UNIVERSIDADE ESTADUAL DO PIAUÍ – UESPI
CAMPUS PROF. ALEXANDRE ALVES DE OLIVEIRA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

DANIEL DE ARAUJO LIMA SOBRINHO

ANÁLISE DE DESEMPENHO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DE EXAMES DE CÂNCER DO COLO DO ÚTERO

PARNAIBA
2018

DANIEL DE ARAUJO LIMA SOBRINHO

ANÁLISE DE DESEMPENHO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DE EXAMES DE CÂNCER DO COLO DO ÚTERO

Trabalho de Conclusão de Curso submetido ao Curso de Bacharelado em Ciência da Computação da Universidade Estadual do Piauí, campus Prof. Alexandre Alves de Oliveira, Parnaíba, Piauí, como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. MSc. Rodrigo A. R. S. Baluz

PARNAIBA

2018

L732a Lima Sobrinho, Daniel de Araujo.
Análise de desempenho de algoritmos de aprendizado de máquina na classificação de exames de câncer do colo do útero / Daniel de Araujo Lima Sobrinho. - 2018.
52f. : il.

Monografia (graduação) – Universidade Estadual do Piauí - UESPI, Curso Bacharelado em Ciência da Computação, *Campus* Prof. Alexandre Alves de Oliveira, Parnaíba-PI, 2018.
"Orientador(a): Prof. Msc. Rodrigo A. R. S. Baluz."

1. Aprendizado de Máquina. 2. WEKA. 3. Exame de Papanicolaou. I. Título.

CDD: 004

Ficha elaborada pelo Serviço de Catalogação da Biblioteca Central da UESPI
Grasielly Muniz (Bibliotecária) CRB 3/1067

DANIEL DE ARAÚJO LIMA SOBRINHO

ANÁLISE DE DESEMPENHO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DE EXAMES DE CÂNCER DO COLO DO ÚTERO

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Estadual do Piauí – UESPI, Campus Prof. Alexandre Alves de Oliveira, como parte das exigências da disciplina de Estágio Supervisionado, requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Me. Rodrigo Augusto Rocha Souza Baluz

Monografia Aprovada em: **28 de julho de 2018**.

BANCA EXAMINADORA:

Prof. Me. Rodrigo Augusto Rocha Souza Baluz
Orientador - UESPI

Prof. Me. Leinylson Fontinele Pereira
Avaliador - UESPI

Prof. Me. Dario Brito Calçada
Avaliador - USP

Dedico aos meus pais, por todo amor, apoio e incentivo que sempre me deram; à minha irmã que, mesmo estando longe, sempre me encorajou com enorme carinho; e principalmente a Deus, que em Seu infinito amor tem me sustentado durante essa jornada.

AGRADECIMENTOS

Agradeço aos meus pais amados, Genes Brito Sobrinho e Eleonora de Araujo Lima Sobrinho, por serem os melhores pais que Deus poderia ter me dado, pela minha vida e por todo amor, carinho e incentivo. Agradeço também por sempre buscarem o melhor para mim e por todo o esforço para que nada me faltasse.

À melhor irmã que o mundo já viu, Fernanda Lima (Rose), por ser uma peça fundamental na minha vida, por acreditar em mim e por estar sempre me enchendo de amor, carinho e apoio. Agradeço também ao meu cunhado, Sérgio Marinho, pois, junto à minha irmã, sempre torceu por mim e me incentivou a fazer aquilo que me traz felicidade.

A toda minha família, que sempre se mostrou preocupada com meus estudos e sempre me motivou a realizar meus sonhos.

Ao meu professor orientador, Rodrigo Baluz, por todo o apoio dado a mim, não só na realização deste trabalho, mas também durante o decorrer do curso, onde sempre motivou os alunos a buscarem ir cada vez mais longe.

Aos meus amigos da melhor turma do mundo, que me fizeram sorrir todos os dias durante esses 4 anos de curso, tornando a jornada muito mais fácil e empolgante. Agradeço especialmente à minha querida amiga Stéfane de Sousa Santos, por todo apoio, parceria, conversas, momentos de desespero, partidas de uno, trabalhos em dupla, enfim, por ser uma pessoa incrível.

À minha outra grande amiga Hariadne Carvalho (Sari) que, apesar de ter me abandonado no meio do caminho, sempre foi uma amiga incrível, tornando minhas tardes muito mais divertidas com nossas conversas, joguinhos, e muitos memes.

Aos professores do curso de Ciência da Computação da UESPI, pelo carinho e parceria, e por todo conhecimento oferecido durante a graduação.

E, principalmente, a Deus, pois se hoje estou aqui é porque Ele, em Seu infinito amor e misericórdia, me salvou e continua a me salvar todos os dias. A Ele toda a honra e glória.

