# Aplicação de Algoritmos de Árvore de Decisão Sobre Uma Base de Dados de Câncer de Mama

**Jéssica Augusti Bonini**

Bacharelado em Ciência da Computação – Centro de Tecnologia, Universidade Federal de Santa Maria (UFSM)
Avenida Roraima – Santa Maria – RS – Brasil

jbonini@inf.ufsm.br

***Abstract.*** *This article describes a data mining experiment using decision trees for the extraction of information from a dataset with breast tumor samples and their characteristics, with the purpose of the classification of the tumor benign or malignant. Detailed up the database and the strategies used to carry out the experiments and application of decision algorithms. The decision tree algorithms selected for the classification of the samples in question were the J48 and the REPtree. After analysis of the trees generated, it is concluded that the quantity attribute in the cytoplasm nucleus was always analyzed to determine both malignant tumors and benign tumors. In addition, there was the standard confirmations for breast tumors.*

***Resumo.*** *Este artigo descreve um experimento de mineração de dados utilizando árvores de decisão para a extração de informações de um dataset com amostras de tumores de mama e suas características, tendo como finalidade a classificação do tumor em benigno ou maligno. Detalhou-se a base de dados e as estratégias utilizadas para realização dos experimentos e aplicação dos algoritmos de decisão. Os algoritmos de árvore de decisão escolhidos para a classificação das amostras em questão foram o J48 e o REPtree. Após analise das árvores geradas, conclui-se que o atributo quantidade de citoplasma no núcleo sempre foi analisado para a determinação tanto de tumores malignos quanto de tumores benignos. Além disso, houve confirmações de padrões para tumores de mama.*

## 1. Introdução

Este artigo refere-se à aplicação de algoritmos de árvores de decisão com o intuito de obter conhecimentos sobre as características de tumores de mama, classificando-os em malignos ou benignos. A classificação é feita analisando-se outros 9 atributos (espessura, tamanho, forma, adesão, células epiteliais, núcleos nus, cromatina, núcleos normais e mitoses).

O experimento busca aplicar conhecimentos da área de informática para resolver problemas na área da saúde, como no caso, auxiliar na descoberta de padrões e informações úteis para o diagnóstico de tumores de mama.

A mineração de dados vem mostrando-se bastante eficiente para a descoberta de conhecimentos em grandes bases de dados, assim, para o experimento foram aplicados algoritmos de mineração de dados, presente no software WEKA, sobre um dataset com 699 amostras de tumores na região da mama. Esse conjunto de dados foi retirado da

Wisconsin Breast Cancer e foram recolhidas pelo Dr. William H. Wolberg, da Universidade de Wisconsin Hospitals.

Os algoritmos de árvore de decisão utilizados para a extração de informações contidas no conjunto de dados foram J48 e REPtree, e a aplicação dos mesmos ao dataset foi feita utilizando o software WEKA.

Além de alguns conceitos breves e alguns testes realizados, são apresentadas as árvores geradas pelos dois algoritmos e os conhecimentos adquiridos a partir da analise das mesmas.

As seções a seguir apresentam a fundamentação teórica, descrição da base de dados escolhida, descrição dos experimentos realizados, analise das árvores geradas, resultados obtidos e conclusão, correspondendo à seção 2, seção 3, seção 4, seção 5, seção 6 e seção 7, respectivamente.

## 2. Fundamentação teórica

A mineração de dados é a exploração e análise de grandes quantidades de dados, a fim de descobrir padrões e regras significativas [Berry e Linoff, 2004]. Essa técnica tem sido muito importante na descoberta de conhecimento implícito em grandes bases de dados, auxiliando na tomada de decisões a partir dos conhecimentos extraídos. Como cita Amorin (2006) a maioria das organizações não usa adequadamente o grande volume de dados obtidos, assim deixa de utilizar o conhecimento contido nessas bases dados. Amorin (2006) fala ainda da importância da mineração de dados na geração de informações e conhecimentos escondidos sob essa grande massa de dados e que não poderiam ser descobertos através do gerenciamento convencional de banco de dados. Ainda, Berry e Linoff (2004) citam que muitas vezes o processo de mineração de dados é referido como descoberta de conhecimento ou KDD (descoberta de conhecimento em bases de dados), porém eles acreditam que a mineração de dados é um processo de criação de conhecimento.

