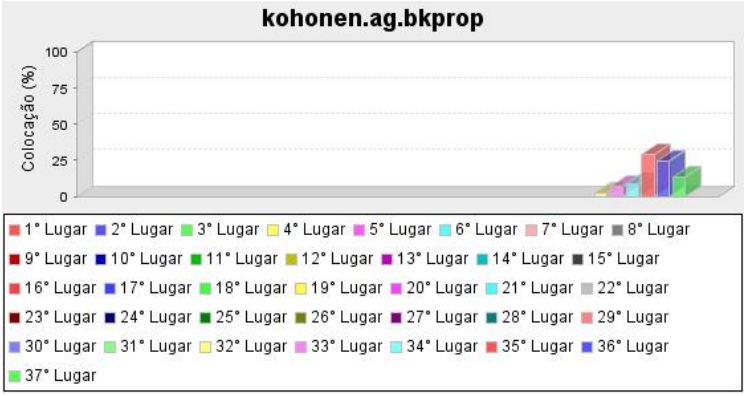
Iris Plants



Pima Indians Diabetes



Wisconsin Breast Cancer



5.3 Comparação dos resultados por base de dados

Todos os resultados de SOARES(2007) foram gerados novamente juntamente com os novos resultados das novas técnicas implementadas neste trabalho. A imputação composta abordada por SOARES (2007) possui 14 combinações de técnicas de regressão, agrupamento e seleção que iremos chamar de *Estratégia de Soares*:

1. Imputação com *Média*
2. Agrupamento com *K-Means* seguido de imputação com *Média*
3. Seleção com *PCA* seguido de agrupamento com *K-Means* seguida de imputação com *Média*
4. Agrupamento com *K-Means* seguido de seleção com *PCA* seguida de imputação com *Média*
5. Imputação com *k-NN*
6. Seleção com *PCA* seguida de imputação com *k-NN*
7. Agrupamento com *K-Means* seguido de imputação com *k-NN*
8. Seleção com *PCA* seguido de agrupamento com *K-Means* seguida de imputação com *k-NN*
9. Agrupamento com *K-Means* seguido de seleção com *PCA* seguida de imputação com *k-NN*
10. Imputação com *Back Propagation*
11. Seleção com *PCA* seguida de imputação com *Back Propagation*
12. Agrupamento com *K-Means* seguido de imputação com *Back Propagation*
13. Seleção com *PCA* seguido de agrupamento com *K-Means* seguida de imputação com *Back Propagation*
14. Agrupamento com *K-Means* seguido de seleção com *PCA* seguida de imputação com *Back Propagation*

Os resultados de todas as combinações acima certamente não serão idênticos ao de SOARES (2007) mas estarão bem próximos porque foram utilizadas as mesmas bases de dados (*Iris Plants, Pima Indians Diabetes, Wisconsin Breast Cancer*), percentuais de ausência (10%, 20%, 30%, 40%, 50%), mecanismo de ausência (*MACR*), sistema (*Appraisal*) e combinações de técnicas de imputação entretanto não foi retirado nenhum registro das bases de dados como foi feito com a base *Wisconsin Breast Cancer*, não foi considerado o comitê e

algumas configurações das técnicas de imputação foram descartadas seguindo as observações de SOARES (2007).

A finalidade não é comparar essas pequenas diferenças e sim os resultados dessas combinações de SOARES (2007) refeitos aqui neste trabalho com as novas combinações de imputação para avaliar possíveis melhorias na qualidade da imputação. Essas novas combinações listadas abaixo iremos chamar de *Estratégia Proposta*:

1. Imputação com *Média*
2. Agrupamento com *Redes de Kohonen* seguido de imputação com *Média*
3. Seleção com *AG* seguido de agrupamento com *K-Means* seguida de imputação com *Média*
4. Seleção com *PCA* seguido de agrupamento com *Redes de Kohonen* seguida de imputação com *Média*
5. Seleção com *AG* seguido de agrupamento com *Redes de Kohonen* seguida de imputação com *Média*
6. Agrupamento com *K-Means* seguido de seleção com *AG* seguida de imputação com *Média*
7. Agrupamento com *Redes de Kohonen* seguido de seleção com *PCA* seguida de imputação com *Média*
8. Agrupamento com *Redes de Kohonen* seguido de seleção com *AG* seguida de imputação com *Média*
9. Imputação com *k-NN*
10. Seleção com *AG* seguida de imputação com *k-NN*
11. Agrupamento com *Redes de Kohonen* seguido de imputação com *k-NN*
12. Seleção com *AG* seguido de agrupamento com *K-Means* seguida de imputação com *k-NN*
13. Seleção com *PCA* seguido de agrupamento com *Redes de Kohonen* seguida de imputação com *k-NN*
14. Seleção com *AG* seguido de agrupamento com *Redes de Kohonen* seguida de imputação com *k-NN*
15. Agrupamento com *K-Means* seguido de seleção com *AG* seguida de imputação com *k-NN*
16. Agrupamento com *Redes de Kohonen* seguido de seleção com *PCA* seguida de imputação com *k-NN*

17. Agrupamento com *Redes de Kohonen* seguido de seleção com *AG* seguida de imputação com *k-NN*
18. Imputação com *Back Propagation*
19. Seleção com *AG* seguida de imputação com *Back Propagation*
20. Agrupamento com *Redes de Kohonen* seguido de imputação com *Back Propagation*
21. Seleção com *AG* seguido de agrupamento com *K-Means* seguida de imputação com *Back Propagation*
22. Seleção com *PCA* seguido de agrupamento com *Redes de Kohonen* seguida de imputação com *Back Propagation*
23. Seleção com *AG* seguido de agrupamento com *Redes de Kohonen* seguida de imputação com Back Propagation
24. Agrupamento com *K-Means* seguido de seleção com *AG* seguida de imputação com *Back Propagation*
25. Agrupamento com *Redes de Kohonen* seguido de seleção com *PCA* seguida de imputação com *Back Propagation*
26. Agrupamento com *Redes de Kohonen* seguido de seleção com *AG* seguida de imputação com *Back Propagation*

5.3.1 Gráficos: Comparação das Técnicas por base de dados

Com a análise de técnicas por base de dados teremos uma visão geral de qual técnica ou combinação de técnicas de imputação foi a melhor para preencher os dados ausentes tanto com as *Estratégias de Soares* quanto as *Estratégias Propostas*, se trata de uma contagem total de técnicas ou combinação de técnicas vencedoras de cada base de dados com percentuais de ausência diferentes.

Nos resultados a base *Iris Plants* teve a técnica de agrupamento como melhor resultado em 80% dos casos nas *Estratégias de Soares* enquanto nas *Estratégias Propostas* o agrupamento teve uma queda de 5% devido a seleção com *Algoritmos Genéticos*, porém não foi o suficiente para vencer o agrupamento já que a seleção apareceu como melhor em apenas 15% dos casos.

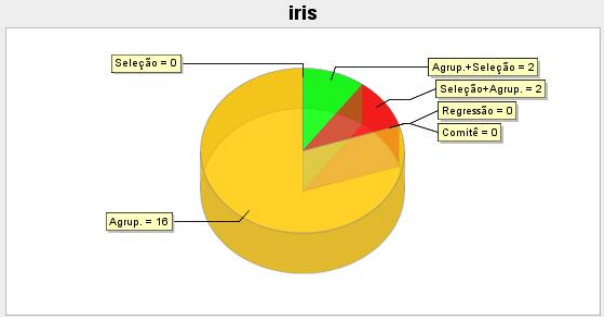
Semelhantemente, a base *Pima Indians Diabetes* obteve um percentual muito próximo a base *Iris Plants* na técnica de agrupamento, 70%, nas *Estratégias Propostas*.

Diferentemente das *Estratégias de Soares* em que a técnica de agrupamento teve sucesso em apenas 37,5% na base *Pima Indians Diabetes*. Percebe-se então que o agrupamento se manteve como melhor técnica nesta base de dados devido o uso de Redes de Kohonen.

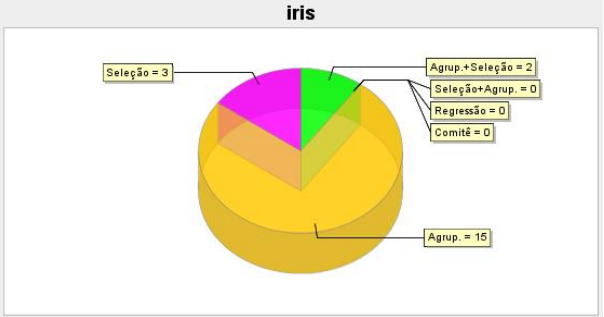
Na base *Wisconsin Breast Cancer* o agrupamento também obteve melhores resultados só que em mais de 90% dos casos tanto nas *Estratégias de Soares* quanto nas *Estratégias Propostas*, entretanto a taxa de erro foi alto nos melhores casos, o que veremos nos gráficos posteriores.

Nesta simples análise vemos que a redução dos registros de dados através do agrupamento antes da imputação de dados é de grande importância na qualidade do processo de complementação de dados ausentes e que o uso de agrupamento com Redes de Kohonen não afetou os resultados obtidos pelas *Estratégias de Soares,* pelo contrário, melhorou esses resultados na base *Pima Indians Diabetes*.

Iris Plants

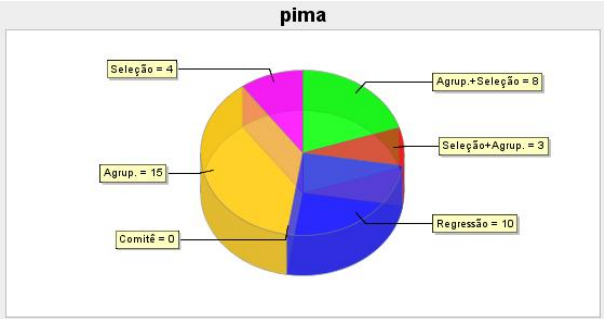


Estratégia de Soares

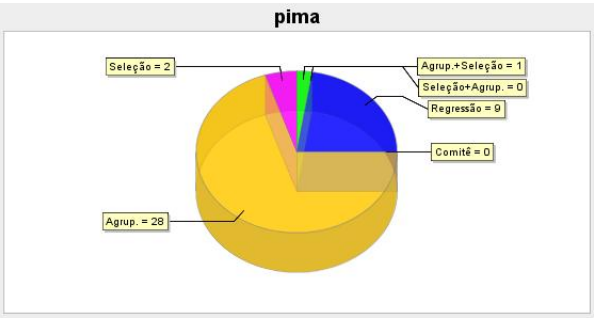


Estratégia Proposta

Pima Indians Diabetes

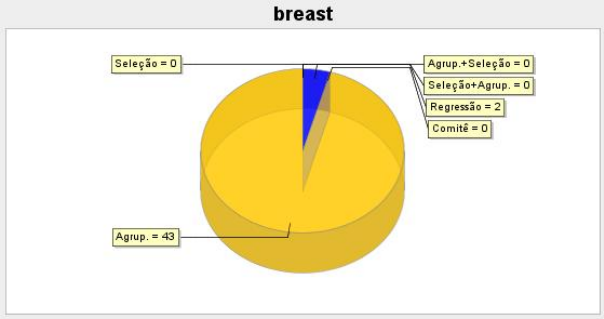


Estratégia de Soares

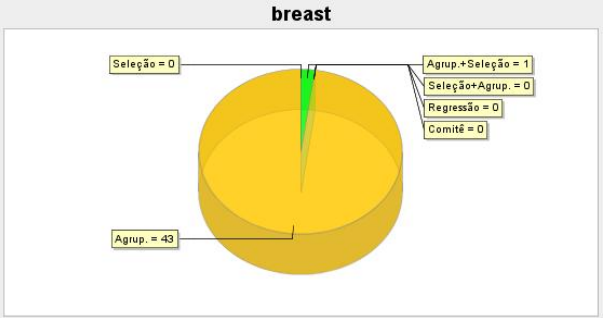


Estratégia Proposta

Wisconsin Breast Cancer



Estratégia de Soares



Estratégia Proposta

5.3.2 Gráficos: Comparação das Estratégias por percentual de ausência

Com a análise de técnicas ou combinação de técnicas por percentual de ausência teremos uma visão de qual técnica ou combinação foi a melhor para preencher os dados ausentes em cada base de dados com percentuais diferentes de ausência tanto para a *Estratégia de Soares* quanto para *Estratégia Proposta*.

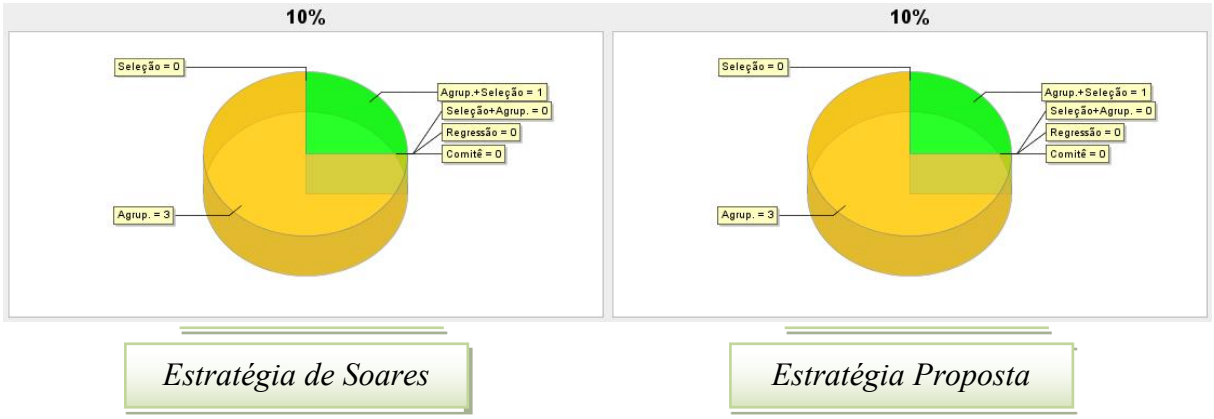
Nessa análise busca-se verificar se a quantidade de dados ausentes por coluna de cada base de dados impacta nos resultados das técnicas, pois quanto maior o percentual de dados ausentes, menos informações poderão ser utilizadas pelas técnicas para complementação dos dados e imaginamos que a imputação com *Back Propagation* e *k-NN* podem sofrer impactos. Além disso, o tamanho dos grupos do agrupamento com *K-Means* e *Redes de Kohonen* podem ser significantemente alterados.

Mas com os gráficos vemos que a base *Iris Plants* não foi impactada nas duas *Estratégias, de Soares* e *a Proposta*, em todos os cinco percentuais de ausência o agrupamento foi considerado a melhor estratégia. Possivelmente os resultados constantes se devem as características dessa base de dados: pouca variação entre os valores, alta correlação entre grande parte dos atributos e poucos registros. Observando mais atentamente, a *Estratégia Proposta* apresentou resultados mais constantes do que a *Estratégia de Soares*, possivelmente pela contribuição do agrupamento com *Redes de Kohonen*, o que poderemos constatar mais adianta durante a análise do ranking estratégias.

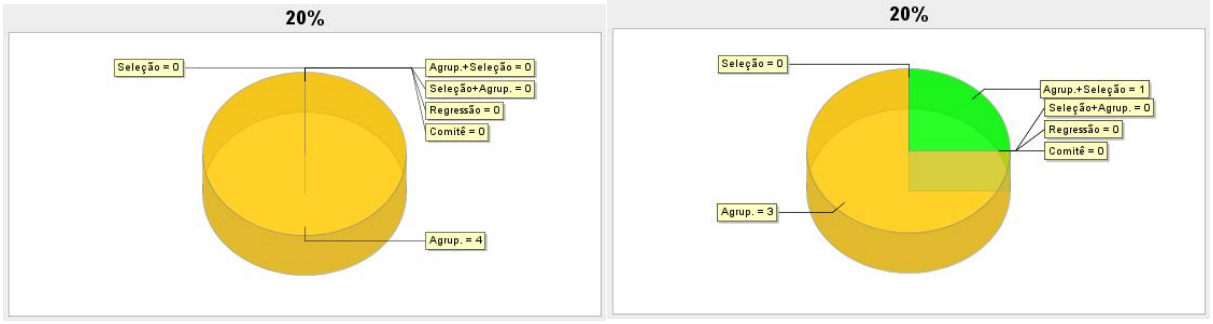
Em todos percentuais o agrupamento foi o vencedor na base *Pima Indians Diabetes* na *Estratégia Proposta,* no entanto a *Estratégia de Soares* apresentou muitas variações nos resultados por percentual de ausência: a técnica de agrupamento venceu em dois percentuais de ausência e empatou com regressão simples em outros dois percentuais. SOARES (2007) ressalta o desempenho da técnica de seleção, e aqui ressaltamos o impressionante desempenho da técnica de agrupamento, especificamente de *Redes de Kohonen*, que manteve obteve melhor qualidade de imputação em todos os percentuais.