“Ele fortalece o cansado e dá grande vigor ao que está sem forças. Até os jovens se cansam e ficam exaustos, e os moços tropeçam e caem; mas aqueles que esperam no Senhor renovam as suas forças. Voam alto como águias; correm e não ficam exaustos, andam e não se cansam. ”
(Isaías 40:29-31)

RESUMO

O câncer do colo do útero é uma doença silenciosa, que apresenta maior ocorrência em países menos desenvolvidos. A melhor forma de prevenção e controle do câncer do colo do útero é a detecção precoce, realizada por meio do Exame de Papanicolaou. As alterações celulares nas células do colo do útero são as principais indicadoras de formação de tumores com suspeita de malignidade. A identificação dessas alterações é uma tarefa que necessita bastante atenção, a fim de minimizar erros de interpretação. Este trabalho consiste na análise do desempenho de algoritmos de aprendizado de máquina (J48, *Random Forest*, *Naive Bayes*, *Multilayer Perceptron*) na classificação de células do colo do útero e na identificação de possíveis anomalias. O desempenho dos algoritmos foi analisado através da ferramenta de mineração de dados *Waikato Environment for Knowledge Analysis* (WEKA), seguindo algumas métricas de avaliação. Os experimentos foram realizados em uma base de dados contendo descrições completas de células do colo do útero, bem como suas 7 possíveis classificações, fornecida pelo Hospital da Universidade de Herlev, na Dinamarca. Além da base original, foram realizados experimentos em uma base secundária, onde a quantidade de classificações possíveis foi reduzida a duas: normal e anormal. Após esses experimentos, foi escolhido o algoritmo de melhor desempenho geral para ser testado utilizando a técnica de seleção de atributos, que analisa a base em busca dos atributos mais relevantes antes de realizar a classificação. O melhor resultado foi obtido com o algoritmo *Multilayer Perceptron*, com a seleção de atributos na base de dados secundária, que obteve uma taxa de acerto de 94,44%, com um índice de concordância considerado excelente. A taxa de falsos positivos para normalidade foi de 3,6%, indicando que poucas células anormais foram classificadas como células normais. Os resultados obtidos mostram que os algoritmos de aprendizado de máquina possuem alta capacidade para identificar padrões e para realizar tarefas de classificação, revelando grande potencial para utilização na área médica.

PALAVRAS-CHAVE: Câncer do Colo do Útero. Exame de Papanicolaou. Aprendizado de Máquina. WEKA.

ABSTRACT

Cervical cancer is a silent disease, which is most prevalent in less developed countries. The best way to prevent and control cervical cancer is early detection through Pap Smear. Cellular changes in the cells of the cervix are the main indicators of the formation of tumors with suspected malignancy. The identification of these changes is a task that needs a lot of attention in order to minimize errors of interpretation. This work consists of the analysis of the performance of machine learning algorithms (J48, Random Forest, Naive Bayes, Multilayer Perceptron) in the classification of cells of the cervix and in the identification of possible anomalies. The performance of the algorithms was analyzed through the data mining tool Waikato Environment for Knowledge Analysis (WEKA), following some evaluation metrics. The experiments were conducted in a database containing complete descriptions of cervical cells, as well as their 7 possible classifications, provided by the Herlev University Hospital, Denmark. In addition to the original database, experiments were performed on a secondary database, where the number of possible classifications was reduced to two: normal and abnormal. After these experiments, the algorithm with the best overall performance was chosen to be tested using the attribute selection technique, which analyzes the base in search of the most relevant attributes before performing the classification. The best result was obtained with the Multilayer Perceptron algorithm, with the selection of attributes in the secondary database, which obtained a 94.44% accuracy rate, with a concordance index considered excellent. The rate of false positives for normality was 3.6%, indicating that few abnormal cells were classified as normal cells. The results show that machine learning algorithms have high capacity to identify patterns and to perform classification tasks, revealing great potential for use in the medical field.

KEYWORDS: Cervical Cancer. Pap Smear. Machine Learning. WEKA.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O câncer do colo do útero, ou cervical, é uma doença causada pela infecção persistente por alguns tipos do Vírus do Papiloma Humano (HPV). Apesar de a infecção genital por este vírus, mesmo sendo muito frequente, não causar a doença na maioria das vezes, em alguns casos podem ocorrer alterações celulares que poderão evoluir para o câncer (INCA, 2018). O desenvolvimento desta doença é lento e pode evoluir para uma lesão maligna em um período de 10 a 20 anos (ARAÚJO, 2017).

Apesar da possibilidade de prevenção com o exame citopatológico, mais conhecido como exame de Papanicolaou, o câncer do colo do útero continua a ser uma causa significativa de mortalidade em países de baixa renda (FERNANDES; CARDOSO; FERNANDES, 2017). É o quarto tipo de câncer mais comum entre as mulheres, com exceção dos casos de câncer de pele não-melanoma, com a estimativa aproximada de 530 mil casos novos por ano no mundo. É a quarta causa mais frequente de morte por câncer em mulheres, responsável por aproximadamente 265 mil óbitos por ano (INCA, 2018).

Segundo o Instituto Nacional do Câncer (INCA), no Brasil, em 2016, foram esperados 16.340 casos novos, com uma taxa estimada de 15,85 casos a cada 100 mil mulheres. Em 2013, ocorreram 5.430 óbitos por esta neoplasia, representando uma taxa de mortalidade ajustada para a população mundial de 4,86 óbitos para cada 100 mil mulheres.