O principal desafio é lidar com grandes volumes de dados e informações, segundo Halmenschlager (2002) as árvores de decisão são facilmente aplicadas a grandes conjuntos de dados e são adequadas tanto para dados contínuos quanto para dados discretos.

A classificação de dados é uma das tarefas mais comuns de mineração de dados e consiste em examinar as características de um objeto recém-apresentado e atribuí-lo a uma classe pré-definida [Berry e Linnof, 2004]. Pode ser feita através de técnicas de aprendizagem supervisionadas e não supervisionadas. As supervisionadas são aquelas onde existe um atributo que especifica a qual classe cada instância pertence. Já nas não supervisionadas as instâncias são utilizadas sem a determinação de um atributo chave [Damasceno]. As árvores de decisão utilizam uma variável que apresenta informações sobre as classes a que pertence cada amostra do conjunto de dados, ou seja, essas são técnicas de reconhecimento de padrões que utilizam a aprendizagem supervisionada [Rodrigues, 2005]. Berry e Linoff (2004) citam que as árvores de decisão são técnicas supervisionadas muito uteis para a construção de perfis, como por exemplo, o comportamento de um cliente.

Árvores de decisão constituem um método de classificação baseado na analise dos valores de atributos com finalidade de representar de forma simples e eficiente o conhecimento obtido a partir de um conjunto de amostras de entrada. Normalmente são

usadas em práticas de mineração de dados, ou seja, atividades de extração de conhecimentos, previamente desconhecidos, de uma base de dados [Caraciolo, 2009].

Em uma árvore de decisão, os nós internos correspondem aos testes realizados sob os valores dos atributos, nos arcos são encontrados os possíveis valores de saída para determinado teste e as folhas mostram os resultados da classificação das amostras.

O algoritmo J48 gera árvores de decisão escolhendo o atributo mais apropriado para cada teste ou situação. As árvores são geradas do topo para base e cada nó analisa a relevância do atributo de forma individual. Esse algoritmo não é capaz de realizar regressão, sendo assim, o atributo classe deve ser discreto. Ainda, utiliza a estratégia "gulosa" para induzir árvores de decisão que serão classificadas posteriormente [Martins, 2009]. Um algoritmo guloso é aquele que escolhe a cada iteração o objeto mais apetitoso [Feofiloff, 2015], ou seja, escolhe a melhor alternativa local esperando que isso leve a uma solução ótima global [Koerich, 2006]. No algoritmo J48 a existência ou significância de cada atributo é avaliada de forma individual em cada nó da árvore [Martins, 2009].

O algoritmo REPtree usa informações de ganho e variância para construir a árvore de decisão. Além disso, usa a técnica de poda por redução de erro para podar os ramos da árvore [KNIME WEKA]. A poda por redução de erro separa as amostras em dois conjuntos, um de treinamento e outro de validação. Os dados de treinamento são usados para construir a árvore de decisão e os de validação são utilizados para verificar os erros de classificação [Paulo, 2012]. Possui também, uma variável que define o número máximo de profundidade da árvore.

A matriz de confusão é a matriz gerada a partir da classificação do conjunto de dados em categorias. Nela são comparados os valores reais obtidos com os valores previstos para cada amostra. Os dados do conjunto são classificados em verdadeiros positivos e negativos, que correspondem às amostras cujos valores reais e previstos são iguais, e falsos positivos e negativos, correspondentes às amostras que possuem valores reais diferentes dos valores previstos [Developer Network].

## 3. Descrição da base de dados escolhida

A base de dados escolhida, contém amostras de cancêr de mama recolhidas periodicamente pelo Dr. William H. Wolberg, da Universidade de Wisconsin Hospitals. As amostras foram coletadas usando o método de punção aspirativa por agulha fina, método que permite a retirada de células de nódulos ou lesões em diversos órgãos e tecidos superficiais ou profundos, através de uma agulha fina, não sendo necessária anestesia ou qualquer preparo prévio na maioria dos casos [], e refletem características dos núcleos celulares.

O dataset possui 699 instâncias, dessas 458 pertencem a classe benigno e 241 a classe maligno. Ainda, existem 11 atributos, com valores no intervalo de 1 a 10, onde 1 representa um estado normal e 10 um estado completamente anormal. Todos os atributos do dataset são originalmente numéricos.