As duas *Estratégias* se mostraram muito semelhantes nos resultados da base *Wisconsin Breast Cancer* pois que a técnica de agrupamento venceu em todos os percentuais de ausência.

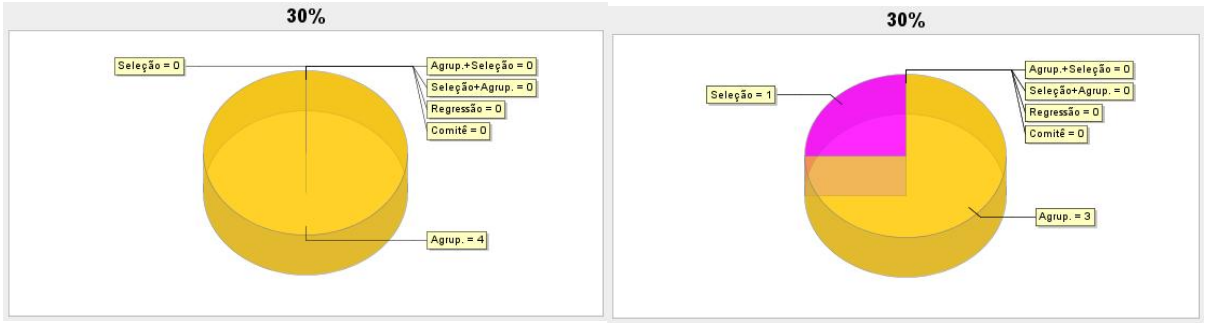
Iris Plants



Estratégia de Soares *Estratégia Proposta*

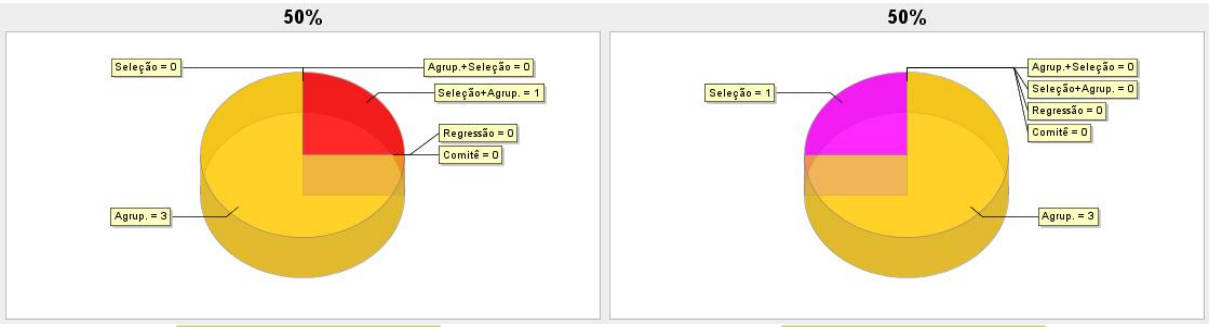


Estratégia de Soares *Estratégia Proposta*



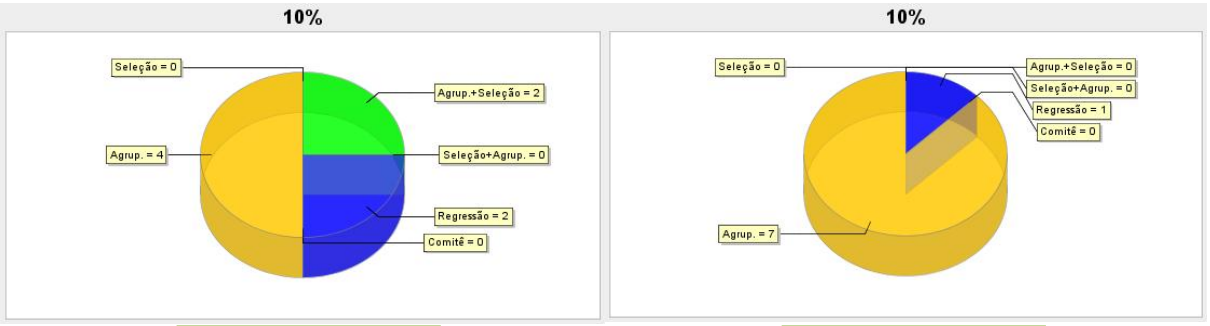
Estratégia de Soares *Estratégia Proposta*

Estratégia de Soares *Estratégia Proposta*

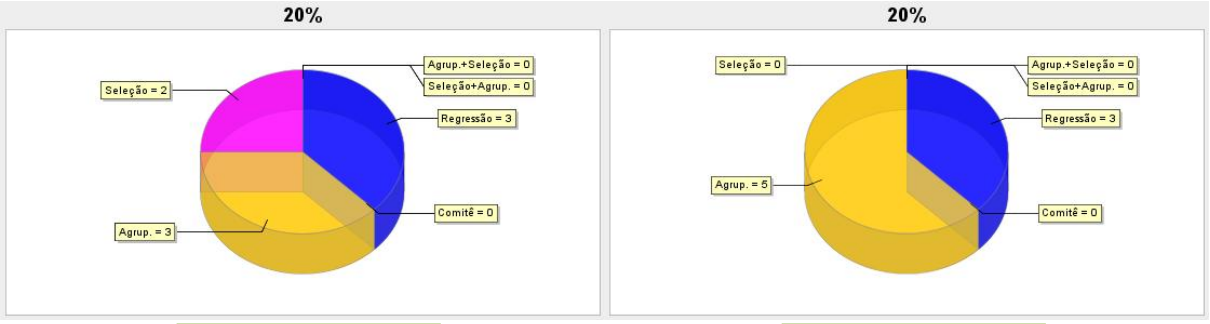


Estratégia de Soares *Estratégia Proposta*

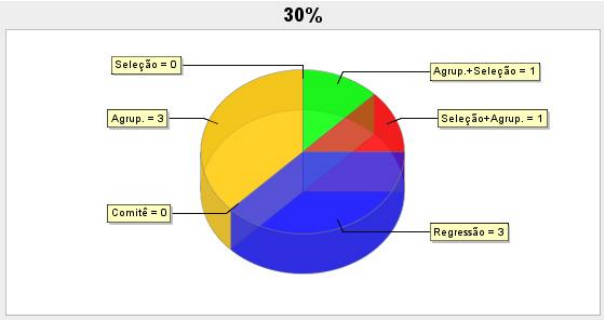
Pima Indians Diabetes



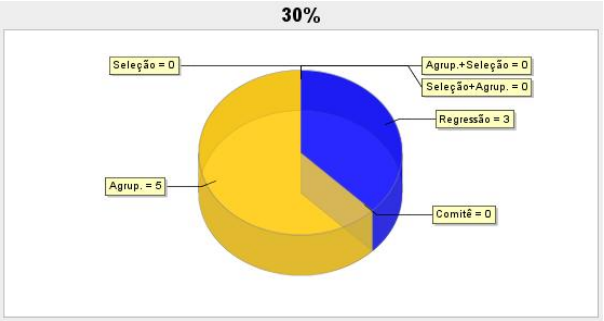
Estratégia de Soares *Estratégia Proposta*



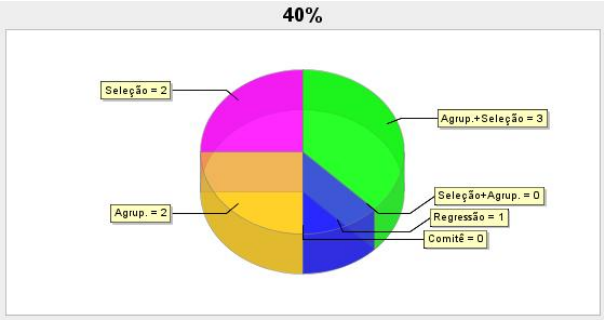
Estratégia de Soares *Estratégia Proposta*



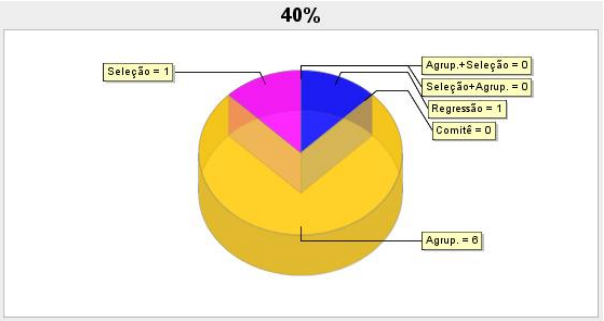
Estratégia de Soares



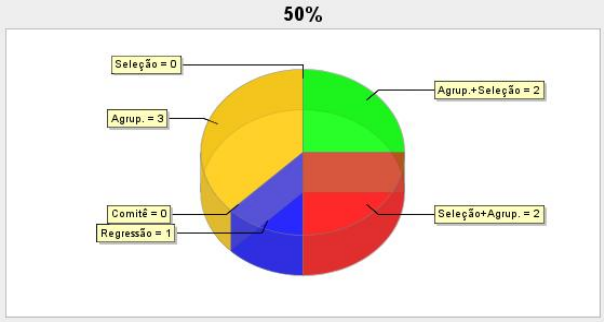
Estratégia Proposta



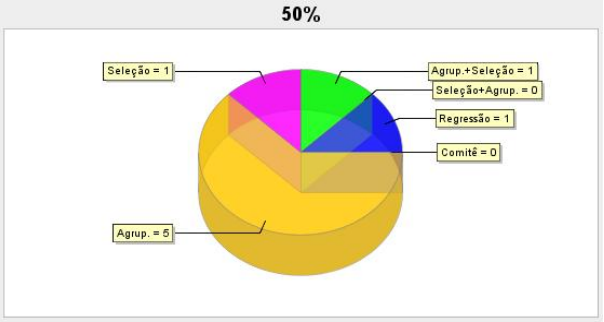
Estratégia de Soares



Estratégia Proposta

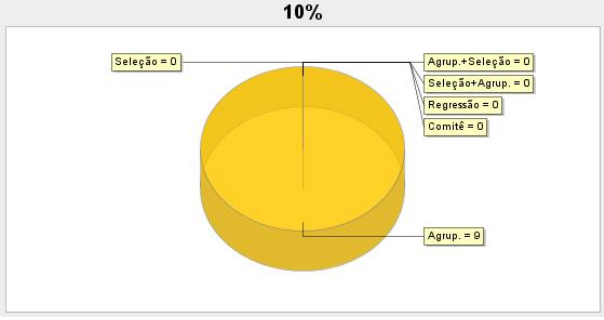


Estratégia de Soares

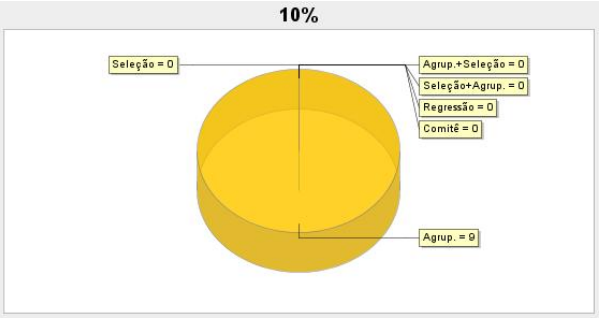


Estratégia Proposta

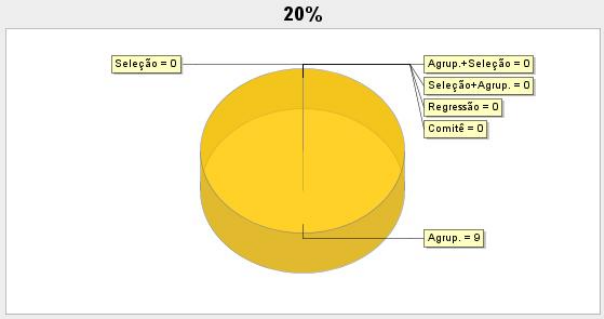
Wisconsin Breast Cancer



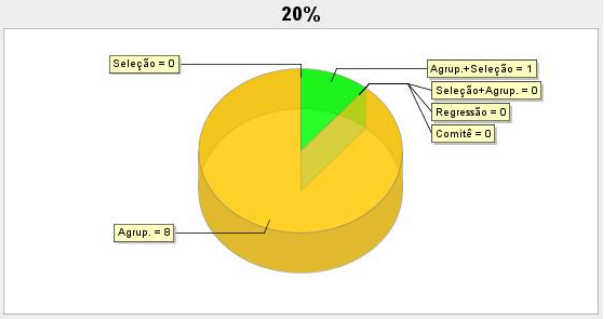
Estratégia de Soares



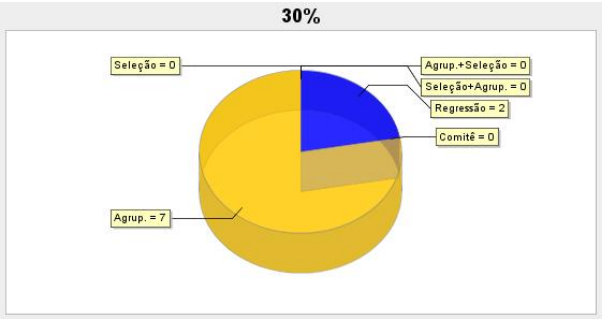
Estratégia Proposta



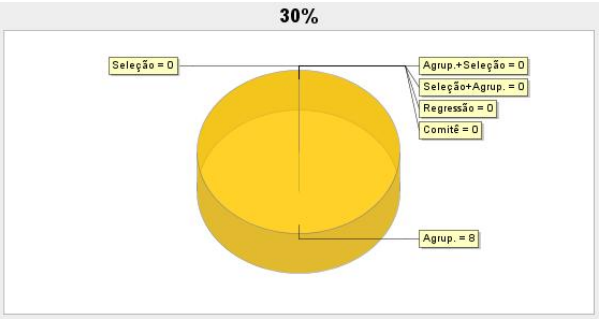
Estratégia de Soares



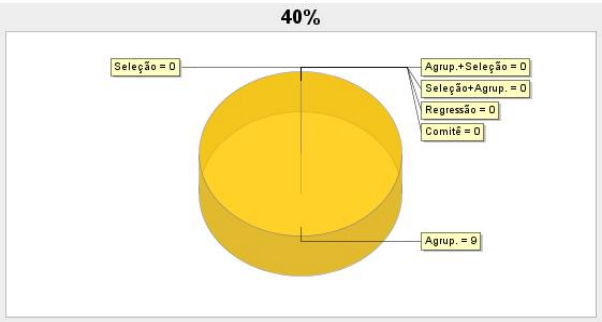
Estratégia Proposta



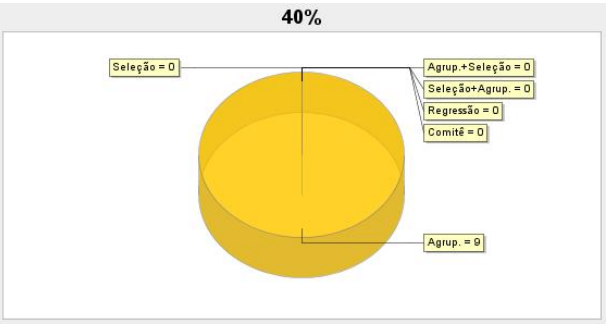
Estratégia de Soares



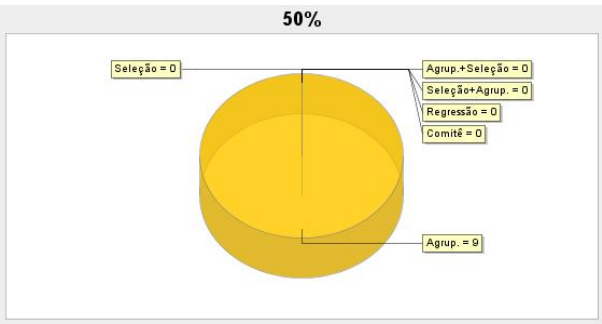
Estratégia Proposta



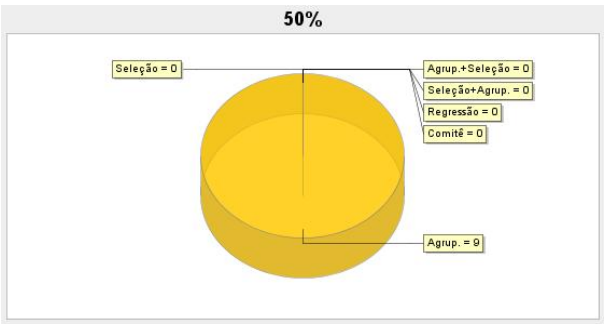
Estratégia de Soares



Estratégia Proposta



Estratégia de Soares



Estratégia Proposta

5.3.3 Gráficos: Comparação das Estratégias por ranking

Nessa análise saberemos quais foram as melhores técnicas ou combinações de técnicas dentre as estratégias utilizadas ao avaliar os gráficos de rankeamento que mostram percentuais de colocação de cada estratégia da *Estratégia de Soares* e da *Estratégia Proposta*.