O maior problema com este tipo de câncer é a dificuldade em seu rastreio, pois os sintomas e outros indicativos são tipicamente notados apenas em estágios mais avançados da doença (SHARMA; GUPTA, 2016). Porém, caso a doença seja diagnosticada em sua fase inicial, na qual os sintomas ainda não são visíveis, a chance de recuperação do câncer podem chegar a 100% (INCA, 2018).

Os maiores índices de ocorrência do câncer de colo uterino são observados nos países menos desenvolvidos, o que indica a ligação deste tipo de câncer com fatores como os baixos índices de desenvolvimento humano, falta de estratégias de conscientização das comunidades, problemas nos sistemas públicos de saúde, entre outros (SANTOS et al., 2014). Uma vez doentes, as mulheres têm seus papéis no mercado de trabalho comprometidos e também acabam sendo privadas do convívio familiar, resultando em um grande prejuízo social (BRENNA et al., 2001).

A realização de exames preventivos com a população em períodos de 3 a 5 anos pode reduzir a incidência do câncer de colo uterino em até 80%. Para maior garantia de sucesso, é necessário um grau elevado de qualidade em cada etapa do processo de diagnóstico. É preciso

fornecer informação, incentivo, exame preventivo de qualidade e, caso necessário, tratamento para as mulheres que precisarem (ARBYN et al., 2010).

A taxa de falso-negativo do exame de Papanicolaou pode variar até 30% dependendo da subjetividade e de outros fatores, como a coleta do material, leitura do esfregaço (lâmina de vidro contendo o material coletado) e a interpretação do exame (BRASIL, 2002). Segundo Mehta, Vasanth e Balachandran (2009) o erro humano é provavelmente a maior ameaça quando se busca uma interpretação precisa. Normalmente, o esfregaço do exame contém de 50.000 a 300.000 células que precisam ser examinadas. Caso a amostra contenha poucas células anormais entre muitas células saudáveis, é possível que as células anormais não sejam notadas.

Nos últimos anos, análises estatísticas e técnicas de Inteligência Artificial, como a mineração de dados e o aprendizado de máquina, têm sido amplamente utilizadas para solucionar problemas na área da saúde (SARWAR et al., 2016). Aplicar a inteligência artificial na medicina possibilita a criação de sistemas capazes de auxiliar os profissionais da saúde nas tomadas de decisões e na melhora de seus serviços (PEREIRA; CHAMORRO; ROMERO, 2012).

O aprendizado de máquina é um processo em que dados já existentes são utilizados para classificar novos dados, através da detecção de padrões, da matemática e da estatística. Desta forma, o aprendizado de máquina na classificação dos resultados de exames preventivos pode simplificar o trabalho dos médicos e diminuir o tempo para a obtenção de um diagnóstico preciso (KURNIAWATI; PERMANASARI; FAUZIATI, 2016).

O objetivo da pesquisa é analisar o desempenho de algoritmos de aprendizado de máquina, aplicados a uma base de dados pré-existente obtida através do exame citológico preventivo para a tarefa de detecção de anomalias em células do colo do útero. Para alcançar o objetivo principal, os seguintes objetivos específicos foram determinados: (i) estudar as áreas de saberes envolvidos no trabalho, especialmente o estado da arte de aprendizado de máquina na área da saúde; (ii) definir qual a melhor base de dados para uso nas etapas de treinamento e testes; (iii) realizar a etapa de pré-processamento na base de dados a fim de prepará-la para ser analisada; (iv) selecionar os algoritmos de aprendizado de máquina a serem utilizados nos experimentos por meio da ferramenta WEKA; (v) definir as métricas a serem utilizadas durante a avaliação dos resultados, a fim de quantificar o desempenho dos algoritmos no problema em questão.

Para a realização deste trabalho, adotou-se inicialmente uma pesquisa bibliográfica sobre as áreas relacionadas, buscando por trabalhos relacionados ao câncer do colo do útero e

à inteligência artificial. Após essa etapa, foi realizada a seleção de uma base de dados, de uma ferramenta de mineração de dados, e dos algoritmos aplicados nos experimentos.

A base de dados utilizada nos experimentos foi produzida pelo Hospital da Universidade de Herlev, na Dinamarca, e contém 917 registros com 20 atributos descritivos de células do colo do útero. A base conta com uma quantidade reduzida de instâncias, quando comparada à quantidade de células obtidas em um esfregaço do exame de Papanicolaou, porém, a proporção entre os atributos alvo (classes) das instâncias é bastante equilibrada, o que é um ponto crucial para garantir resultados satisfatórios nos testes. Além da base original, com 7 classes, foi preparada uma base secundária, onde as 7 classes foram resumidas em apenas duas, permitindo a classificação das células como sendo normais ou anormais.

A ferramenta escolhida para a realização dos experimentos foi a *Waikato Environment for Knowledge Analysis* (WEKA). A ferramenta foi escolhida por ser um software livre, apresentar interface simples de operar, e por contar com uma grande variedade de algoritmos de classificação, dos quais foram escolhidos: J48, *Random Forest*, *Naive Bayes e Multilayer Perceptron*. Estes algoritmos foram selecionados por serem populares na literatura, e por apresentarem propostas diferentes para treinamento e classificação.