Dos 11 atributos existentes no dataset, apenas 10 são relevantes para o experimento, ou seja, o atributo id não possui valor significativo na classificação das amostras. Os atributos diagnóstico, espessura, tamanho, forma, adesão, células epiteliais, núcleos nus, cromatina, nucléolo normal e mitoses trazem informações importantes para a classificação dos tumores.

O atributo **id**, como citado anteriormente, não possui importância sobre a classificação das instâncias, pois apresenta somente o número da amostra.

As classes possíveis para classificação dos tumores são dadas através do atributo **diagnóstico**, sendo que o valor 2 representa tumores benignos e 4 tumores malignos.

Quanto ao atributo de **espessura**, células normais crescem formando apenas uma camada, ou seja, tendem a agrupar-se formando uma monocamada, já em células tumorais o crescimento é contínuo e forma pilhas de células sobrepostas, formando multicamadas [Alessi, 2011]. O valores para esse atributo são 1 para células totalmente monocamadas, 2 - 5 para 90-50% das células monocamadas, 5 - 9 para 50-20% das células monocamadas e 10 para células totalmente multicamadas.

No caso de células malignas, essas normalmente sofrem variações de tamanho e forma, assim os atributos correspondentes a essas duas características possuem uma significância alta [Alessi, 2011]. Para o atributo de uniformidade de **tamanho** os valores são 1 corresponde a células completamente uniformes, 2 - 9 a 90-20% das células uniformes e 10 a células não uniformes. Já para o atributo de uniformidade de **forma** o valor 1 representa células completamente uniformes, 2 - 9 representa 90-20% das células uniformes e 10 células não uniformes.

O atributo **adesão** representa uma das características mais importantes das neoplasias malignas, pois a perda de adesão entre as células vizinhas permite o desprendimento e disseminação das células [Alessi, 2011]. Dessa forma, um tumor maligno terá células com menor adesão entre si. Assim, valores para esse atributo correspondem a 1 para células completamente grudadas, 2 - 9 para 90-20% das células grudadas e 10 para células completamente desgrudadas.

Para o atributo **células epiteliais**, leva-se em conta que células cuja dilatação é significativa podem indicar tumores malignos. Sendo assim, 1 representa células com tamanho normal, 2 - 9 representa 20-90% das células maiores e 10 representa 100% das células maiores.

**Núcleos nus** correspondem a relação núcleo/citoplasma. Se essa relação está aumentada existem grandes chances do tumor ser considerado maligno [Alessi, 2011]. Os valores são 1 para núcleos desprovidos de citoplasma, 2 - 9 para 20-90% dos núcleos têm citoplasma e 10 para todos os núcleos com citoplasma.

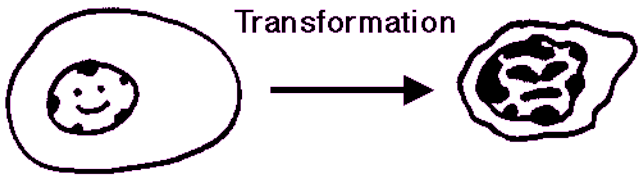


**Figura 1. Transformação de núcleo normal para núcleo anormal.**

No caso do atributo **cromatina**, células benignas costumar possuir uma cromatina bem formada, delicada, já células malignas possuem cromatina grosseira, mal formada [Alessi, 2011]. Nesse caso os valores representam 1 para cromatina completamente bem formada, 2 - 9 para 20- 90% mal formada e 10 para cromatina completamente deformada.

Em células normais os nucléolos são estruturas pequenas e quase invisíveis localizadas no núcleo, nas células cancerosas os nucléolos tornam-se evidentes e múltiplos [Alessi, 2011]. Para o atributo **nucléolos normais** 1 representa nucléolos completamente normais, 2 - 9 representa 20-90% dos nucléolos são anormais e 10 nucléolos completamente anormais.

As mitoses em células normais são raras e típicas, já em em células malignas tornam-se frequentes e atípicas [Alessi, 2011]. O atributo **mitoses** possui os seguintes valores, 1 para atividade mitótica completamente normal, 2 - 9 para 20- 90% da atividade mitótica anormal e 10 para atividade mitótica completamente anormal.