A partir dos gráficos logo vemos que a simples imputação com *Média* foi a última colocada na base *Iris Plants*, apresentando maiores casos de erro. Já a técnica de agrupamento com *K-Means* seguida de imputação com *k-NN* foi a vencedora com um percentual ótimo de 75% na primeira colocação, depois dela a técnica de agrupamento com *K-Means* seguida de seleção com *AG* e imputação com *k-NN* também obteve quantidade de acerto impressionante, 50%. Outras combinações de técnicas envolvendo sempre agrupamento com ou sem seleção tiveram resultados satisfatórios sendo que *Redes de Kohonen* e *K-Means* estão empatados por cada um aparecer em seis de 12 combinações com melhores resultados enquanto *AG* vence *PCA* por aparecer em quatro de seis combinações dos melhores resultados envolvendo seleção.

Analisando os gráficos da base *Pima Indians Diabetes* vemos que a *Média* também foi a última colocada porém não foi em disparada pois que as outras estratégias tiveram um desempenho geral regular. Novamente as estratégias envolvendo agrupamento antes da imputação tiveram bons resultados e em todos os melhores casos a imputação com *k-NN* se sobressaiu estando presente em seis das seis melhores combinações. Em três melhores combinações envolvendo seleção, *AG* aparece em duas, enquanto Redes de Kohonen aparece em duas combinações mas dentre as seis melhores.

Já a base *Wisconsin Breast Cancer* obteve mais de uma estratégia com pior resultado e em nenhum delas a *Média* se encontra. E novamente ressaltamos a influência das redes *Redes de Kohonen* nos melhores resultados, pois essa técnica aparece em quatro das oito combinações com melhores resultados da base *Wisconsin Breast Cancer*. Nessa base tivemos três melhores colocados: em primeiro com 55% o agrupamento com *K-Means* seguindo de imputação com *k-NN*, em segundo com 40% o agrupamento com *Redes de Kohonen* seguido de imputação com *k-NN* e em terceiro com 30% agrupamento com *K-Means* seguindo de imputação com *k-NN*. SOARES (2007) havia demonstrado ser instável a imputação com *k-NN* porém com melhores resultados do que as imputações com *Média* e *Back Propagation*, mas na *Estratégia Proposta* nas três bases de dados a imputação com *k-NN* esteve mais presente

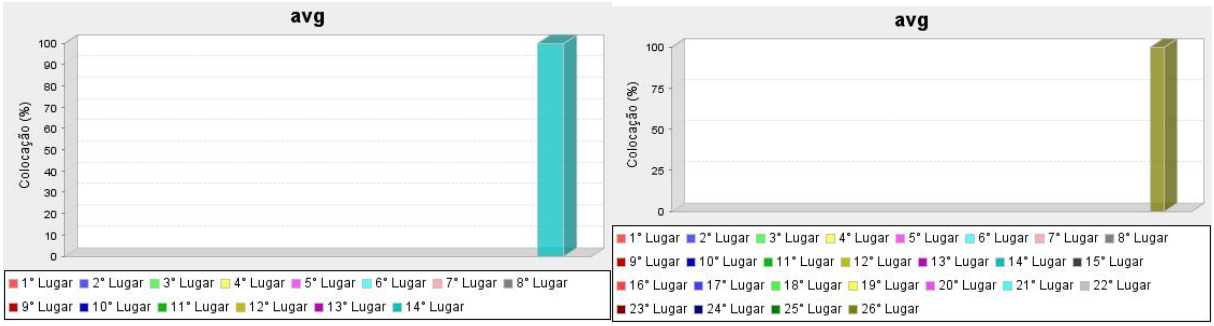
nas estratégias com melhores resultados, sendo que na base *Pima Indians Diabetes* teve também bom resultado com apenas a imputação simples com *k-NN.*

E claramente observamos a contribuição do agrupamento com *Redes de Kohonen* no aumento do percentual de acerto de imputação e a seleção com *AG* proporcionando também uma melhora na qualidade da imputação só que com uma participação mais reduzida.

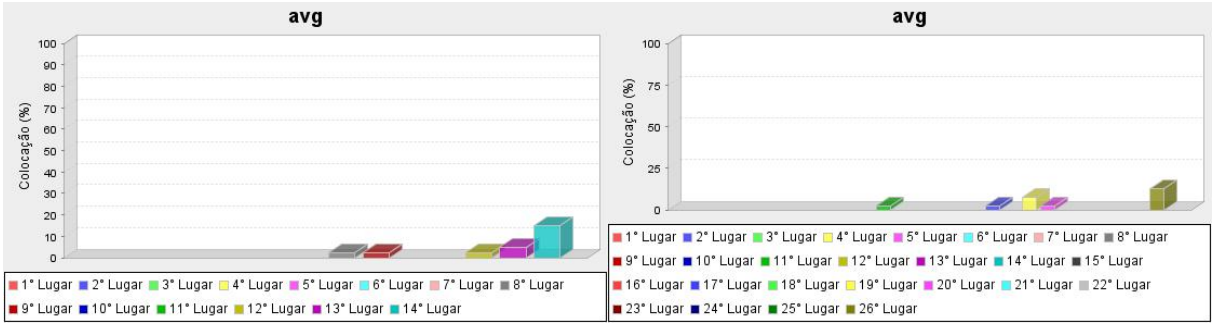
Estratégia 1: Imputação com *Média*

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* foram similares, ambos ficaram próximos das últimas colocações no ranking. As bases *Iris Plants* teve o pior desempenho, máximo grau na última colocação do ranking.

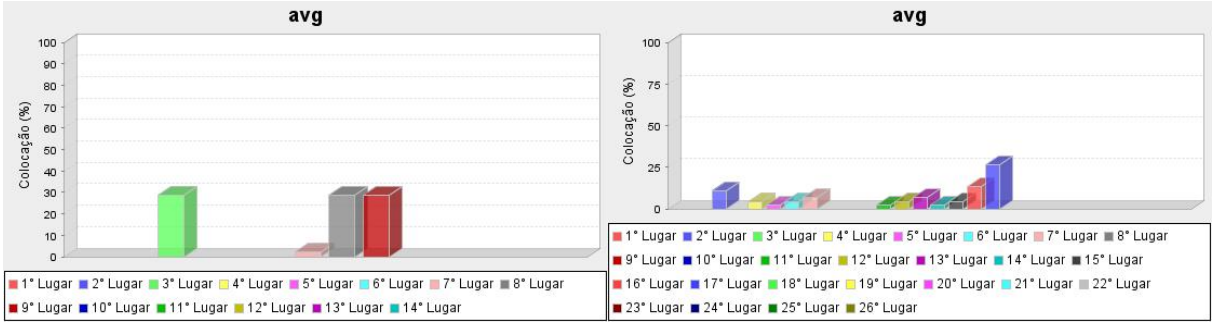
Iris Plants



Pima Indians Diabetes



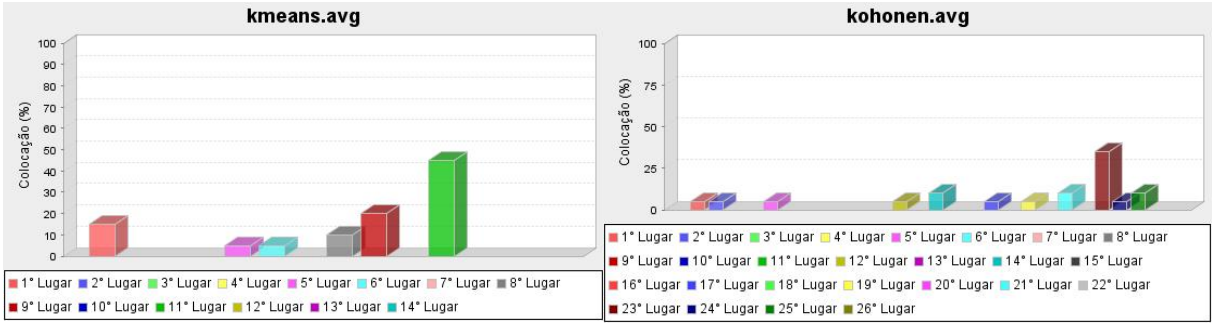
Wisconsin Breast Cancer



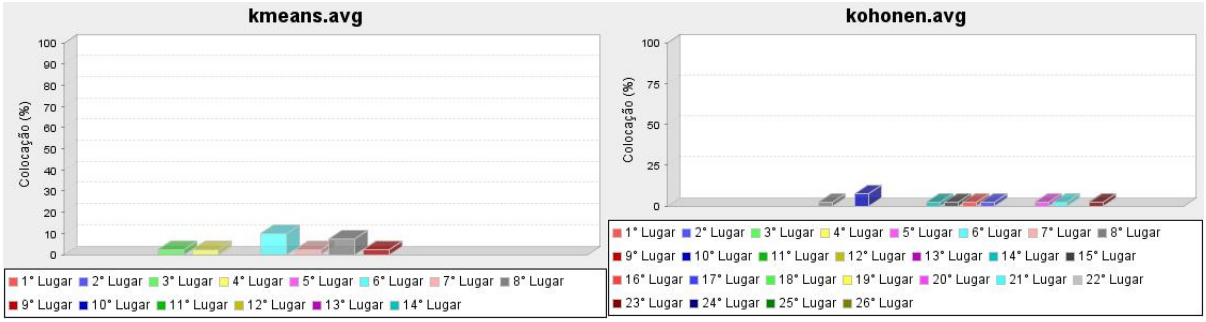
Estratégia 2: Agrupamento e Imputação com *Média*

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* foram similares nas bases *Iris Plants* e *Wisconsin Breast Cancer* onde alcançaram a primeira colocação, sendo que *Estratégia de Soares* se sobressaiu ao resultar um percentual mais alto no agrupamento com *K-Means* do que *Redes de Kohonen*. A base *Pima Indians Diabetes* não obteve bons resultados com *K-Means* nem com *Redes de Kohonen*.

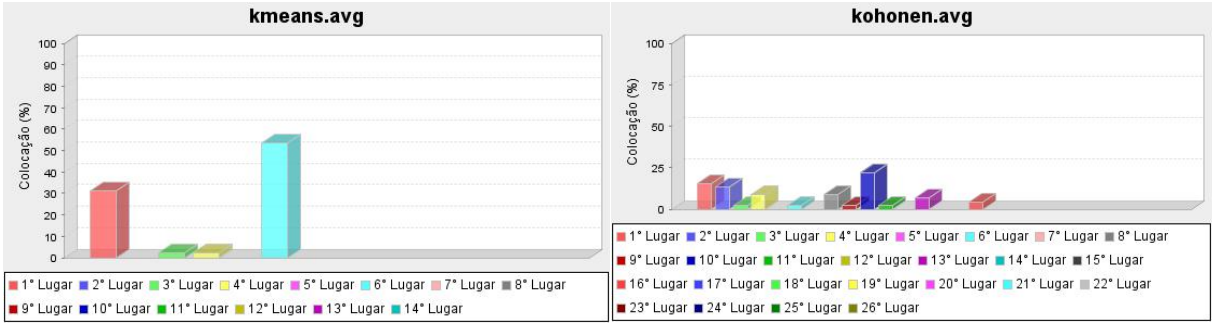
Iris Plants



Pima Indians Diabetes



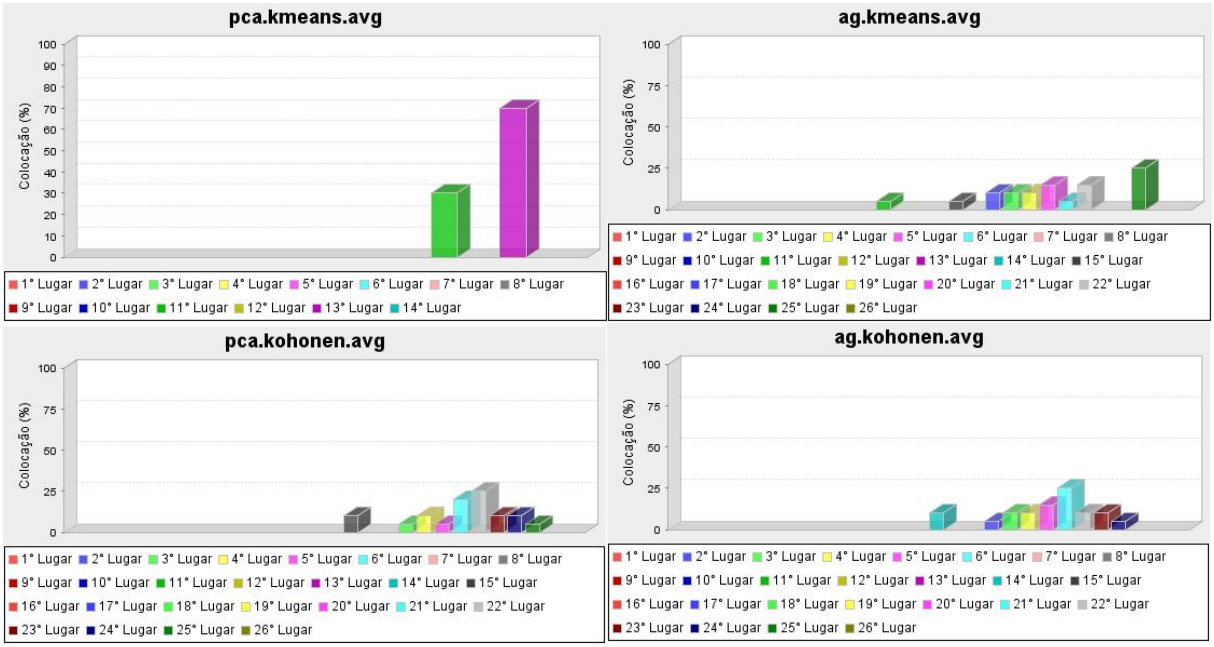
Wisconsin Breast Cancer



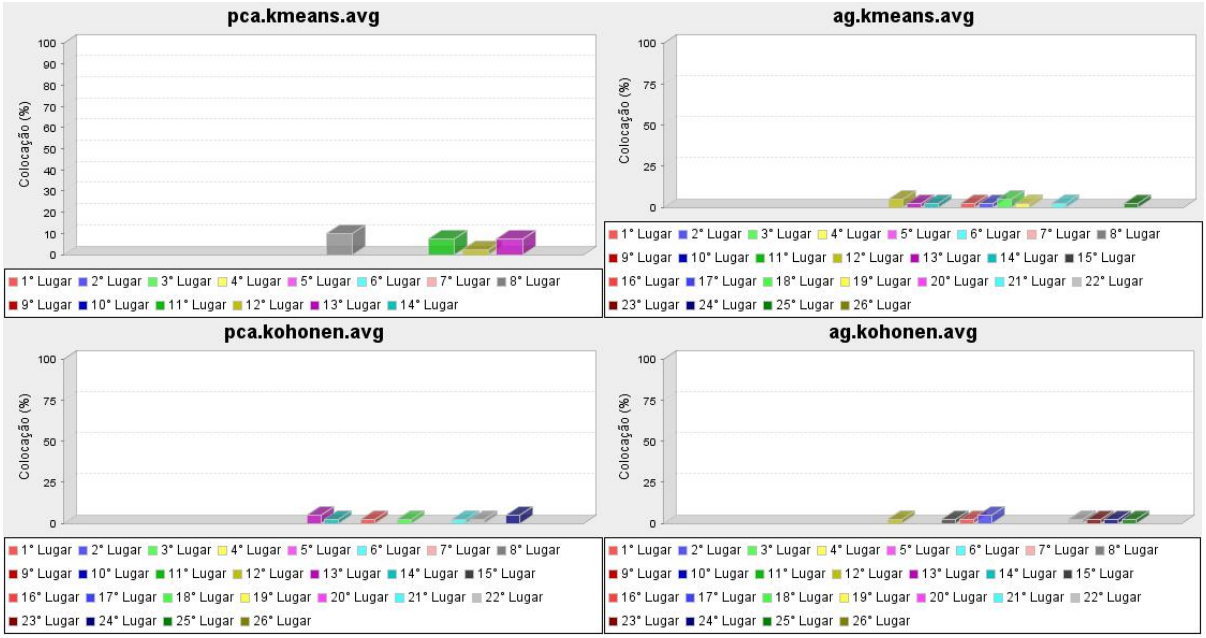
Estratégia 3: Seleção, Agrupamento e Imputação com *Média*

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* foram similares, ambos ficaram próximos das últimas colocações no ranking.