Além deste capítulo introdutório, onde foi apresentada uma visão geral deste trabalho, o presente trabalho está organizado da seguinte maneira:

No capítulo 2, Câncer do Colo do Útero e o Exame de Papanicolaou, descreve-se conceitos relacionados ao câncer do colo do útero, assim como o principal método de prevenção da doença: o Exame de Papanicolaou.

No capítulo 3, Aprendizado de Máquina, são abordados conceitos relacionados ao Aprendizado de Máquina, bem como trabalhos que utilizaram suas técnicas a fim de solucionar problemas na área da saúde.

No capítulo 4, Materiais e Métodos, foi destacada a metodologia empregada na execução da pesquisa, com uma descrição da ferramenta e dos algoritmos utilizados, e também das métricas utilizadas para avaliação dos experimentos.

No capítulo 5, Análise e Discussão dos Resultados, são apresentados os principais resultados obtidos na pesquisa, relatando os pontos de relevância nos resultados.

O trabalho encerra com as Considerações Finais, onde se apresenta os principais resultados alcançados e sugestões para possíveis trabalhos futuros.

2 CÂNCER DO COLO DO ÚTERO E O EXAME DE PAPANICOLAOU

As décadas iniciais do século XX foram muito importantes no que tange ao estudo do câncer do colo do útero, pois muitas pesquisas inovadoras relacionadas ao diagnóstico e tratamento da doença foram realizadas. Dentre estas pesquisas temos o trabalho realizado por Papanicolaou e Traut em 1941, que introduziu a citologia cervical como método de diagnóstico (THULER, 2012). Um declínio anual constante de 70% na mortalidade por câncer do colo do útero tem sido observado desde a introdução do exame citológico, hoje conhecido como o exame de Papanicolaou. O exame ajudou a reduzir significativamente as taxas de incidência do câncer através da detecção de lesões pré-malignas, possibilitando o diagnóstico precoce da doença e maiores chances de recuperação (VALDESPINO, V. M.; VALDESPINO, V. E., 2006).

2.1 O COLO DO ÚTERO

O útero é um órgão do sistema reprodutor feminino que se localiza no abdome inferior, entre a bexiga e o reto e é dividido em corpo e colo, conforme pode ser observado na Figura 1 (BRASIL, 2002). Pode apresentar variações de forma, tamanho, localização e estrutura, dependendo de fatores como a idade, o número de gestações, e a estimulação hormonal (FILHO, 2011).

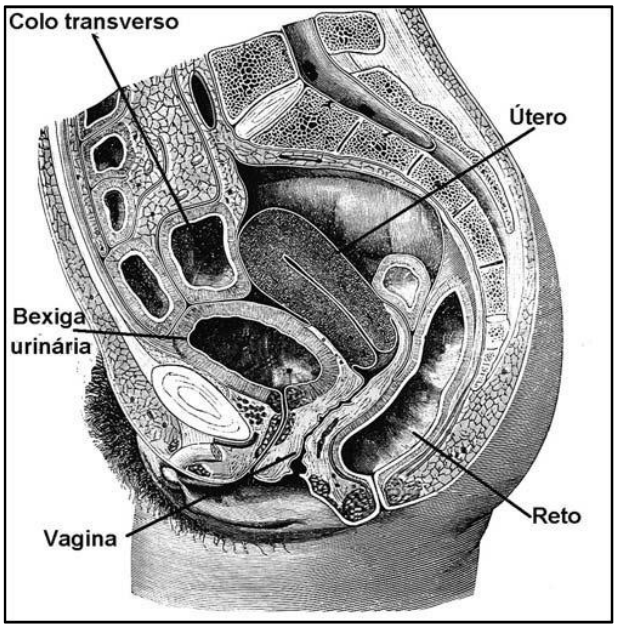

Figura 1 – Visão lateral do útero.
Fonte: FILHO, 2011.

O colo, também conhecido como cérvix, é a porção inferior do útero e está situado dentro da cavidade vaginal, conforme ilustrado na Figura 2 (BRASIL, 2002). Possui um formato cilíndrico ou cônico e mede de 3 cm a 4 cm de comprimento e 2,5 cm de diâmetro (SELLORS; SANKARANARAYANAN, 2004).

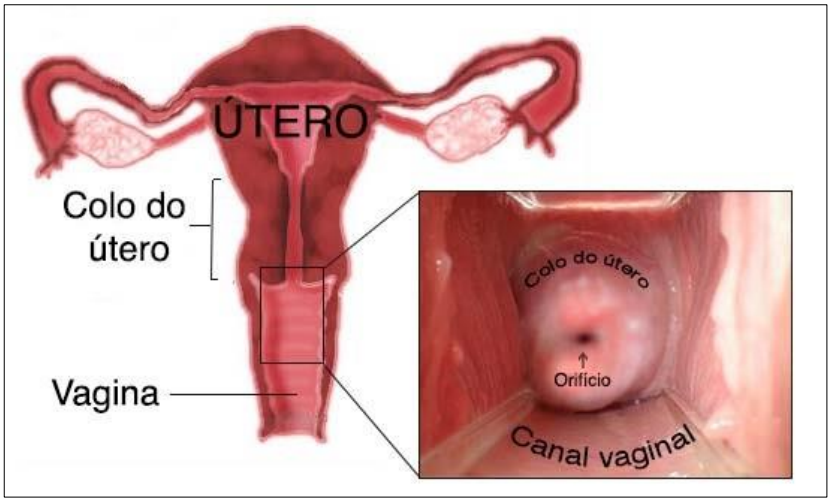

Figura 2 – O colo do útero.
Fonte: PINHEIRO, 2017.