## 4. Descrição dos experimentos realizados

O primeiro passo para a realização dos experimentos foi a aplicação de estratégias de pré-processamento. Técnicas de pré-processamento são utilizadas para aumentar a qualidade e o poder de expressão dos dados que posteriormente serão minerados. As principais tarefas da fase de pré-processamento são a limpeza, integração e transformação dos dados.

No caso do dataset escolhido, depois de analisar os atributos foi decidido pela exclusão dos atributos ID e Group, pois estes eram irrelevantes para a classificação do tumor. Foi necessário também, mudar o atributo diagnóstico de numérico para nominal, utilizando o filtro não supervisionado "NumericToNominal" para a aplicação do algoritmo J48.

Atributos nominais são aqueles que possuem valores que são apenas nomes diferentes, ou seja, apenas fornecem informação para diferenciar uma instância da outra. Valores nominais não possuem uma ordem e somente testes de igualdade e diferença podem ser aplicados sobre eles.

Além da transformação do atributo diagnóstico, citada anteriormente, foi testada a possibilidade de aplicação dos algoritmos de árvore de decisão com todos os atributos transformados de numérico para nominal, usando novamente o filtro "NumericToNominal". A árvore obtida não foi satisfatória, pois, foram geradas arestas para todos os valores de cada atributo (valores de 1 a 10) e ainda número de atributos analisados para a obtenção do diagnóstico final, benigno ou maligno, foi pequeno, levando a resultados bastante genéricos. Desta forma, ficou decidido pelo uso apenas do atributo diagnóstico como nominal, os outros atributos permaneceram numéricos.

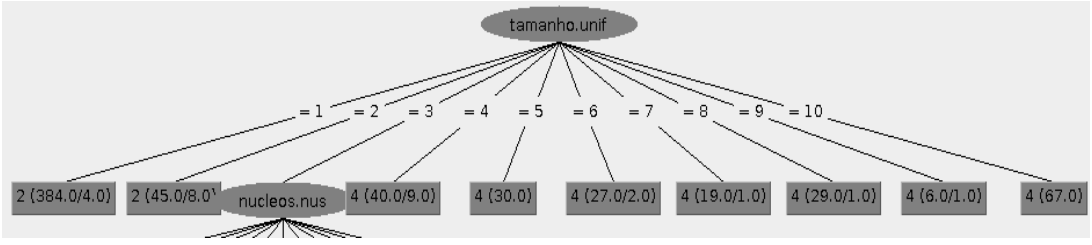


**Figura 2. Imagem de um nó da árvore gerada com todos os atributos nominais.**

O termo confiança é utilizado para avaliar a força e a validade de uma regra. Considere os itens A e B, a taxa de confiança é dada pela razão entre o número de registros contendo A e B e o número de registros contendo A. A confiança junto com o

suporte, razão entre o número de registros contendo A e B e o número total de registros, permite eliminar regras que possuem pouca significância. Durante a realização do experimento os valores do parâmetro de confiança foram alterados, com valores de confiança altos, como 0.5, por exemplo, houve repetição de atributos nos nós o que prejudica a classificação correta das amostras. Assim, após alguns testes, o melhor valor encontrado foi 0.2, gerando menos falsos positivos e falsos negativos na matriz de confusão.

Um falso positivo é uma instância onde seu valor previsto foi positivo, mas seu valor real é negativo. Da mesma forma, um falso negativo é uma instância que teve valor previsto negativo, mas que apresentou um valor real positivo.

Outro parâmetro que teve os valores alterados durante o teste foi o de número mínimo de instâncias por folha. Os testes mostraram que valores altos desse parâmetro resultam em árvores mais genéricas, menores e com menos atributos sendo analisados para a classificação das amostras. Com valores próximos de um, as árvores geradas são mais precisas e mais atributos são analisados para classificar a amostra. O valor final escolhido foi dois.

## 5. Analise das árvores geradas

Nas próximas seções, são apresentas as analises das árvores geradas pelos algoritmos J48 e REPtree, mostrando o desempenho em relação ao número de instâncias classificadas corretamente e informações sobre os atributos considerados por cada algoritmo para a classificação final dos tumores.