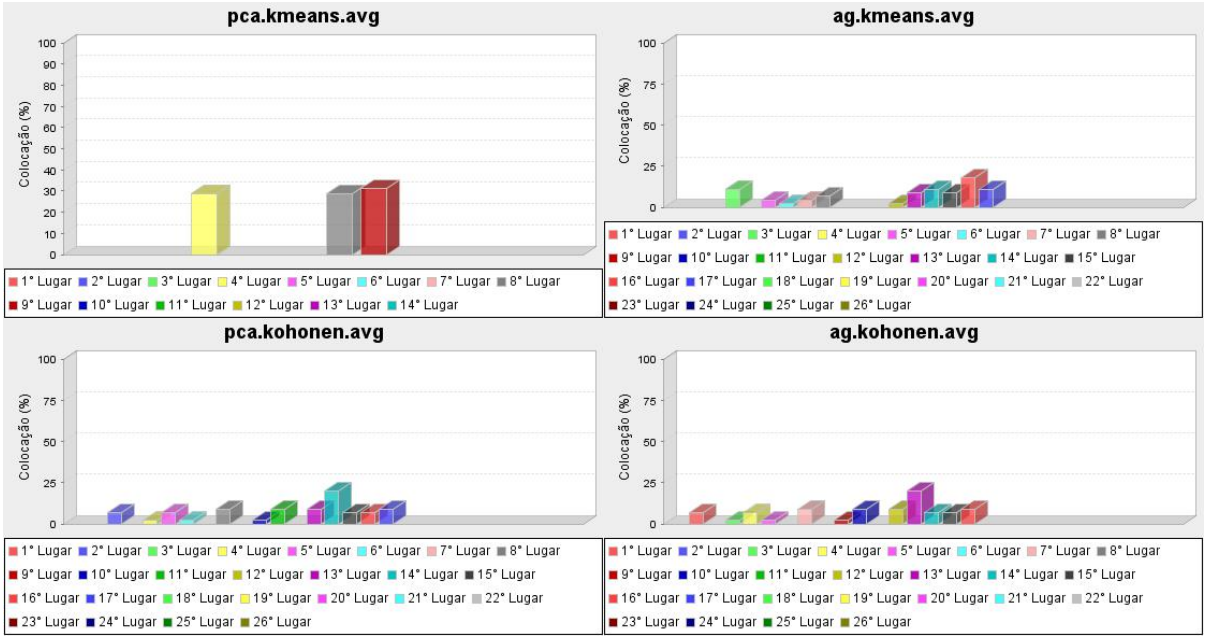
Iris Plants



Pima Indians Diabetes



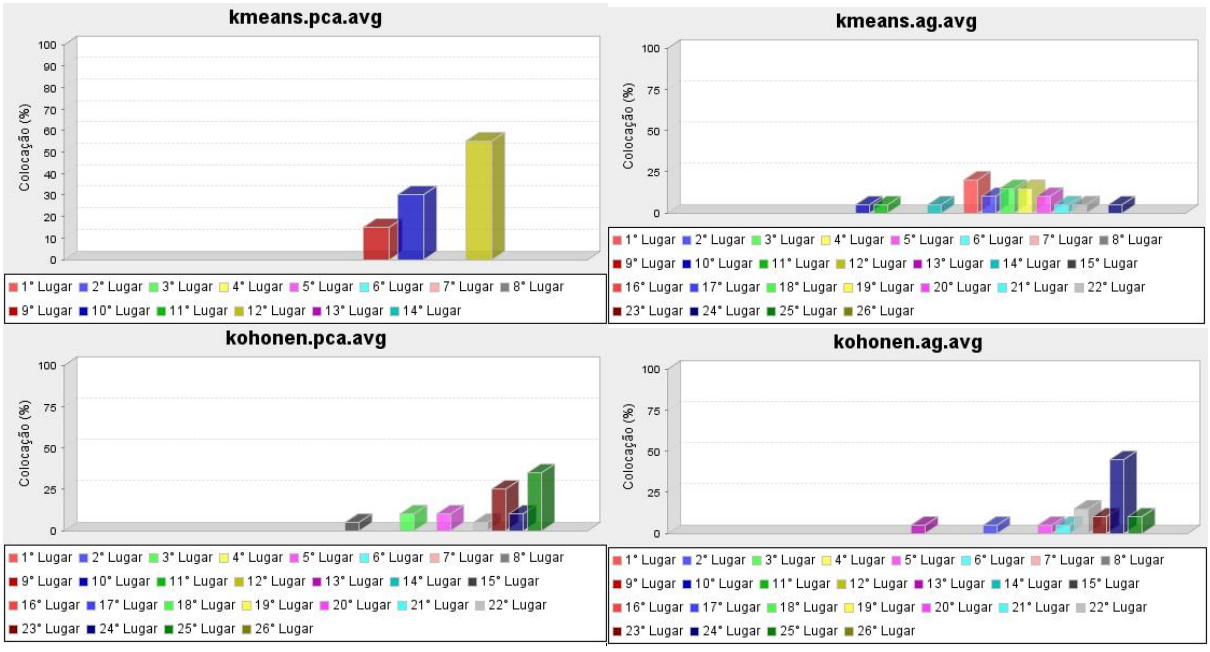
Wisconsin Breast Cancer



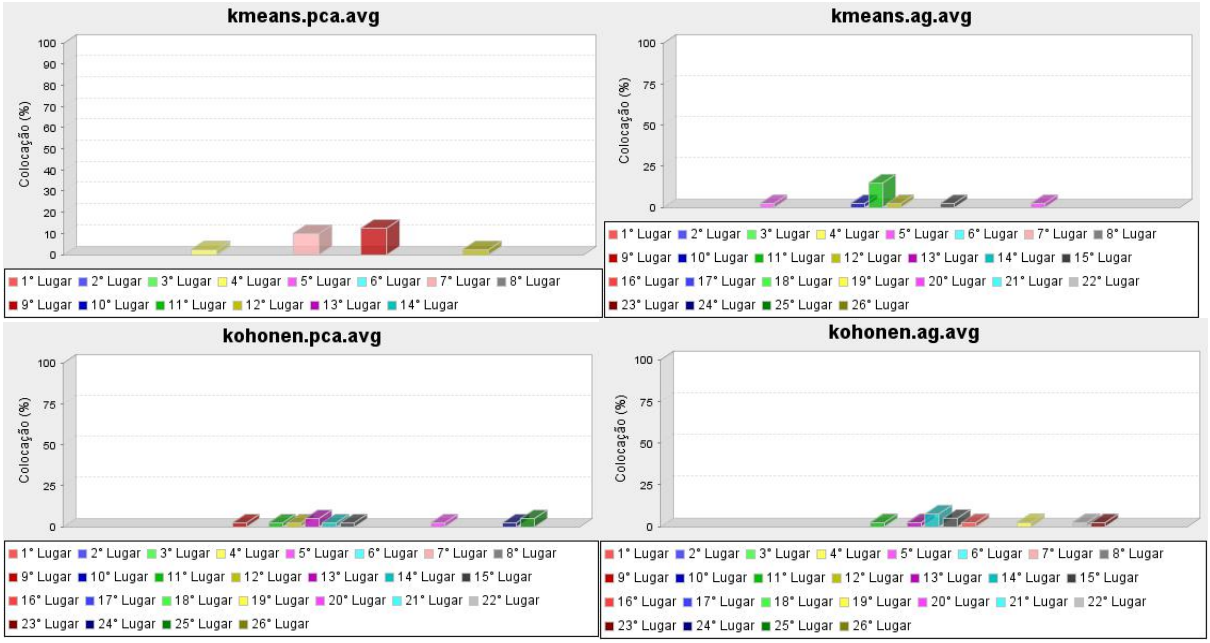
Estratégia 4: Agrupamento, Seleção e Imputação com Média

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* foram similares, ambos ficaram próximos das últimas colocações no ranking bases *Iris Plants* e *Pima Indians Diabetes*. Na base *Wisconsin Breast Cancer* ambos ficaram próximos das primeiras colocações, até no 1º lugar com agrupamento com *K-Means* e com *Redes de Kohonen* seguido de seleção com *AG*.

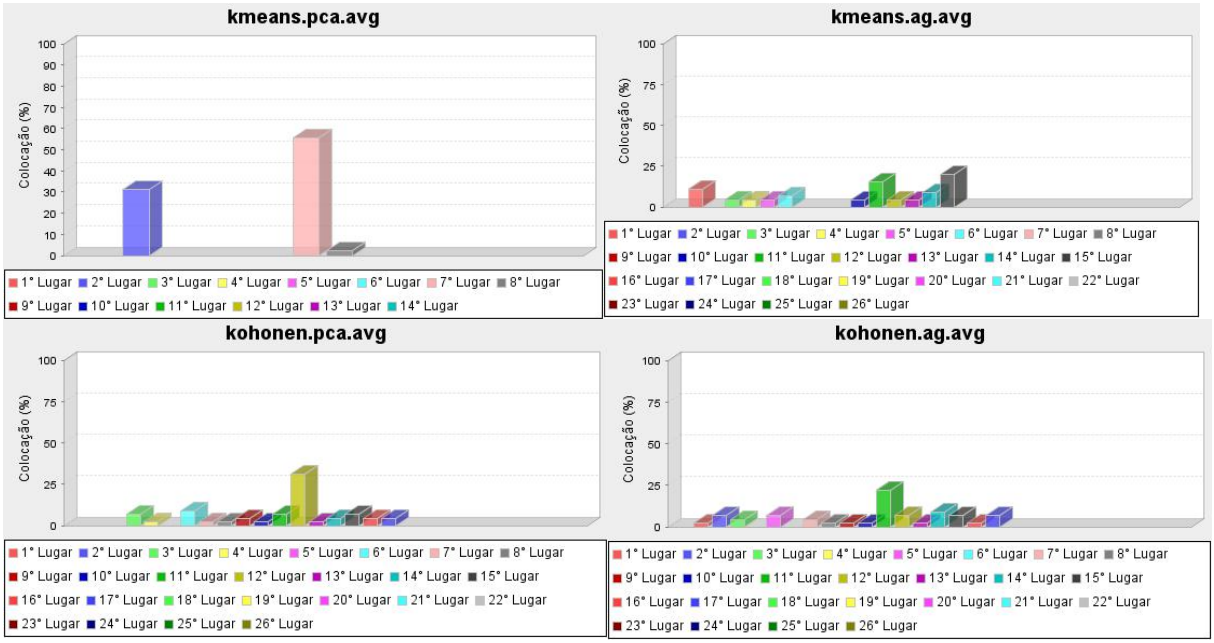
Iris Plants



Pima Indians Diabetes



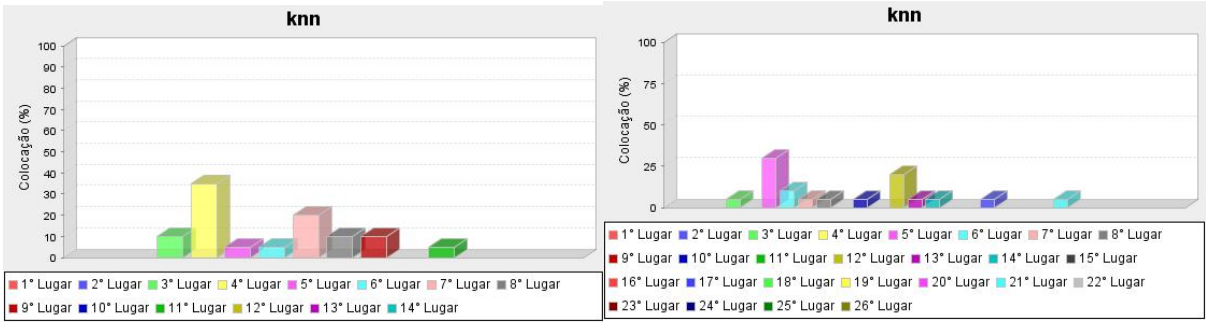
Wisconsin Breast Cancer



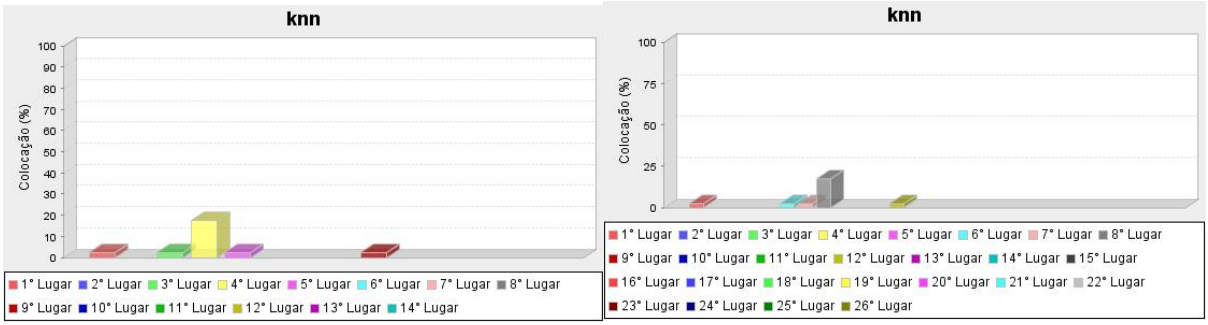
Estratégia 5: Imputação com *k*-NN

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* foram similares, ambos ficaram próximos da primeira colocação no ranking.

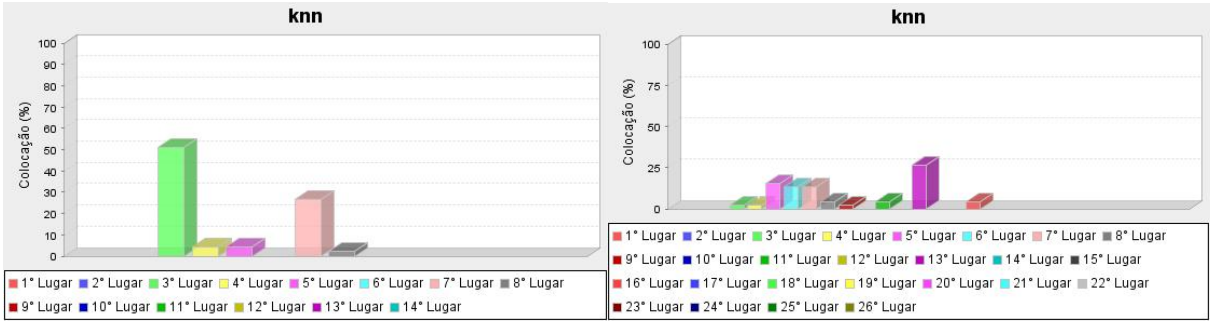
Iris Plants



Pima Indians Diabetes



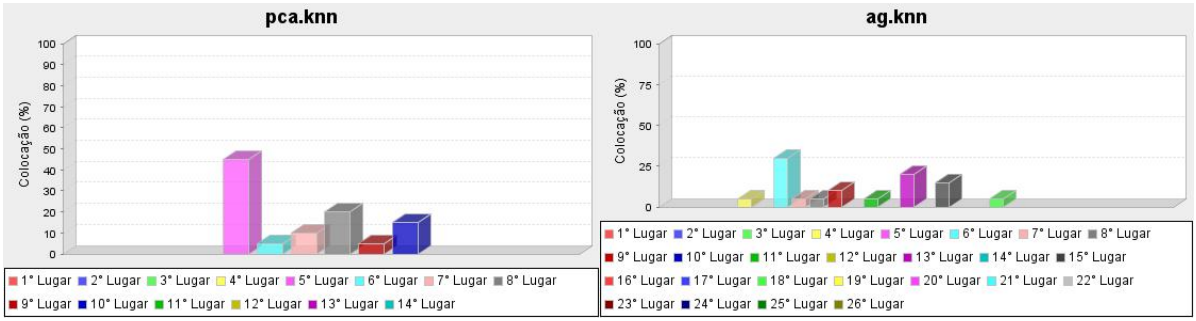
Wisconsin Breast Cancer



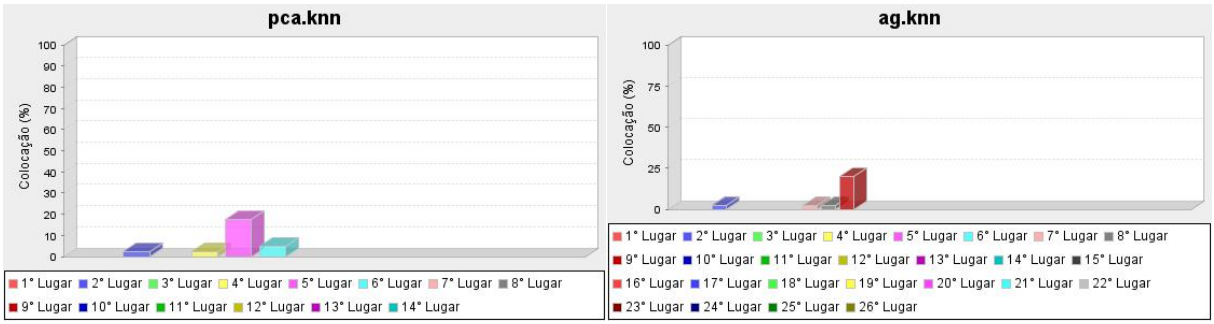
Estratégia 6: Seleção e Imputação com *k*-NN

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* foram regulares, não alcançaram os primeiros lugares e nem os últimos. A base *Iris Plants* e *Wisconsin Breast Cancer* apresentaram uma pequena melhora na colocação com a *Estratégia Proposta* através da seleção com *AG* e imputação com *k-NN*.