O colo do útero possui uma parte interna, chamada de endocérvice, que é revestida pelo epitélio colunar simples, que consiste em uma camada única de células cilíndricas produtoras de muco. A parte externa, chamada de ectocérvice, está em contato com a vagina e é revestida pelo epitélio escamoso, que é um tecido de várias camadas de células planas. (BRASIL, 2002; FILHO, 2011). A localização destas partes pode ser observada na Figura 3.

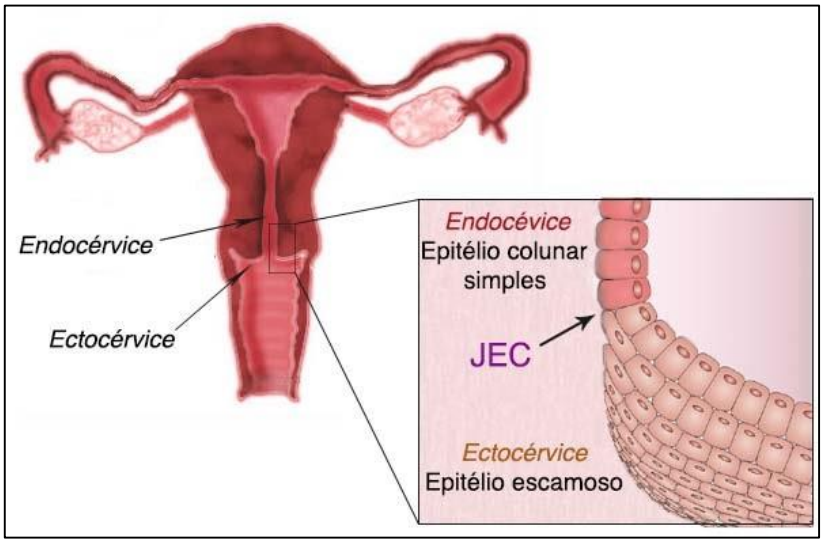

Figura 3 – Revestimento do colo do útero.
Fonte: PINHEIRO, 2017.

Na junção dos dois epitélios encontra-se a junção escamo-colunar (JEC). Esta junção apresenta variações na localização, podendo estar tanto na ectocérvice como na endocérvice, dependendo do estado hormonal, gestacional, parto vaginal e trauma.

No colo do útero, as células escamosas, presentes na ectocérvice, estão distribuídas entre 4 camadas distintas: A camada basal, a parabasal, a intermediária e a superficial, conforme ilustrado na Figura 4. As células mais jovens estão presentes na camada basal, logo acima da lâmina basal. Quando as células amadurecem, elas se movem para as camadas superiores, sofrendo aumento em seus citoplasmas e diminuição de seus núcleos (NORUP, 2005).

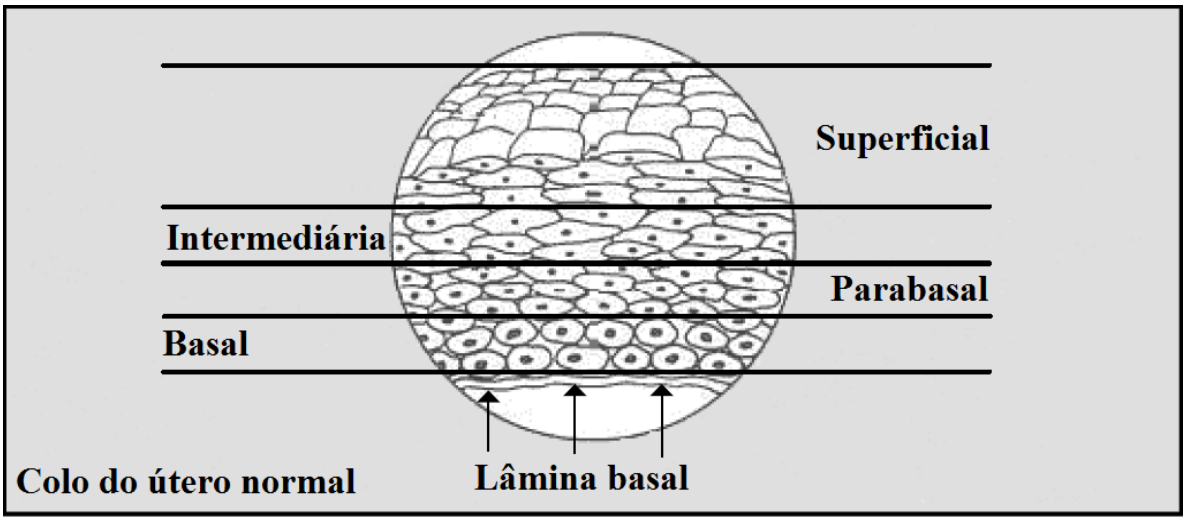

Figura 4 – As camadas do epitélio escamoso.
Fonte: Adaptada de NORUP, 2005.