### 5.1. Analise da árvore gerada pelo algoritmo J48

O tamanho da árvore gerada pelo algoritmo J48 foi 27, com o número de folhas igual a 14. Classificou 685 instâncias corretamente, aproximadamente 98% das amostras, e 14 incorretamente, 2% das amostras, além disso, foram encontrados 448 verdadeiros positivos e 10 falsos positivos, 237 verdadeiros negativos e 4 falsos negativos.

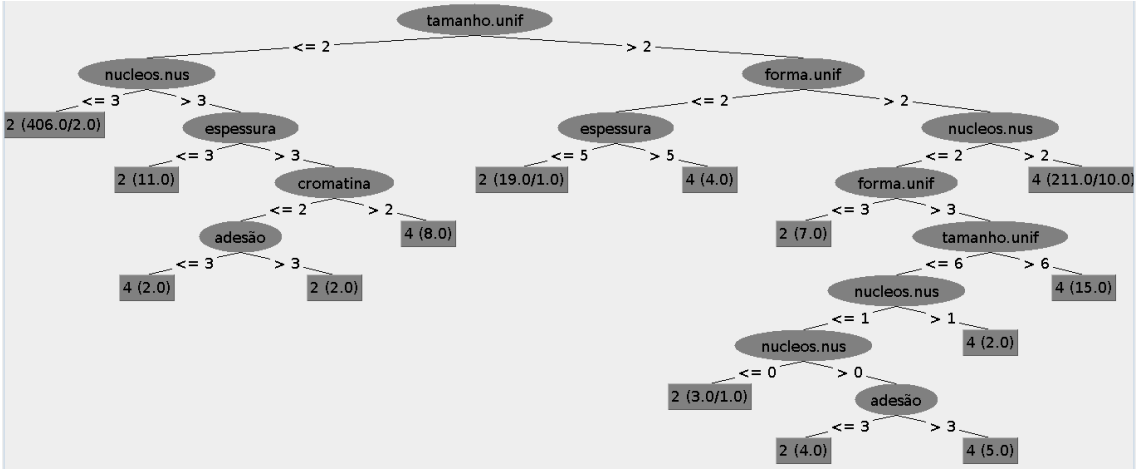


**Figura 3. Árvore gerada pelo algoritmo J48.**

Na árvore gerada pelo algoritmo J48, a primeira característica a ser avaliada é a uniformidade de tamanho das células, onde 1 representa uniformidade total de tamanho e 10 não uniformidade de tamanho. No caso de tumores onde a uniformidade de

tamanho das células é total ou igual a 90% o próximo atributo a ser avaliado é a quantidade de células com citoplasma no núcleo. Um tumor onde 30% ou menos das células têm citoplasma no núcleo é classificado como benigno, já um tumor com mais de 30% das células contendo citoplasma no núcleo, precisa da avaliação do atributo espessura para ser classificado. Tumores com todas, 90% ou 80%das células monocamada são considerados benignos, caso contrário a cromatina das células é analisada. Nos casos onde mais de 20% de deformação da cromatina das células os tumores foram classificados como malignos e nos casos onde a deformação era de 20% ou menos, o atributo adesão foi avaliado. Células com adesão total, 90% ou 80% são classificadas como malignas e células com menos de 80% de adesão são classificadas como benignas.

No ramo direito da árvore são classificadas as células com menos de 90% de uniformidade do tamanho. Nesse caso é considerada a uniformidade da forma, células totalmente ou 90% uniformes são classificadas analisando a espessura das mesmas. Tumores quem contém células com espessura média, 50% monocamada, ou menor são classificados benignos, já os tumores com espessura maior são considerados malignos. Considerando células com menos de 90% de uniformidade são classificadas analisando os núcleos nus, ou seja, a quantidade de célula com citoplasma núcleo. Tumores com mais de 20% de células com citoplasma no núcleo são considerados malignos, no caso de tumores com 20%, ou menos, de células com citoplasma no núcleo é preciso analisar a uniformidade da forma. Células totalmente, 90% ou 80% uniformes são consideradas benignas, já as células com menos de 80% de uniformidade são classificadas analisando a uniformidade do tamanho. Tumores com uniformidade de tamanho menor que 50% são classificados malignos. Para tumores com 50%, ou mais, de uniformidade de tamanho é necessário analisar a quantidade de células com citoplasma no núcleo, se mais de 10% das células contém citoplasma, o tumor é classificado como maligno caso 10% das células contenha citoplasma, é necessário analisar a adesão. Por fim, tumores que contém uma quantidade insignificante de células com citoplasma no núcleo e/ou células com uma adesão menor ou igual a 80% são considerados benignos e tumores com adesão maior que 80% são considerados malignos.