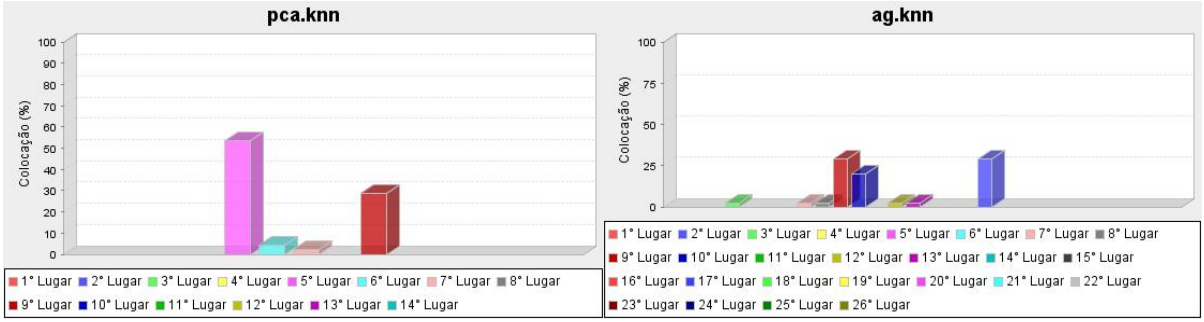
Iris Plants



Pima Indians Diabetes



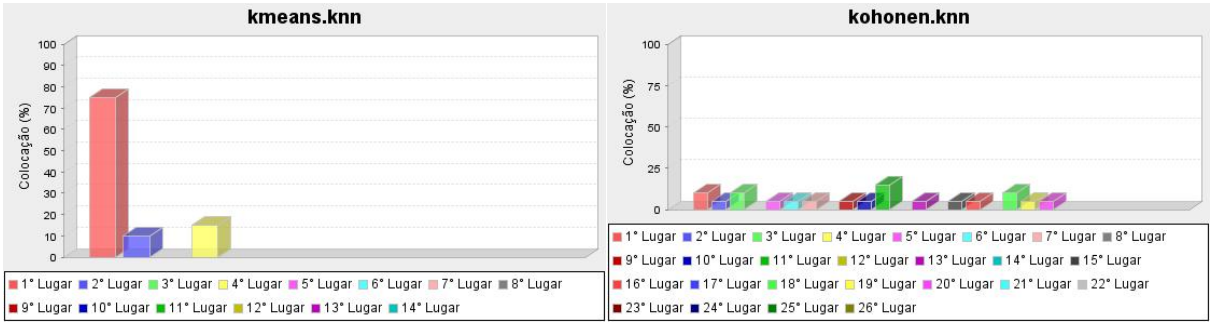
Wisconsin Breast Cancer



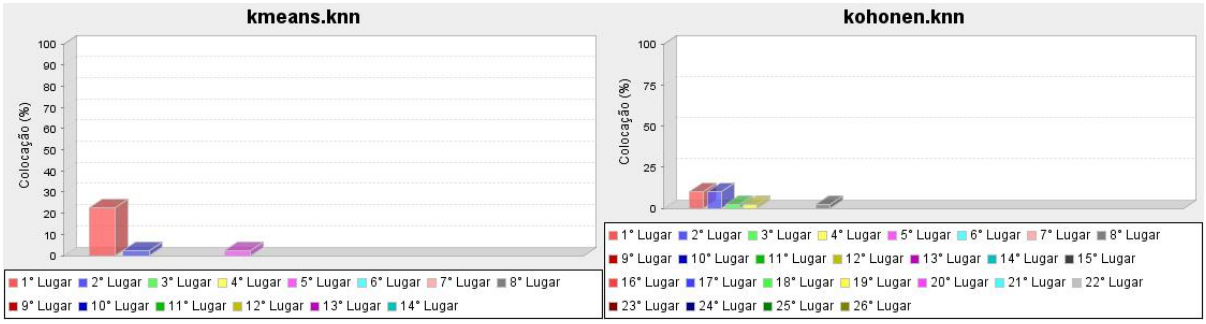
Estratégia 7: Agrupamento e Imputação com *k*-NN

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* foram excelentes, ambos ficaram na primeira colocação no ranking sendo que a *Estratégia de Soares* se sobressaiu pois o agrupamento com *K-Means* apresentou os maiores graus de colocação em todas as bases de dados.

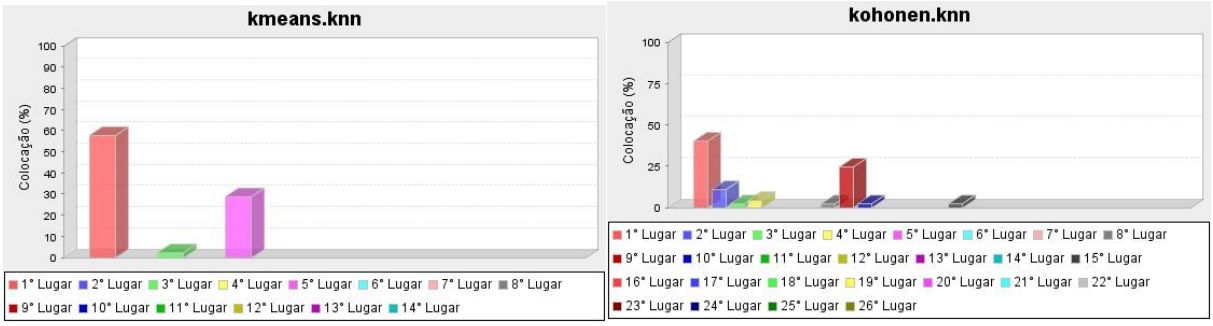
Iris Plants



Pima Indians Diabetes



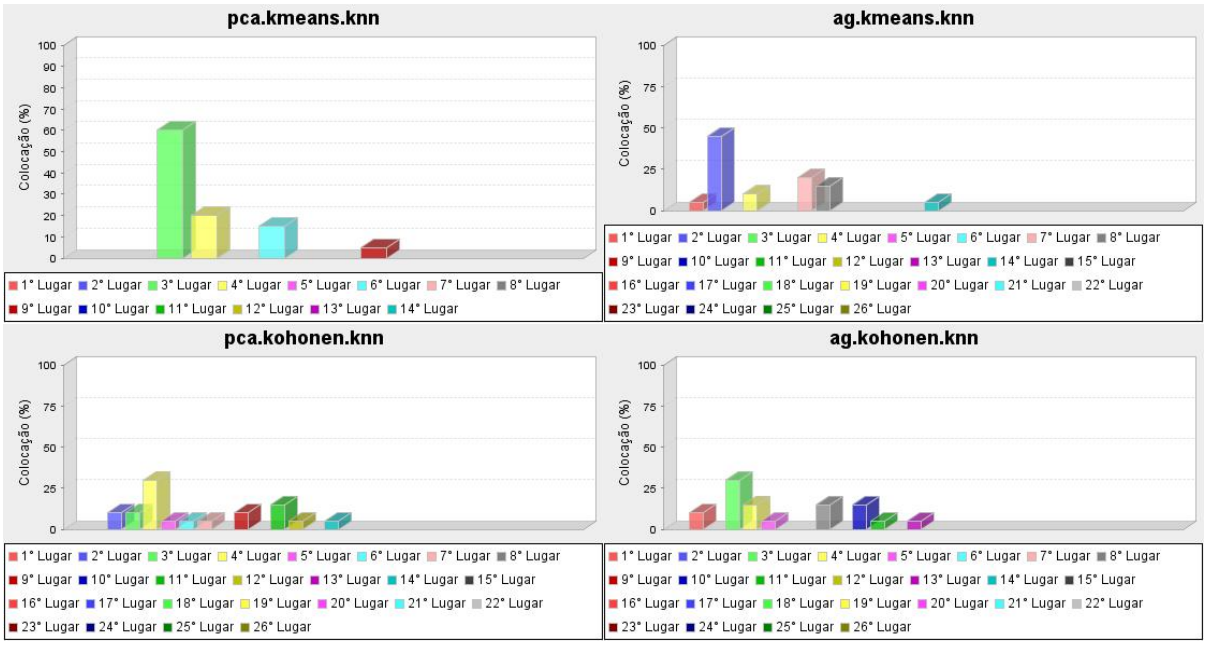
Wisconsin Breast Cancer



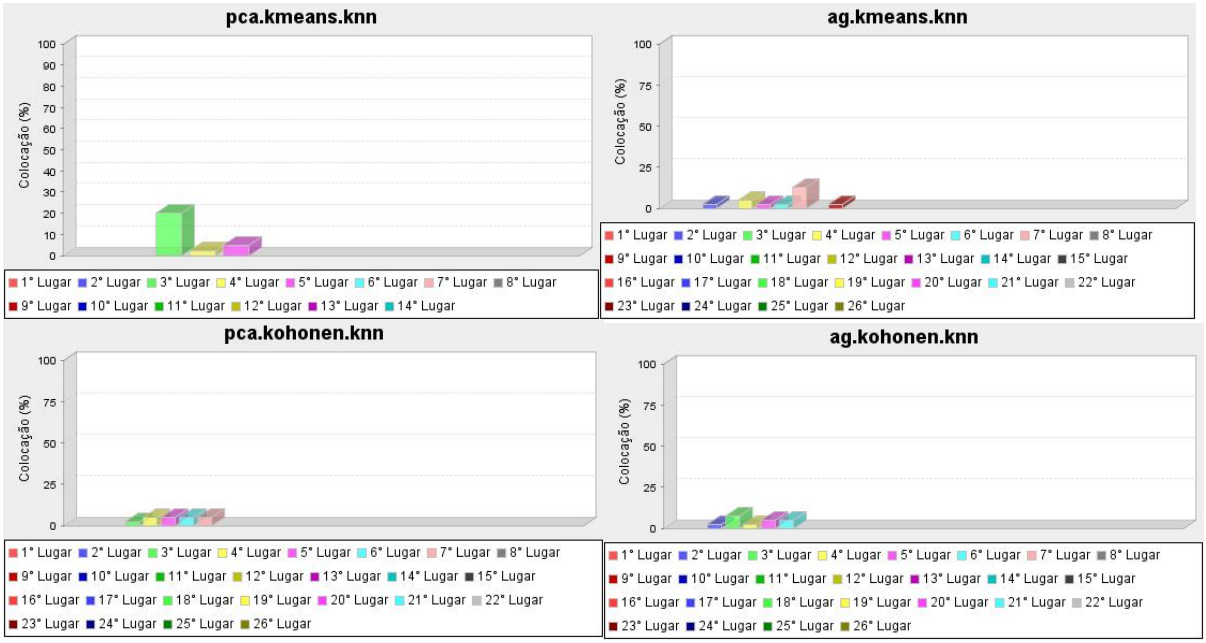
Estratégia 8: Seleção, Agrupamento e Imputação com *k*-NN

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* se mostraram próximos da primeira colocação no ranking. Em todas as bases de dados a seleção com *PCA* seguida de agrupamento com *K-Means* e imputação com *k-NN* (*Estratégia de Soares*) não alcançou a primeira colocação, chegou ao máximo a terceira e segunda colocações, entretanto a seleção com *AG* seguida de agrupamento com *K-Means* ou *Redes de Kohonen* e imputação com *k-NN* (*Estratégia Proposta*) trouxe melhora nos resultados conseguindo o primeiro lugar nas bases *Iris Plants* e *Wisconsin Breast Cancer* e segundo na base *Pima Indians Diabetes*.