As células colunares, ou glandulares, se localizam na endocérvice e estão dispostas em uma camada única, a camada basal. Estas células possuem formato cilíndrico, com o citoplasma alongado e um núcleo largo presente em uma das extremidades. Na JEC, também conhecida como zona de transformação, as células colunares estão sendo constantemente transformadas em células escamosas (NORUP, 2005).

De acordo com a Sociedade Americana do Câncer (*American Cancer Society* - ACS), aproximadamente 90% dos casos de câncer do colo do útero são do tipo que acomete as células escamosas, principalmente as que estão mais próximas da zona de transformação. A maioria dos casos restantes ocorre nas células colunares da endocérvice, chamados de Adenocarcinomas. Há ainda os tipos mais raros, chamados Carcinomas Adenoescamosos, que combinam características dos dois tipos.

2.2 CÂNCER DO COLO DO ÚTERO

O câncer do colo do útero é uma neoplasia maligna que se inicia com transformações intra-epiteliais progressivas, caso não seja detectado antecipadamente. É uma doença que se desenvolve apenas em mulheres, e em seu estágio inicial os sintomas são normalmente inexistentes, por isso acredita-se que caso o tratamento não seja realizado a tempo, o câncer pode invadir outros órgãos e estruturas do corpo (LINARD; SILVA; SILVA, 2002).

Apesar de ser uma doença de caráter progressivo, possui grandes chances de cura caso os tratamentos sejam realizados. O câncer do colo do útero é uma doença silenciosa, e tende a se estabelecer em um período de 10 a 20 anos. Quanto mais cedo a doença for diagnosticada e tratada, maiores são as chances de sobrevivência e menores os custos de tratamento (ARAÚJO, 2017; FILHO, 2011).

2.2.1 Fatores de Risco

O principal fator relacionado ao surgimento do câncer do colo do útero é a infecção pelo HPV. Estudos recentes mostram que o HPV, presente na quase totalidade dos casos da doença, possui grande parte no desenvolvimento das células do colo do útero em células cancerosas. A carga viral e o tipo do HPV são aspectos que irão determinar se a infecção persistente pelo vírus pode levar ao desenvolvimento da doença. A Tabela 1 apresenta os tipos do HPV, classificados de acordo com o grau de risco (BRASIL, 2002; FILHO, 2011; MELO et al., 2009; SILVA et al., 2006).

Tabela 1 – Tipos do HPV de acordo com grau de risco.

Baixo Risco	6, 11, 26, 40, 42, 53-55, 57, 59, 66, 68
Médio – Alto Risco	16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 59

Fonte: BRASIL, 2002.

Vários outros fatores também estão relacionados a ocorrência do câncer do colo do útero. Atualmente são conhecidos os seguintes fatores de risco que podem levar ao surgimento e desenvolvimento das lesões no cérvix: DSTs; número de gestações; promiscuidade sexual; tabagismo; uso de contraceptivos orais. Há também os fatores que não estão diretamente relacionados ao surgimento da doença, mas dificultam sua identificação precoce, como o medo

da mulher em realizar o exame preventivo, vergonha, nervosismo, falta de conhecimento e dificuldade de acesso aos serviços de saúde (BEZERRA et al., 2005; MELO et al., 2009).

2.2.2 Lesões Precursoras do Câncer do Colo do Útero

As células do colo do útero estão distribuídas em camadas, de maneira bastante organizada. Durante as displasias intra-epiteliais, a distribuição das células fica desordenada e também ocorrem alterações nas células, dando origem às lesões precursoras do câncer. O termo displasia significa "crescimento desordenado". As displasias das células cervicais são divididas em 3 tipos: leve, moderada e severa, descrevendo o risco que as células possuem de se tornarem células cancerígenas (BRASIL, 2002; NORUP, 2005).

As lesões precursoras do câncer do colo do útero possuem caráter evolutivo, e são classificadas como neoplasia intra-epitelial cervical (NIC) de graus I (lesões de baixo grau), II e III (lesões de alto grau) (MELO et al., 2009). A progressão dessas lesões e seus efeitos na estrutura das células do colo do útero pode ser observada na Figura 5.