### 5.2. Analise da árvore gerada pelo algoritmo REPtree

Para o algoritmo REPtree o tamanho da árvore gerada foi 21, com o número de folhas igual a 14. O algoritmo classificou das 699 amostras, 681 corretamente, aproximadamente 97%, e 18 incorretamente, em torno de 3% das amostras, ainda foram encontrados 447 verdadeiros positivos e 11 falsos positivos, 234 verdadeiros positivos e 7 falsos negativos.

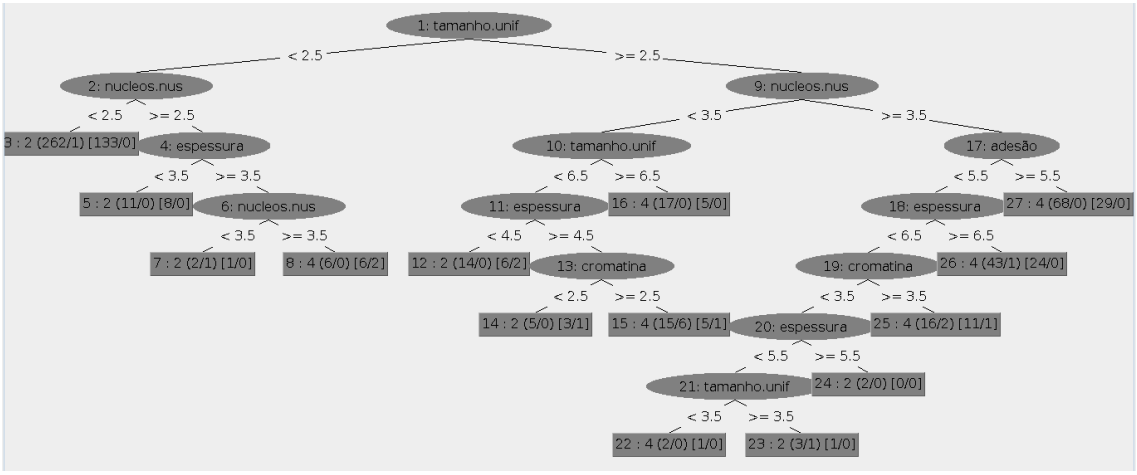


**Figura 4. Árvore gerada pelo algoritmo REPtree.**

No caso do algoritmo REPtree o primeiro atributo a ser avaliado foi também a uniformidade de tamanho. Na aresta esquerda, são classificadas tumores com uniformidade de tamanho das células total ou cerca de 90 a 80%. Para a classificação o próximo atributo avaliado é a quantidade de células com citoplasma no núcleo. No caso em que maioria, 90% ou 80% das células tinham núcleos desprovidos de citoplasma os tumores foram classificados como benignos e no caso em que 30% ou mais das células tinham citoplasma no núcleo foi necessário avaliar o atributo espessura, tumores onde 85%, ou mais, das células são monocamadas foram considerados benignos e tumores com menos de 85% das células monocamadas são classificadas analisando o atributo de quantidade de células com citoplasma no núcleo. Se menos de 30% das células continham citoplasma no núcleo, os foram classificadas como benignos, caso contrário foram classificadas como malignos.

Na aresta direita são classificados os tumores com uniformidade de tamanho das células menor que 90%. O próximo atributo avaliado é quantidade de células com citoplasma no núcleo. Tumores com menos de 35% das células contendo citoplasma no núcleo foram classificados avaliando novamente a uniformidade de tamanho. Aqueles que contêm uniformidade de tamanho das células menor ou igual a 50% são considerados malignos, já tumores com uniformidade de tamanho maior que 50% foram classificados avaliando a espessura das células. Os tumores com mais de 70% das células monocamadas foram classificados como benignos e os que continham em torno de 70%, ou menos, das células monocamadas foram classificados a partir do atributo cromatina. Para esse atributo, células com cromatina completamente, ou 95% dela, bem formada são consideradas benignas, já as células que possuem alguma porcentagem de anormalidade na cromatina foram consideradas malignas. No caso de tumores onde 30%, ou mais, das células contém citoplasma no núcleo, a próxima característica a ser avaliada é a adesão. Se a adesão é menor que 60%, os tumores são considerados malignos e se a adesão for maior que 60% o atributo espessura deve ser avaliado. Tumores onde 60%, ou mais, das células são multicamadas são considerados malignos e tumores com menos de 60% das células multicamadas são classificados analisando o atributo cromatina. Para tumores com 30%, ou mais, da cromatina das células mal formada são considerados malignos, já para tumores com células onde menos de 30% da cromatina está deforma são classificados avaliando a característica espessura.