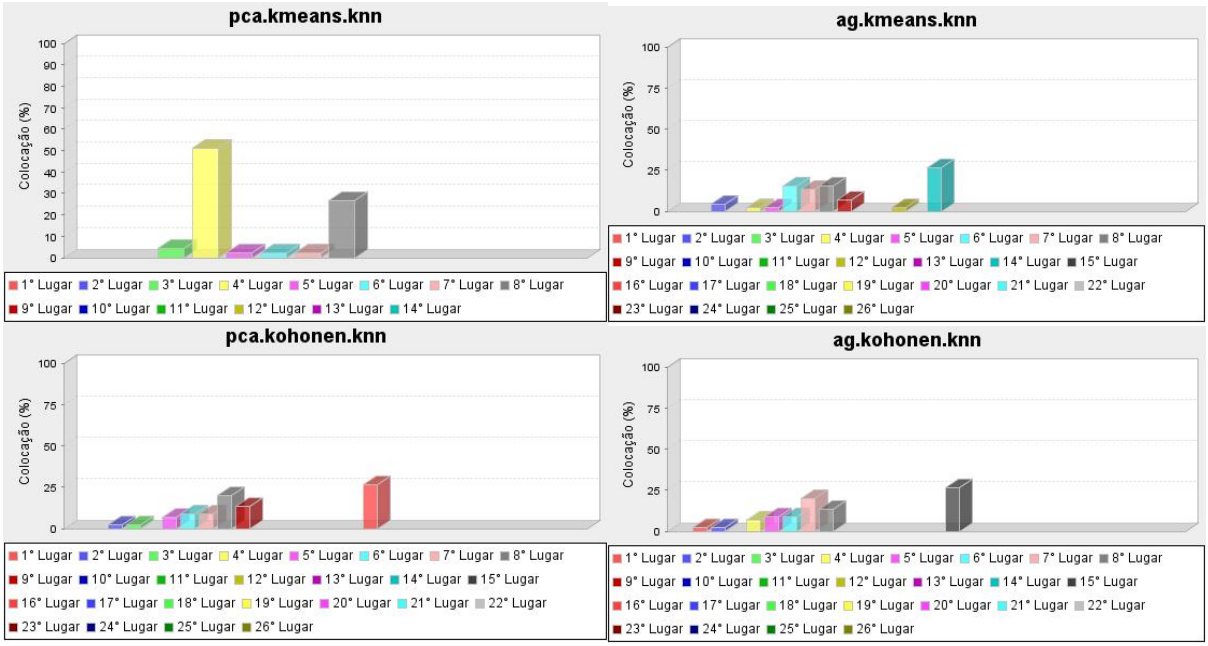
Iris Plants



Pima Indians Diabetes



Wisconsin Breast Cancer

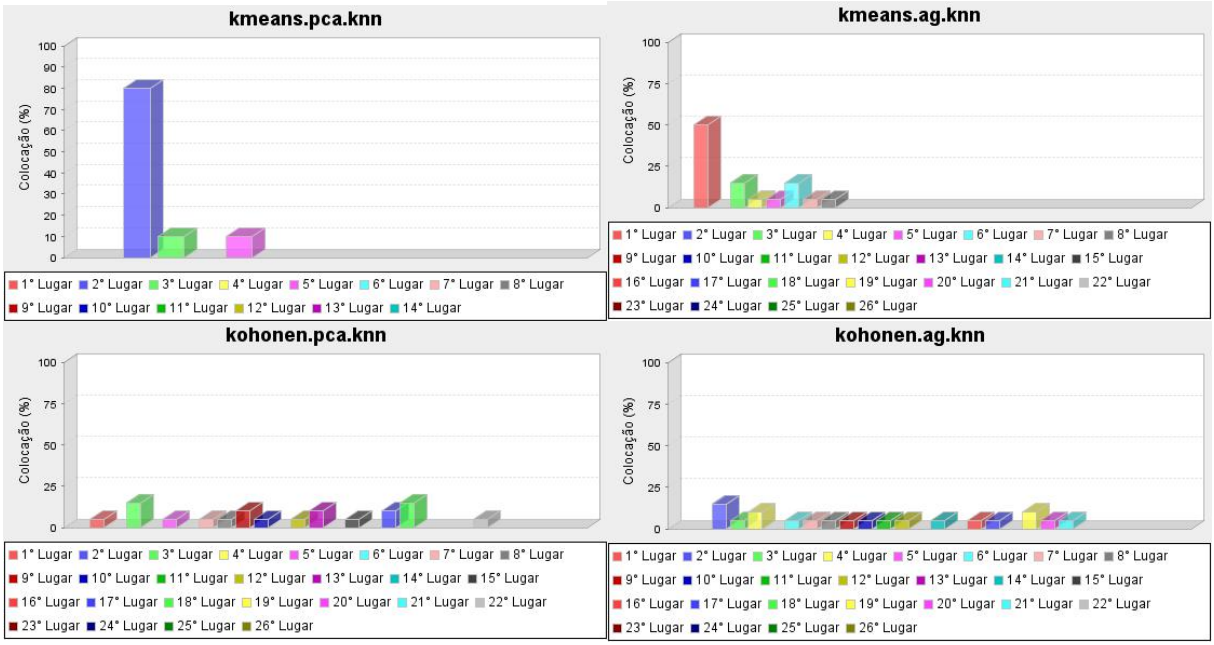


Estratégia 9: Agrupamento, Seleção e Imputação com *k*-NN

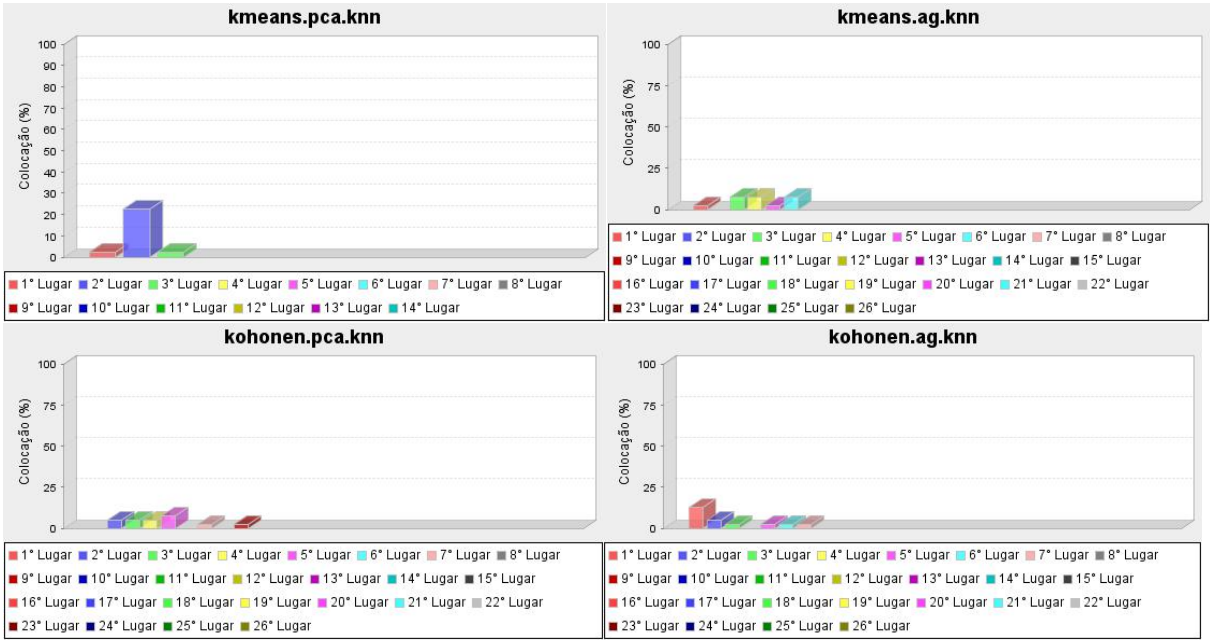
Os resultados da *Estratégia de Soares* e *Estratégia Proposta* estão sempre próximos da primeira colocação no ranking. Nas bases *Iris Plants* e *Wisconsin Breast Cancer*, o agrupamento com *K-Means* seguido de seleção com *PCA* e imputação com *k-NN* (*Estratégia de Soares*) trouxe alto percentual no segundo lugar do ranking, entretanto o uso de seleção com *AG* (*Estratégia de Proposta*) elevou a colocação para o primeiro lugar do ranking em elevado grau. Além disso, a *Estratégia de Proposta* também apresentou primeiro lugar em baixo grau no agrupamento com *Redes de Kohonen* seguido de seleção com *PCA* e imputação com *k-NN* na base *Iris Plants*.

Na base *Pima Indians Diabetes* o agrupamento com *K-Means* seguido de seleção com *PCA* e imputação com *k-NN* (*Estratégia de Soares*) obteve o mesmo grau de primeiro lugar no ranking do que o agrupamento com *K-Means* seguido de seleção com *AG* e imputação com *k-NN* (*Estratégia Proposta*), porém o agrupamento com *Redes de Kohonen* seguido de seleção com *AG* e imputação com *k-NN* (*Estratégia Proposta*) apresentou um grau maior na primeira colocação.

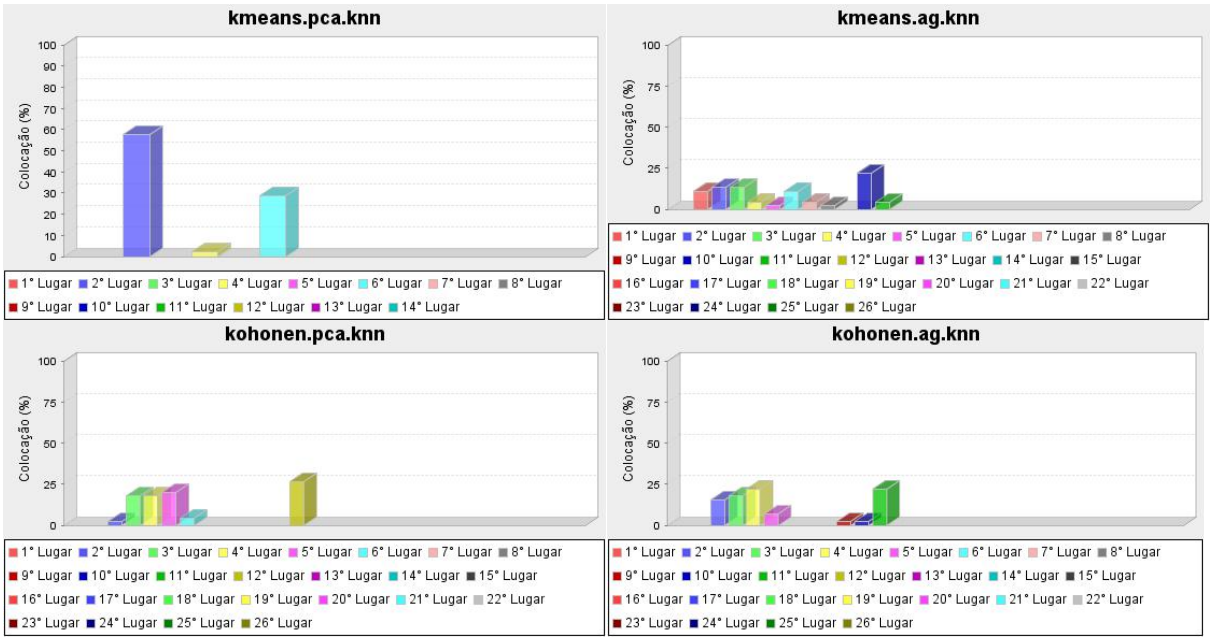
Iris Plants



Pima Indians Diabetes



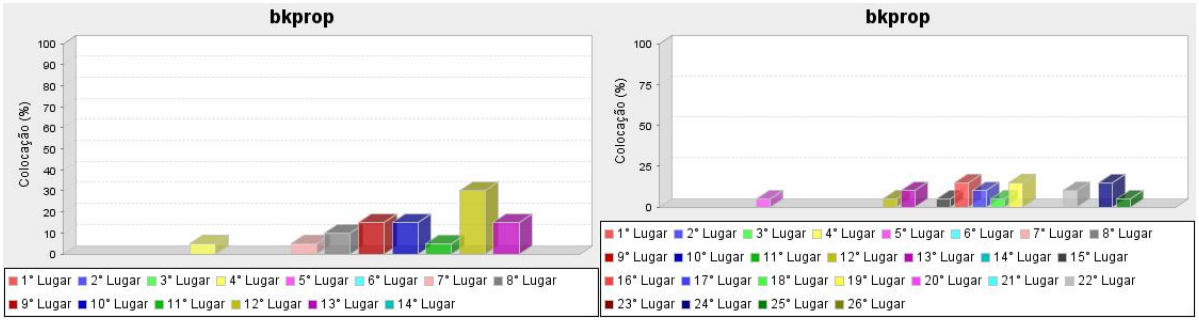
Wisconsin Breast Cancer



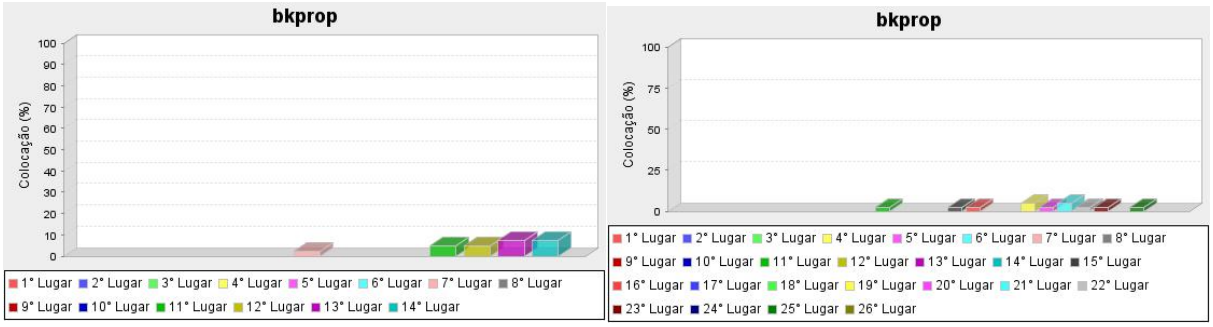
Estratégia 10: Imputação com *Back Propagation*

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* foram similares, distantes da primeira colocação no ranking.

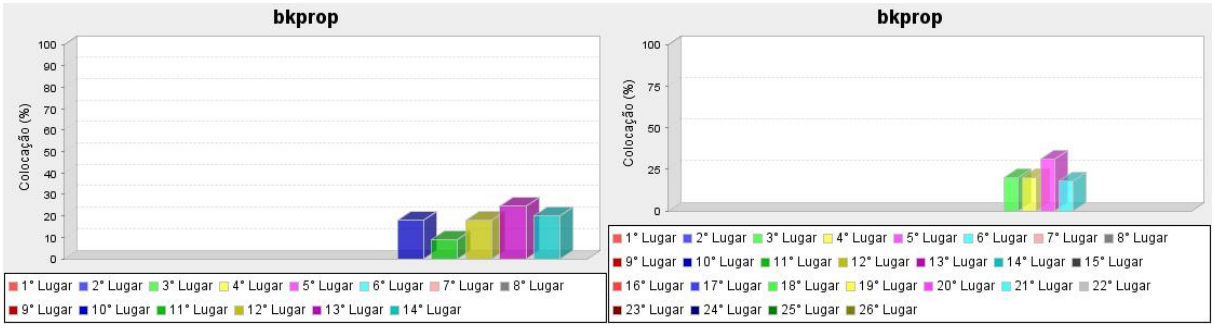
Iris Plants



Pima Indians Diabetes



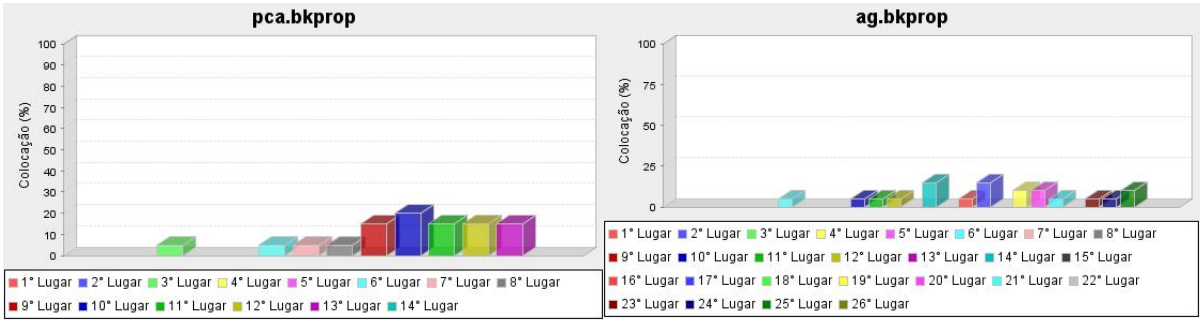
Wisconsin Breast Cancer



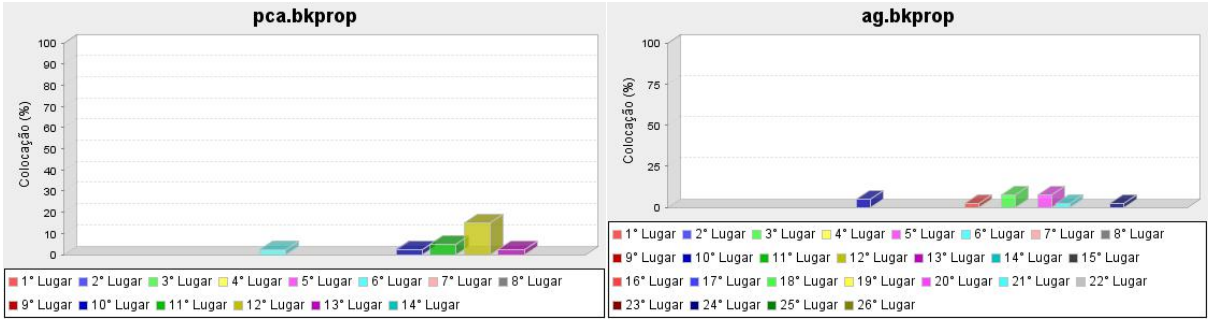
Estratégia 11: Seleção e Imputação com *Back Propagation*

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* foram similares, distantes da primeira colocação no ranking.

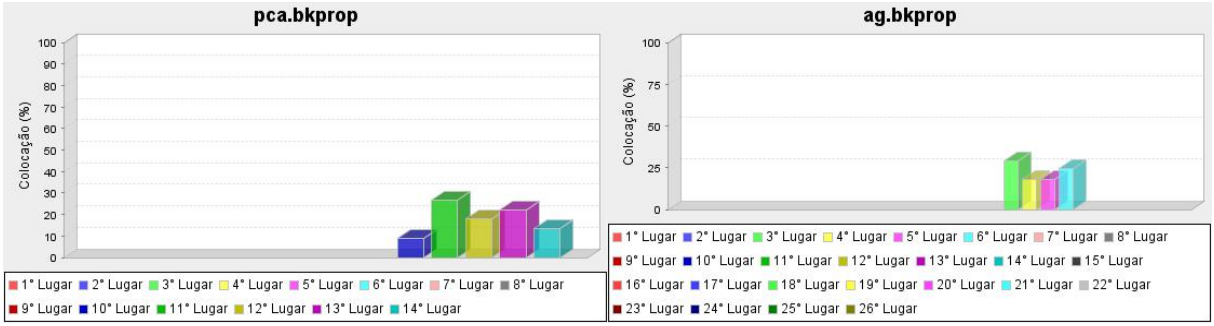
Iris Plants



Pima Indians Diabetes



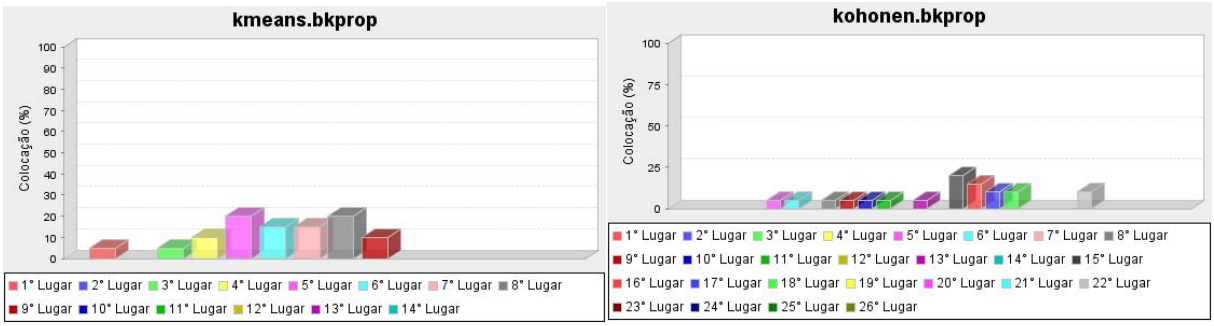
Wisconsin Breast Cancer



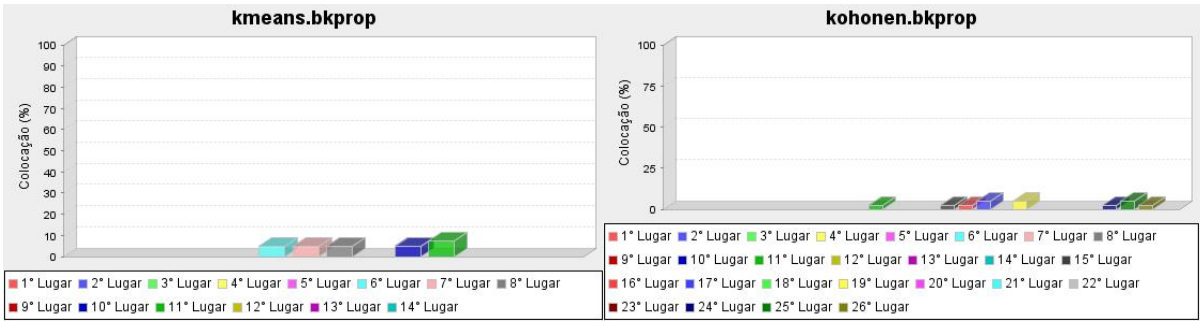
Estratégia 12: Agrupamento e Imputação com *Back Propagation*

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* foram similares, distantes da primeira colocação no ranking, com exceção da base *Iris Plants* onde a *Estratégia de* Soares usando o agrupamento com *K-Means* obteve a primeira colocação num baixo percentual.