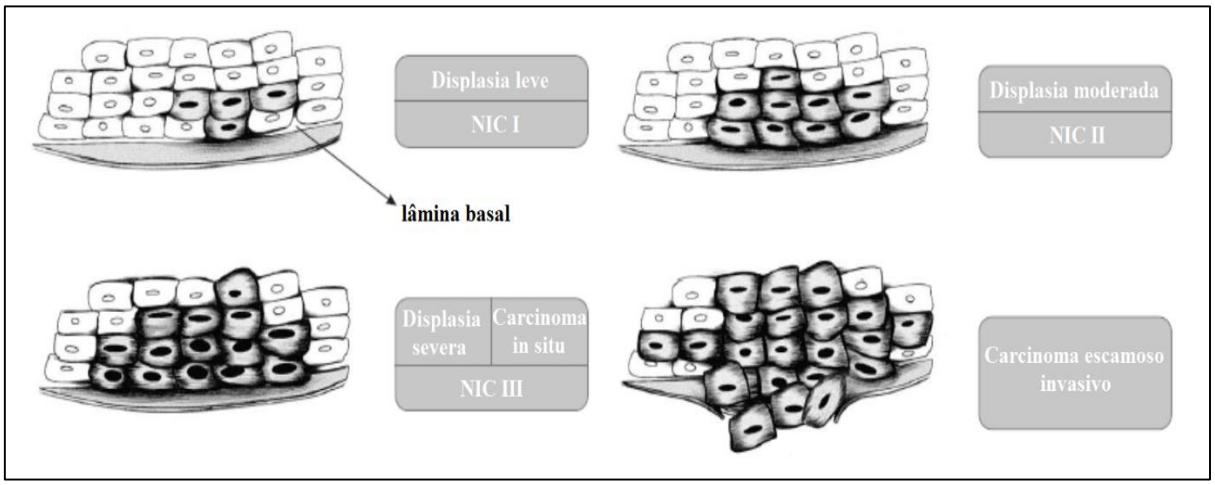

Figura 5 – Progressão das lesões precursoras do câncer do colo do útero.
Fonte: BRASIL, 2002.

A NIC I apresenta uma displasia leve e a desordem nas células é observada nas camadas mais próximas da lâmina basal. Na grande maioria dos casos, a NIC I regride espontaneamente ou permanece sem evoluir. Já a NIC II é caracterizada por uma displasia moderada, onde a desordem avança para camadas superiores, mas ainda sem atingir as camadas mais superficiais. No caso da NIC III, caracterizada por uma displasia severa, a desorganização pode ser vista em todas as camadas do epitélio (BRASIL, 2002).

O termo carcinoma in situ corresponde a “câncer não invasivo”. No passado, era comumente tradado como um problema muito mais sério que a displasia severa, quando na verdade são essencialmente a mesma coisa (NORUP, 2005).

Caso a NIC III não seja tratada a tempo, as alterações celulares se intensificam até o ponto em que as células atravessam a lâmina basal e invadem o tecido conjuntivo do colo do útero, caracterizando um carcinoma invasivo (BRASIL, 2002).

2.2.3 Sintomas

Os sintomas são normalmente inexistentes nos estágios iniciais, só aparecendo quando o câncer se torna invasivo e invade tecidos próximos. Quando isso ocorre, os sintomas mais comuns são: Sangramento vaginal anormal, como sangramento após atividade sexual; Fluxos menstruais mais intensos; Corrimento de odor fétido; Aumento da frequência urinária; Dores nas costas; Dores abdominais; Dor durante atividade sexual. (ACS, 2016b; SELLORS; SANKARANARAYANAN, 2004)

Alguns desses sintomas não são obrigatoriamente sinais de câncer do colo do útero, visto que sangramentos e dores também podem ser causados por infecções (ACS, 2016b).

2.3 O EXAME DE PAPANICOLAOU

Dependendo do estágio da doença e do dano sofrido pelas células, a progressão do câncer pode ser interrompida. Por este motivo, a realização do exame preventivo se torna um fator de extrema relevância no combate à doença (BRASIL, 2002).

O exame de Papanicolaou, também conhecido como citologia oncótica ou exame citopatológico, é um exame preventivo utilizado por vários países como forma de detecção do câncer do colo do útero em mulheres assintomáticas (SILVA et al., 2006). Recebeu este nome em homenagem ao Dr. George N. Papanicolaou, que descreveu a técnica pela primeira vez em 1928. Desde seu surgimento, o exame contribuiu na redução da mortalidade e da incidência do câncer do colo do útero em 75% (MEHTA; VASANTH; BALACHANDRAN, 2009). O exame possibilita a detecção do câncer ainda em sua fase não-invasiva, onde as chances de cura são muito elevadas (LEAL et al., 2003).

O exame consiste no estudo das células da ectocérvice e da endocérvice, extraídas através da raspagem do colo do útero. Para a realização da coleta, a combinação mais eficaz é o uso da escova para coleta do material endocervical e de uma espátula tipo ponta longa para

coleta do material ectocervical. Após a coleta, é preparado o esfregaço, que consiste em espalhar o material coletado em uma lâmina de vidro para que possa ser analisado através de um microscópio. O processo de raspagem e a confecção do esfregaço podem ser observados na Figura 6 (ANDRADE et al., 2001; BEZERRA et al., 2005).