Tumores com 60%, ou menos, das células monocamadas são considerados benignos. No caso de tumores com mais de 60% das células monocamadas é necessário analisar do atributo de uniformidade de tamanho. Tumores onde 80%, ou mais, das células tem uniformidade de tamanho são classificados como malignos e tumores com células que contém menos de 80% de uniformidade são considerados benignos.

## 6. Resultados obtidos

O câncer de mama é um tumor maligno consequência de alterações genéticas em algum grupo de células que passam a dividir-se indefinidamente [Naoum, 2008]. Algumas das alterações celulares causadas pelo aparecimento do tumor são alterações de tamanho, alterações no núcleo, citoplasma e cromatina das células [Sandra, 2008].

Analisando os resultados obtidos através do experimento é possível observar que o principal fator para a classificação dos tumores em benignos ou malignos é a quantidade de citoplasma presente no núcleo das células. No caso da análise da árvore gerada pelo J48 pode-se concluir que tumores benignos possuem uma uniformidade de tamanho regular, em torno de 90%, e que menos de 30% das células possuem citoplasma no núcleo. Nos tumores considerados malignos as células possuem uniformidade de tamanho baixa e a maioria delas contém citoplasma no núcleo. Para o algoritmo REPtree, os resultados para tumores benignos foram semelhantes aos obtidos pelo J48, as células possuem uniformidade de tamanho regular e, na maior parte, seus núcleos são desprovidos de citoplasma. Os tumores malignos foram classificados levando em conta, além do tamanho e citoplasma, a adesão e a quantidade de células multicamadas, ou seja, tumores foram considerados malignos quando a uniformidade de tamanho das células era menor que 90%, a quantidade de células com citoplasma no núcleo foi maior ou igual a 30% e a porcentagem de adesão das células foi menor ou igual que 60%, ou ainda, quando a adesão das células era maior que 60% e menos de 50% das células eram multicamadas. Os resultados obtidos confirmam padrões que ajudam na descoberta e classificação de tumores, tornando possível um diagnóstico mais rápido, e consequentemente um tratamento precoce e mais efetivo.

Durante os experimentos e análises dos resultados, foi constatada a necessidade do uso de um dataset mais atual, que permita não só a confirmação de padrões já existentes, mas sim a descoberta de novos conhecimentos capazes de contribuir para as pesquisas na área de câncer de mama. Além disso, torna-se importante o uso de mais algoritmos de árvore de decisão que podem trazer resultados mais detalhados.

## 7. Conclusão

A técnica de mineração de dados tem exercido papel importante na obtenção de conhecimento na área da saúde, contribuindo na confirmação e descoberta de padrões em grandes bases de dados. Dessa forma, a mineração torna-se essencial para a exploração e analise de dados sobre o câncer de mama, uma das causas mais comuns de morte entre as mulheres de 35-54 anos [GBECAM].

A decisão pelo uso de algoritmos de árvore de decisão foi tomada devido à utilidade desse tipo de técnica supervisionada na construção de perfis. No experimento foram aplicados sobre um conjunto de 699 amostras de tumores de câncer de mama, algoritmos presentes no software WEKA. A base de dados continha amostras de tumores malignos e benignos, com característica de deformação e/ou alteração celular, recolhidas pelo Dr. Wolberg durante consultas periódicas.

Os algoritmos escolhidos, J48 e REPTree, levaram a resultados relevantes, como a significância da quantidade de citoplasma no núcleo para a classificação dos tumores, e mostraram-se eficientes, classificando mais de 96% das amostras corretamente.