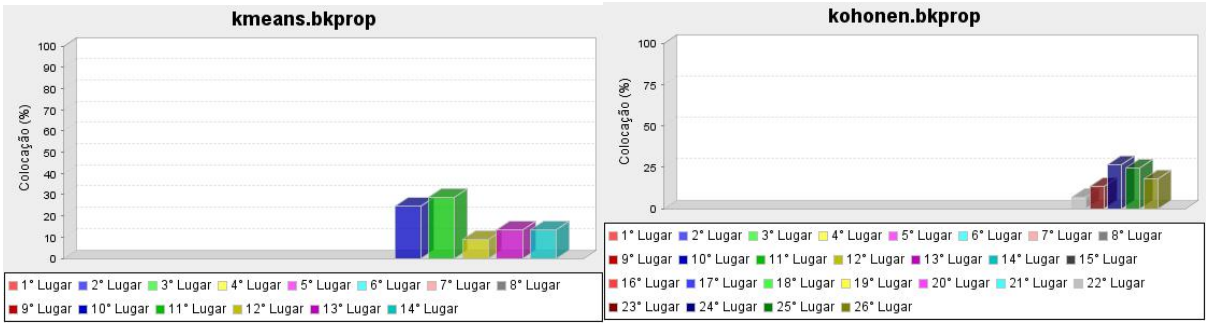
Iris Plants



Pima Indians Diabetes



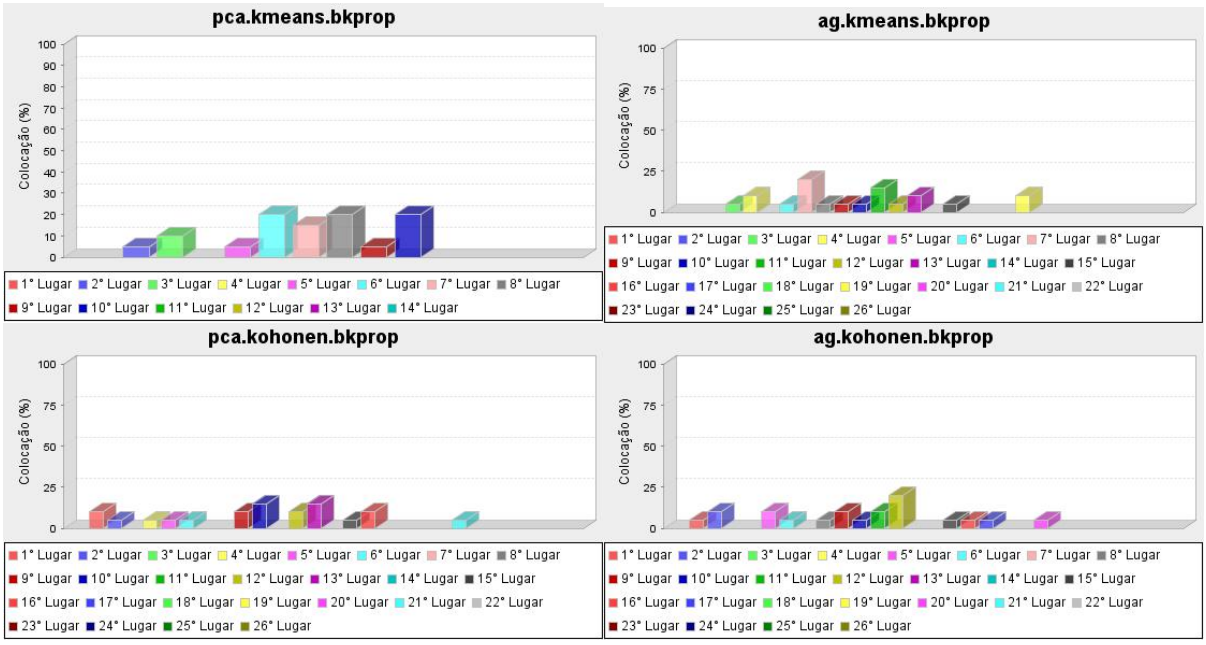
Wisconsin Breast Cancer



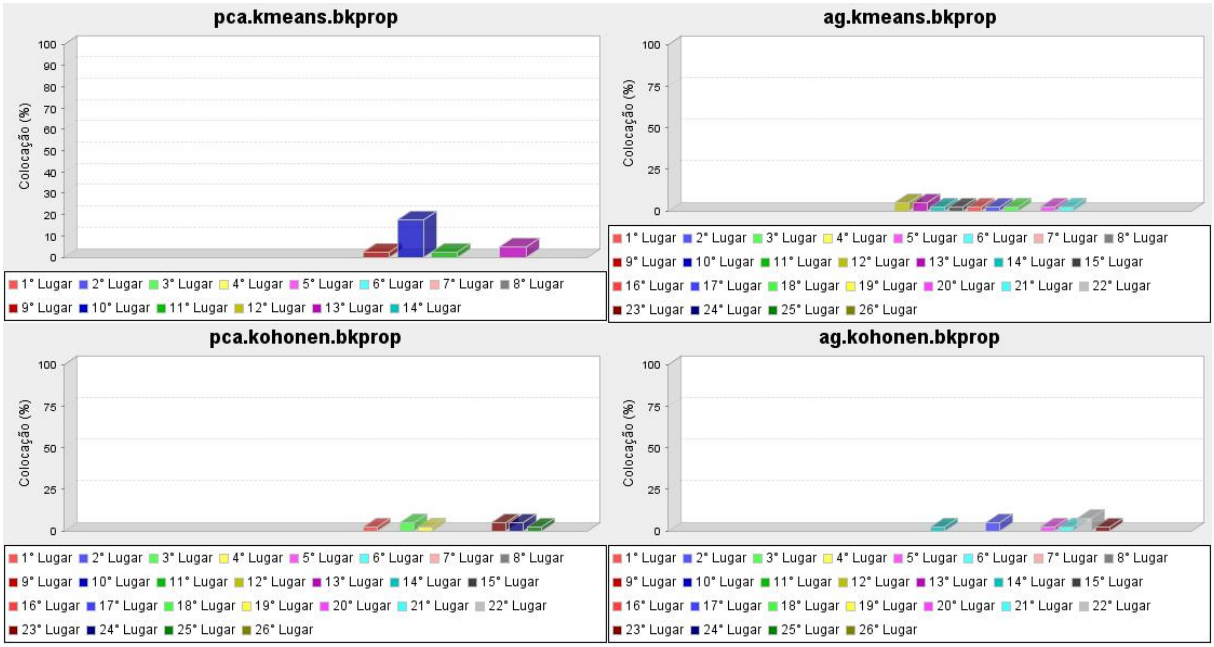
Estratégia 13: Seleção, Agrupamento e Imputação com *Back Propagation*

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* foram similares, distantes da primeira colocação no ranking, com exceção da base *Iris Plants* onde a *Estratégia Proposta* obteve a primeira colocação, mesmo que com um percentual baixo, nas duas combinações envolvendo agrupamento com *Redes de Kohonen*.

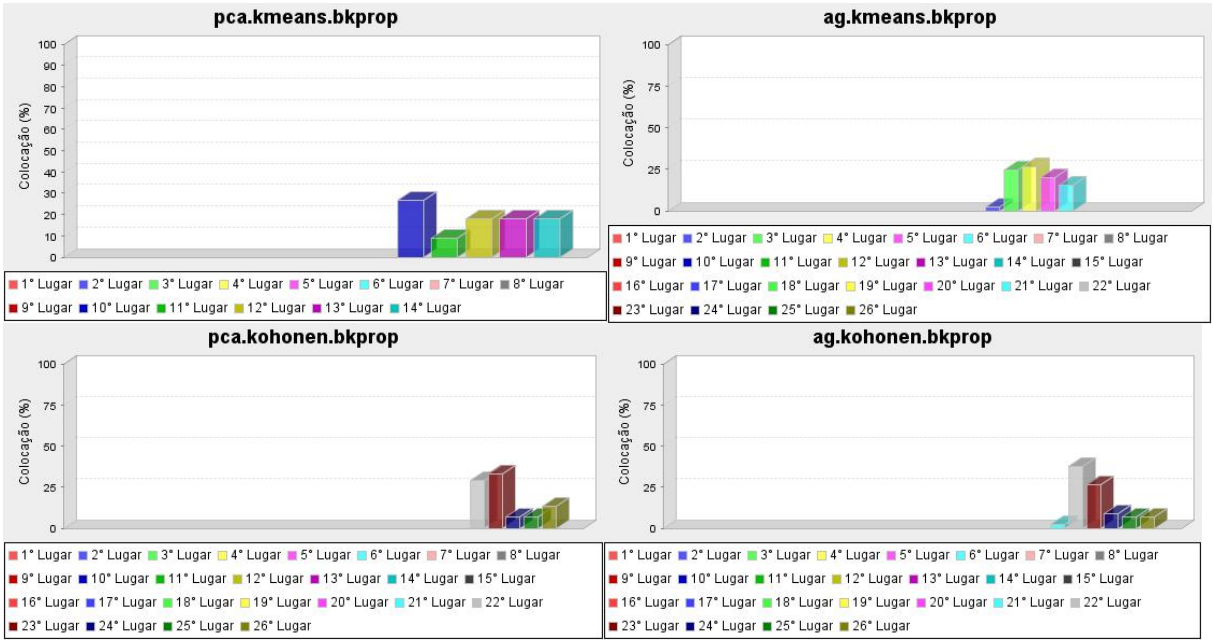
Iris Plants



Pima Indians Diabetes



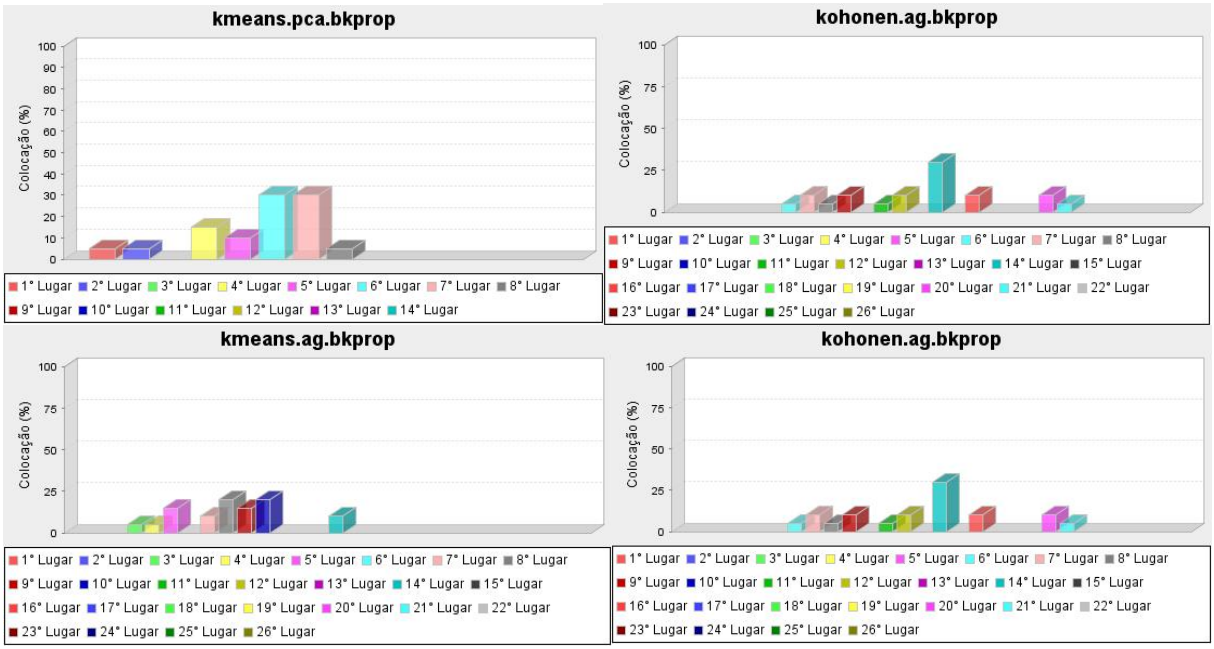
Wisconsin Breast Cancer



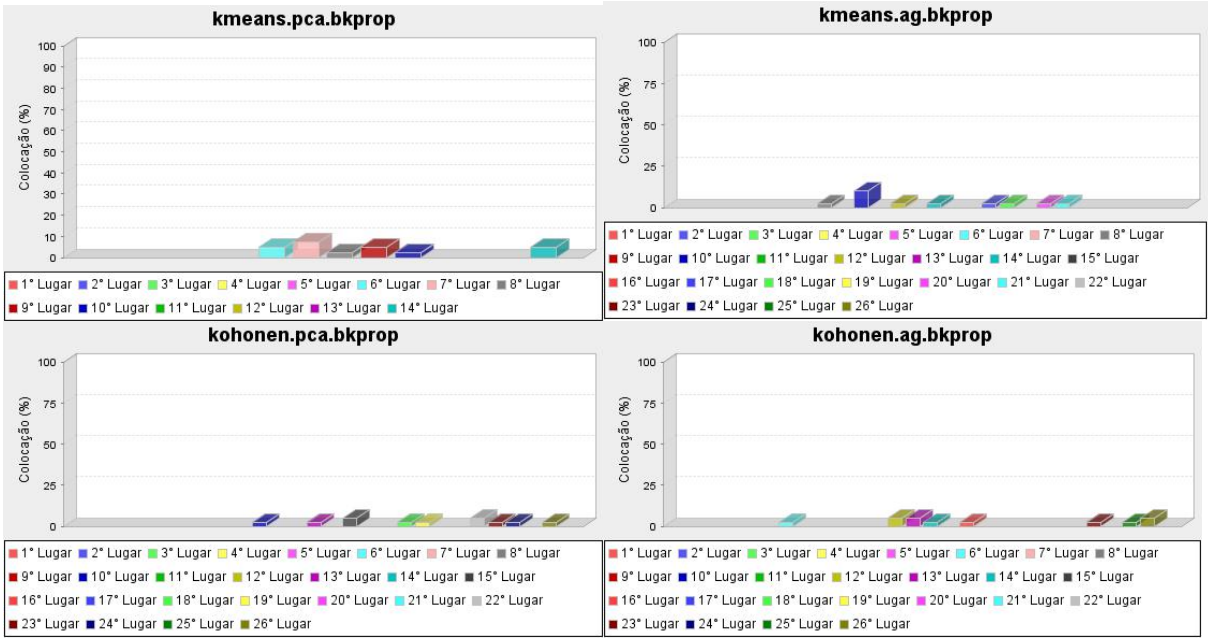
Estratégia 14: Agrupamento, Seleção e Imputação com *Back Propagation*

Os resultados da *Estratégia de Soares* e *Estratégia Proposta* foram similares, distantes da primeira colocação no ranking, com exceção da base *Iris Plants* onde a *Estratégia de* Soares se sobressaiu devido ao agrupamento com *K-Means* seguido de seleção com *PCA* e imputação com *Back Propagation* obtendo a primeira colocação mesmo que com um baixo percentual.

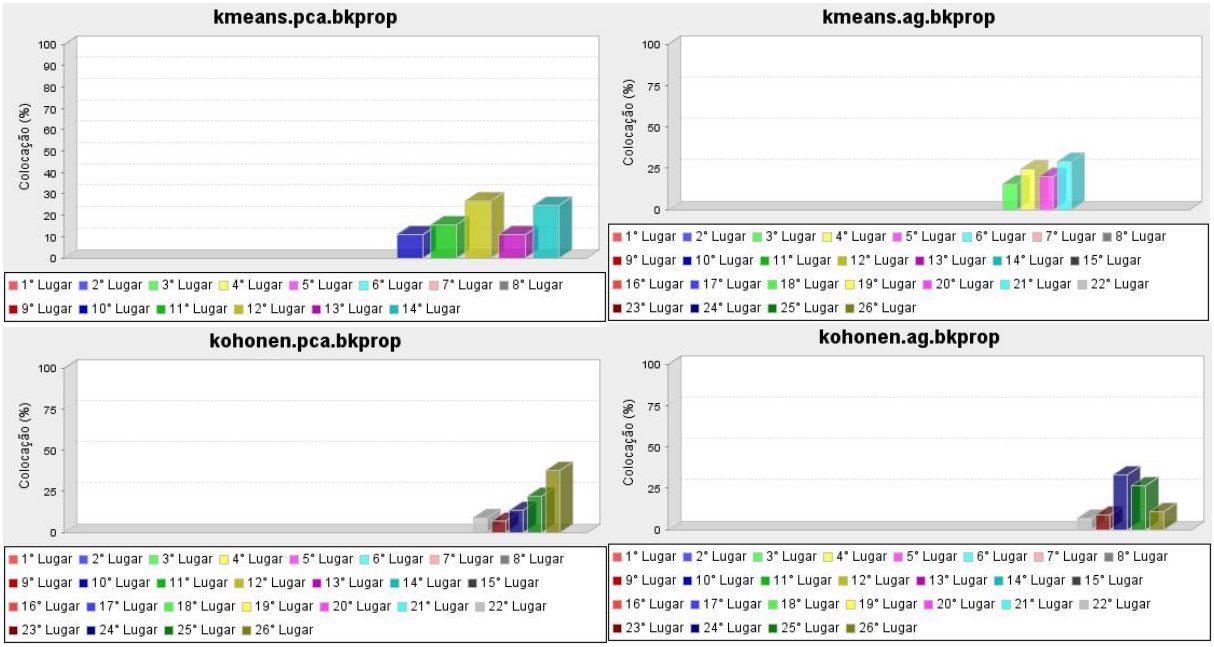
Iris Plants



Pima Indians Diabetes



Wisconsin Breast Cancer



5.4 Resultados Consolidados

A partir das análises dos resultados completos (*Estratégia Proposta unida a Soares*) e dos resultados comparativos (*Estratégia de Soares* e *Estratégia Proposta*) podemos concluir que:

- A imputação simples com *Média* continuou apresentando o pior resultado na base Iris Plants em 100% e a imputação simples com *k-NN* manteve sua colocação próxima aos primeiros lugares do ranking.

- A seleção com *AG* assim como com *PCA* seguidas de imputação não deram resultados satisfatórios analisando o ranking por estratégia.

- O agrupamento com *Redes de Kohonen* seguido de imputação com *Média* e *k-NN* (*hot-deck*) apresentou bons e ótimos resultados aparecendo em 1º lugar, entretanto não com um percentual de colocação maior do que *K-Means*. Na base *Iris Plants* a diferença foi de 65%, mas na base *Pima Indians Diabetes* e *Wisconsin Breast Cancer* a diferença foi de apenas 10% e 20%. Dessa forma a pontuação geral do agrupamento aumentou na base *Pima Indians Diabetes* diminuindo a qualidade de todas as outras combinações de técnicas.

- O agrupamento com *Redes de Kohonen* seguido de imputação também tornou os resultados mais constantes em duas bases (Iris e Pima) ao avaliar a imputação nos cinco percentuais de ausência (10%, 20%, 30%, 40% e 50%) e manteve a constância já existente na base *Wisconsin Breast Cancer*.

- A seleção seguida de agrupamento obteve melhores resultados nas combinações envolvendo seleção com *AG* e imputação com *k-NN* mas não foram suficientes para resultar numa melhora dessa combinação estratégica no resultado geral.

- O agrupamento seguido de seleção resultou grande melhora com o uso de *K-Means* e *AG*, alcançando até 50% na 1ª colocação na base *Iris Plants*. Mas no resultado final essa combinação fez diferença apenas na base *Wisconsin Breast Cancer* ficando no lugar da imputação simples (regressão).

CAPÍTULO 6
CONSIDERAÇÕES FINAIS

Este capítulo traça as considerações finais obtidas ao decorrer dos capítulos e projeta trabalhos que podem ser realizados futuramente de forma complementar.

6.1 Resumo

Inicialmente vimos a importância da complementação de dados ausentes na descoberta de conhecimento, pois que os valores ausentes podem prejudicar substancialmente as informações descobertas e consequentemente as decisões baseadas nessas informações.

Complementando o trabalho de SOARES (2007), ao utilizar novas técnicas de seleção e agrupamento baseadas em CONDE (2005) e FERLIN (2008) respectivamente, apresentamos mais opções de melhores práticas no processo de preparação de dados que levassem a analise das relações intrínsecas entre os dados da base de dados em consideração, para concluir se a técnica terá bons resultados também para outras bases de dados com similaridade nestas relações dos dados.

SOARES (2007) desenvolveu um sistema chamado *Appraisal* que contém os módulos para gerar bases de dados ausentes, processar as estratégias de complementação de dados e gerar os gráficos dos resultados processados pelas estratégias. Foram 14 estratégias utilizadas, que são combinações de três técnicas de imputação (*Média, k-NN e Back Propagation*), uma técnica de seleção (*PCA*) e uma técnica de agrupamento (*k-Means*). Neste trabalho estendemos o estudo, reexecutando os experimentos originais de SOARES (2007), e gerando novos resultados a partir da aplicação de uma nova técnica de seleção (*AG*) e uma de agrupamento (redes de *Kohonen*) gerando 23 novas possibilidades.

Utilizamos as mesmas bases de dados usadas por SOARES (2007) *(Iris Plants*, *Pima Indians Diabetes* e *Wisconsin Breast Cancer)*, derivando novos arquivos em função de um determinada frequência de ausência, segundo o padrão completamente aleatório (MCAR). Adotamos as configurações que SOARES (2007), CONDE (2005) e FERLIN (2008)

adotaram em seus testes buscando também optar pelas configurações que obtiveram melhor resultado final.

Analisamos os resultados através dos gráficos gerados pelo módulo *Reviewer* do sistema *Appraisal*, concluímos que o agrupamento com *Redes de Kohonen* melhorou a estratégia de agrupamento seguida de imputação aumentando tornando estra estratégia mais constante com relação ao percentual de ausência. Também vimos assim como SOARES (2007), que a imputação com k-NN apresentou o melhor resultado em todas as bases de dados e concordamos com SOARES (2007) que muito provavelmente foi por causa do princípio usado por esse algoritmo que só usa para imputação os dados dos atributos com maior grau de semelhança. Continuando as comparações semelhantes, a imputação simples com *Média* gerou pior resultados principalmente na base *Iris Plant*, com 100% de colocação no último lugar.

A seleção com *AG* seguida de imputação assim como a seleção com *PCA* não forneceu resultados satisfatórios. Já o agrupamento seguido de seleção resultou grande melhora com o uso de *K-Means* e *AG*, mas no resultado geral essa combinação ficou no lugar da imputação simples (regressão) na base *Wisconsin Breast Cancer*.

6.2 Contribuições

Descrevemos abaixo algumas contribuições deste trabalho:

1. Desenvolvimento de uma imputação composta utilizando combinações técnicas agrupamento e seleção de dados;
2. Implementação de Algoritmos Genéticos para seleção de dados na complementação de dados ausentes;
3. Implementação de Redes Neuronais Kohonen para agrupamento de dados na complementação de dados ausentes;
4. Análise de resultados que indicam melhora na qualidade da imputação ao utilizar combinações de técnicas, principalmente pelo uso da técnica de agrupamento;
5. Demonstração da possibilidade de utilização do sistema *Appraisal* para desenvolver e executar outras técnicas e suas combinações;
6. Comprovação de que os diferentes percentuais de ausência não interferem na qualidade dos resultados.

6.3 Trabalhos futuros

Diversos trabalhos podem ser realizados baseados nesse a fim de complementar ou aprimorar as técnicas implementadas:

1. Utilização de outras técnicas de seleção e/ou agrupamento e/ou imputação para comparação dos resultados, assim como este trabalho fez com SOARES (2007);
2. Escolha de outra técnica de AG ou combinações de técnicas, já que aqui foi implementada apenas a técnica padrão;
3. Aumento no percentual de ausência para avaliar se o resultado da imputação composta se mantém constante;
4. Geração de ausência em mais de um atributo da mesma tupla para aplicar a imputação multivariada;
5. Adoção de outro padrão de ausência que não seja MCAR.
6. Desenvolvimento de imputação composta repetida, ou seja, realizar duas ou mais vezes a seleção dos dados ou duas ou mais vezes o agrupamento.
7. Assim como SOARES (2007) sugeriu, seria interessante reestruturar o que implementamos para processar atributos categóricos para analisar se os resultados variam com a natureza dos atributos ou usar as mesmas bases de dados retirando um atributo que tenha menor correlação com os demais para avaliar os impactos;
8. E por último mas não menos importante, sugerimos assim como SOARES (2007) o processamento em bases de dados com mais registros e/ou com mais atributos das que foram aqui utilizadas.

REFERÊNCIAS

AHA, D. W., KIBLER, D., ALBERT, M.. Instance-based Learning Algorithms, Machine Learning. 1991. v.6, p. 37-66.

ANDERSON, R. L.. Missing plot techniques. Biometrics, 1946.

BATISTA, G. E. A. P. A., MONARD, M. C.. Um Estudo Sobre a Efetividade do Método de Imputação Baseado no Algoritmo k-Vizinhos Mais Próximos. In: *IV Workshop on Advances & Trends in AI Problem Solving*, Chilán: Chile, 2003. p. 1–6.

BRÁS, L. P., MENEZES, J. C.. Improving cluster-based missing value estimation of dna microarray data. Biomolecular Engineering. 2007.

CARVALHO, L. A. V., FISZMAN, A., FERREIRA, N. C.. A Theoretical Model for Autism. Journal of Theoretical Medicine. Estados Unidos, 2001. v.25, p. 123-140.

COSTA, Bruno Conde de Miranda. Seleção de Variáveis na Descoberta de Conhecimentos em Base de Dados. Rio de Janeiro, 2005. 137p. Dissertação (Projeto final de bacharelado) – Escola de Ciências Exatas e Tecnologia, Universidade da Cidade do Rio de Janeiro.

DASARATHY, B.. Nearest Neighbor (NN) norms: NN pattern classification Techniques. 1. ed. IEEE Computer Society Press: Los Alamitos. 1990.

DEMPSTER, A. P., LAIRD, N. M., e RUBIN, D. B.. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. 1977.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P.. From Data Mining to Knowledge Discovery in Databases. AI Magazine, American Association for Artificial Intelligence. 1996. p. 37-54.

FERLIN, Claudia. Imputação Multivariada: Uma abordagem em cascata. Rio de Janeiro: RJ, 2008. 256p. Dissertação (Tese de Pós-doutorado) - COPPE, Universidade federal do Rio de Janeiro.

FREITAS, Alex. Data Mining and knowledge Discovery with Evolutionary Algorithms. Nova York: Springer, 2002. 265p.

GOLDSCHMIDT, R. e PASSOS, E., Data Mining: Um Guia Prático - Conceitos, Técnicas, Ferramentas, Orientações e Aplicações. Rio de Janeiro: RJ, 2005. v.1.

HAYKIN, S.. Neural Networks: A Comprehensive Foundation. Macmillan College Publishing Company. Estados Unidos: Nova York, 1994.

HOLLAND, J. H. Adaptation in Natural and Artificial Systems An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. Bradford Books. 1995.

HRUSCHKA, E. R., JUNIOR, E. R. H., EBECKEN, N. F. F.. Missing values prediction with K2, Intelligent Data Analysis Jounal. 2002b. v.6, pp. 557-566.

HRUSCHKA, E. R., JUNIOR, E. R. H., EBECKEN, N. F. F.. A Nearest-Neighbor Method as a Data Preparation Tool for a Clustering Genetic Algorithm. In: Anais do 18º Simpósio Brasileiro de Bando de Dados (SBBD). Amazonas: Manaus, 2003a. 6.v, pp. 319-327.

HRUSCHKA, E. R., JUNIOR, E. R. H., EBECKEN, N. F. F.. A Nearest-Neighbor Method Method to Substitute Continuous Missing Values. In: The 16th Australian Joint Conference on Artificial Inteligence. Spring-Verlag: Heidelberg, 2003b. v.2903, pp. 723-734.

KIM, K. Y., LEE, B. J., YI, G S.. Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinformatics. 2004.

KOHONEN, T.. “Self-organization and Associative Memory”. Berlin: Springer-Verlag, 1984.

LAKSHMINARAYAN, K., HARP, S. A., SAMAD, T.. Imputation of Missing Data in Industrial Databases. 1999. v.3, v.11, pp. 259-275.

LINDEN, Ricardo. Algoritmos Genéticos, Uma importante ferramenta da inteligência computacional. Rio de Janeiro: RJ, 2006. 348p.

LITTLE, R. e RUBIN, D.. Statistical Analysis With Missing Data. *Technometrics*, 2003. pp. 346-365.

MAGNANI, M.. Techniques for Dealing with Missing Data in Knowledge Discovery Tasks. Disponível na Internet. http://pdf.aminer.org/000/371/480/techniques_for_dealing_with_missing_values_in_classification.pdf, 20 de outubro de 2012. 10p.

MCCULLOCH, W. e PITTS, W.. Activity Bulletin of Mathematical Biology, A Logical Calculus of the Ideas Immanent in Nervous. 1943. 133p.

MCQUEEN, J.. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967. p. 281–297.

MYRTVEIT, I., STENSRUD, E., OLSSON, U. H.. Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods. *IEEE Transactions on Software Engineering*. v.27, n.11, Nov. 2001.

MITCHELL, T. M.. Machine Learning. Ed. McGraw-Hill. 1997.

MONTEIRO, Diego Saraiva. Imputação Composta Categórica em Bases de Dados. 2008. Rio de Janeiro: RJ, 2008. Trabalho de Conclusão de Curso (Graduação em Bacharelado em Ciência da Computação) - Centro Universitário da Cidade.

NEWMAN, D. J., HETTICH, S., BLAKE, C. L., MERZ, C. J.. *UCI Repository of Machine Learning Databases*. Disponível na internet. http://www.ics.uci.edu/~mlearn/MLRepository.html, 12 de outubro de 2006.

OGLIARI, Paulo José, ANDRADE, Dalton Francisco, BARBETTA, Pedro Alberto, PACHECO, Juliano Anderson. Modelos de Regressão: Desenvolvimento Teórico e Aplicações. 2003. 9p.

PREECE, A. D.. Interatives procedures for missing values in experiments. Technometrics, 1971.

PUC, RIO. Certificado Digital No 0310411/CA. Disponível na Internet. http://www.maxwell.lambda.ele.puc-rio.br/7070/7070_3.PDF, 21 de julho de 2012.

QUINLAN, J. Ross. Induction of decision trees. Machine Learning. 1993.

RIBEIRO, Lívia de Souza. Utilizando proveniência para complementação de dados no contexto do processo de ETL. Rio de Janeiro, 2010. 106p. Dissertação (Mestrado em Sistemas e Computação) - Departamento de Ciência e Tecnologia, Instituto Militar de Engenharia.

RIBEIRO, Rafael Castaneda. Um Ambiente de Imputação Sequencial para Cenários Multivariados. Rio de Janeiro, 2008. 79p. Dissertação (Mestrado em Sistemas e Computação) - Departamento de Ciência e Tecnologia, Instituto Militar de Engenharia.

RUBIN, D. B.. Formalizing subjective notion about the effects of nonrespondents in samples surveys. Jounal of the American Statistical Association. 1977.

RUBIN, D. B.. Multiple imputation for non responses in surveys. Estados Unidos: New York:, 1987.

RUBIN, D. B.. An Overview of Multiple Imputation, In: Proceedings of the Section on Survey Research Methods. American Statistical Association. 1988. pp. 79-84.

RUMELHART, D.E., HINTON, G.E., WILLIAMS. Learning Internal Representations by Error Back-propagation. Rio de Janeiro, 1986. v.1.

SCHAFER, J.. Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC, 2000.

SCHAFER, J.., GRAHAM, J. W.. Missing Data: Our View of the State of the Art. 2002. V.7, n.2, pp.147-177..

SCHÖNER, H.. Working with Real-World Datasets. Technische Universität Berlin, 2004.

SHLENS, J.. A Tutorial on Principal Component Analysis. 2005. Disponível na Internet. http://www.cs.cmu.edu/~elaw/papers/pca.pdf, em 30 de novembro de 2011.

SILVA, Jonathan. Substituição de valores ausentes: uma abordagem baseada em um algoritmo evolutivo para agrupamento de dados. São Paulo: São Carlos, 2010. 120p. Dissertação (Tese de Mestrado) - USP. Disponível na Internet. http://www.teses.usp.br/teses/disponiveis/55/55134/tde-07062010-144250/pt-br.php, em 1 de agosto de 2012.

SOARES, Jorge de Abreu. Pré-processamento em Mineração de Dados: Um Estudo Comparativo em Complementação. 2007. 240p. Dissertação (Tese de Pós-doutorado) - COPPE, Universidade federal do Rio de Janeiro.

SONG, Q., SHEPPERD, M., CARTAWRIGHT, M.. A Short Note on Safest Default Missingness Mechanism Assumptions". *Empirical Software Engineering,* v 10, n2, pp.235-243, Apr. 2005.

SMITH, L. I.. A Tutorial on Principal Component Analysis. 2002. Disponível na Internet. http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf, em 14 de abril de 2007.

TWALA, B., CARTWRIGHT, M., SHEPPERD, M.. Comparison of Various Methods for Handling Incomplete Data in Software Engineering Databases. In: *2005 International Symposium on Empirical Software Engineering*, pp. 105-114, Nov. 2005.

TSENG, S. WANG, K., LEE, C.. A preprocessing Method to Deal with Missing Values by Integrating Clustering and Regression Techniques. *Applied Artificial* Intelegence. v. 17, n. 5, May-Jun, pp. 535-544. 2003.

WAYMAN, J. C.. Multiple For Missing Data: What is It And How Can I Use it?. In: Proceedings of the Annual Meeting of the American Educational Research Association. Estados Unidos: Chicago, 2005.

WITTEN, I. H., FRANK, E.. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. 2 ed. San Francisco: CA: Morgan Kaufmann Publishers. 2005.