O intuito final do trabalho era a descoberta e confirmações de padrões comumente observados em tumores de câncer de mana, buscando tornar a compreensão sobre as características dessa doença mais fácil e clara. Dessa forma, inserindo pessoas ligadas a informática no contexto da saúde e mostrando a importância da tecnologia na vida humana.

Os resultados obtidos levaram a confirmação de padrões de alteração celular em tumores, auxiliando no estudo, reconhecimento e tratamento de tumores considerados malignos.

## Referências


Alessi, A. "Distúrbios do crescimento", 2011. Disponível: tumor_rot_aulas_teor_2011_vet.doc - Unesp, Acesso: setembro/2015.

Amorin, T. "Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados", 2006. Disponível: http://www.cin.ufpe.br/~tg/2006-2/tmas.pdf, Acesso: julho/2015.

Berry, M and Linoff, G. "Data Mining Tehniques – For Marketing, Sales and Customer Relationship Management. Second Edition". United States: Wiley Computer Publishing, 2004.

Caraciolo, M. "Introdução a Árvores de decisão para classificação e mineração de dados", 2009. Disponível: http://aimotion.blogspot.com.br/2009/04/artigo-introducao-arvores-de-decisao.html, Acesso: julho/2015

Damasceno, M. "Introdução a Mineração de Dados Utilizando o WEKA". Disponível: http://connepi.ifal.edu.br/ocs/index.php/connepi/CONNEPI2010/paper/viewFile/258/207, Acesso: setembro/2015

Developer Network. "Matriz de classificação (Analysis Services - Mineração de dados)". Disponível: https://msdn.microsoft.com/pt-br/library/ms174811(v=SQL.120).aspx, Acesso: setembro/2015

Feofiloff, P. "Algoritmos gulosos", 2015. Disponível: http://www.ime.usp.br/~pf/analise_de_algoritmos/aulas/guloso.html, Acesso: setembro/2015

GBECAM. "Câncer de Mama". Disponível: http://www.gbecam.org.br/, Acesso: setembro/2015

Halmenschlager, C. "Um Algoritmo Para Indução De Árvores E Regras De Decisão", 2002. Disponível: http://www.lume.ufrgs.br/bitstream/handle/10183/2755/000325797.pdf?sequence=1, Acesso: julho/2015

KNIME WEKA. "REPTree". Disponível: https://www.knime.org/files/nodedetails/weka_classifiers_trees_REPTree.html, Acesso: setembro/2015

Koerich, A. “Estratégia Gulosa, Técnicas de Projeto de Algoritmos”, 2006. Disponível: http://www.ppgia.pucpr.br/~alekoe/PAA/20061/14-EstrategiaGulosa-PAA2006.pdf, Acesso: setembro/2015

Martins, A., Marques, J. e Costa, P. “Estudo Comparativo De Três Algoritmos De Machine Learning Na Classificação De Dados Electrocardiográficos”, 2009. Disponível: http://www.dcc.fc.up.pt/~ines/aulas/0910/MIM/trabs_ano_anterior/noname-1.pdf, Acesso: setembro/2015

Naoum, P. “Biologia do Câncer”, 2008. Disponível: http://www.portaleducacao.com.br/biologia/artigos/2102/biologia-do-cancer, Acesso: agosto/2015

Nievola, J. “Mineração de Dados: Dados”. Disponível: http://www.ppgia.pucpr.br/~fabricio/ftp/Aulas/Mestrado/IA/Nievola/MD/MD-02-Dados.pdf, Acesso: agosto/2015

Paulo, M. “Comparação dos Atributos Escolhidos pelo Treinamento de Classficadores de Árvores de Decisão com Seleção de Atributos por Filtro”. Disponível: http://wiki.dpi.inpe.br/lib/exe/fetch.php?media=wiki:mauriciodepaulo:podas_geodma.pdf, Acesso: setembro/2015

Rodrigues, M. “Árvores de Classificação, 2004/2005”. Disponível: http://www.amendes.uac.pt/monograf/monograf05arvoreClass.pdf, Acesso: setembro/2015

Sandra. “Citologia Mamária”, 2008. Disponível: http://fapi.br/conteudo/conteudo_programatico/farmacia/citologia_clinica_4ano_sandra02.pdf, Acesso: agosto/2015

William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A.