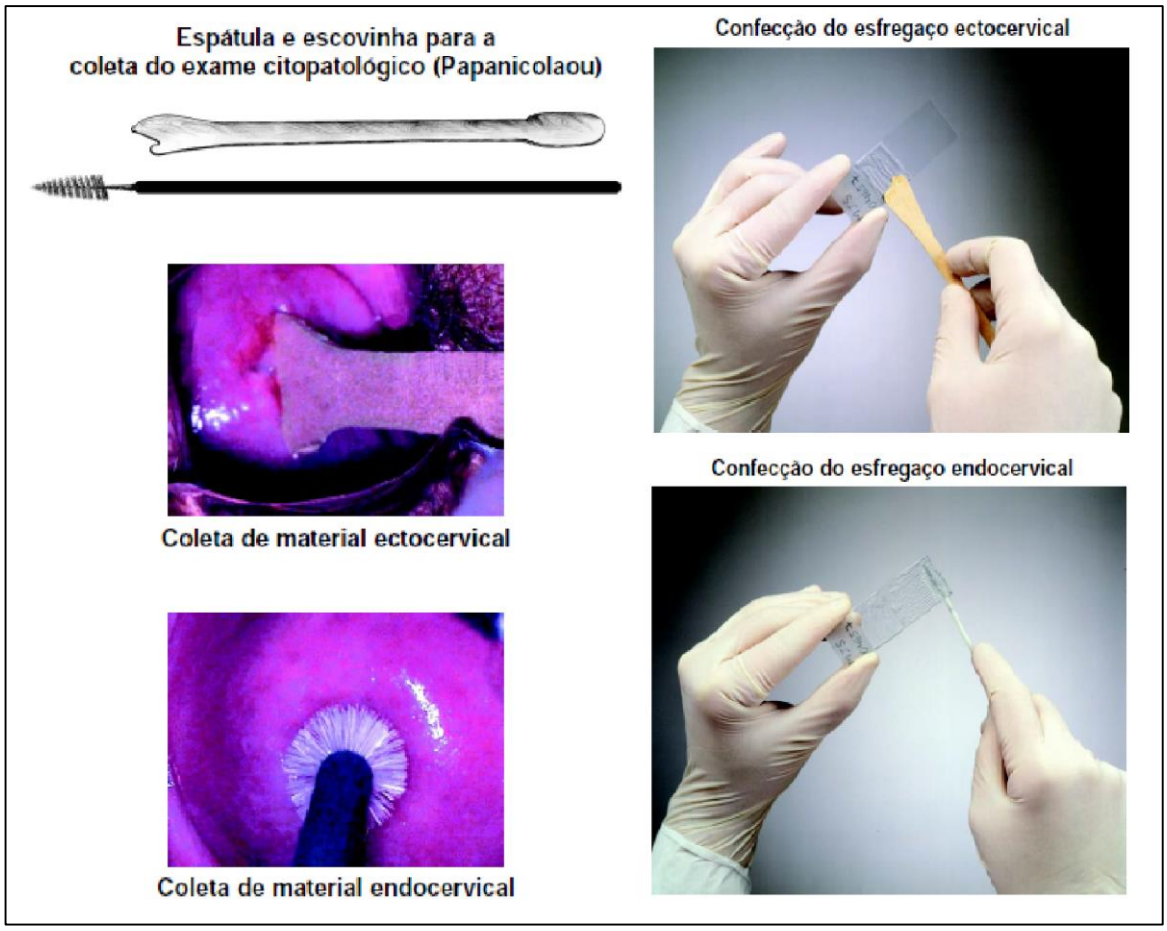

Figura 6 – Coleta do exame de Papanicolaou e preparo do esfregaço.
Fonte: BRASIL, 2002.

Por ser uma técnica altamente eficaz, normalmente indolor e de baixo custo, o exame de Papanicolaou é considerado ideal para a população brasileira, podendo ser realizado de maneira simples e rápida por qualquer profissional qualificado. O exame é oferecido gratuitamente no Brasil através do Programa Nacional de Controle do Câncer do Colo do Útero (BEZERRA et al., 2005; BRASIL, 2002).

O diagnóstico do exame citopatológico não é de certeza, e algumas vezes requer confirmação através de um exame histopatológico, também conhecido como biópsia, que consiste em uma análise mais aprofundada do tecido do colo do útero. A principal finalidade do exame de Papanicolaou é detectar antecipadamente as alterações celulares pré-malignas na

mucosa do colo do útero, permitindo assim que as medidas necessárias sejam tomadas a tempo, a fim de evitar a evolução das alterações celulares para um câncer invasivo (BRASIL, 2002; PINHEIRO, 2017).

O sistema de classificação do exame tem sido refinado com o passar dos anos. O sistema atual para classificação dos achados é o sistema de Bethesda, que foi introduzido em 1988 e posteriormente atualizado em 1999. Os achados do exame e suas interpretações de acordo com o sistema podem ser observados na Tabela 2 (MEHTA; VASANTH; BALACHANDRAN, 2009).

Tabela 2 – Classificação Bethesda para o exame citopatológico.

Achados do Exame Citopatológico	**Interpretação**
Negativo para lesão intra-epitelial ou malignidade	Exame normal
Células escamosas atípicas de significado indeterminado (ASCUS)	Células escamosas anormais, porém não são classificadas como uma lesão intra-epitelial escamosa
Lesão intra-epitelial escamosa de baixo grau (LSIL)	NIC I, alterações normalmente atribuídas ao HPV
Lesão intra-epitelial escamosa de alto grau (HSIL)	NIC II e NIC III
Carcinoma	Presença do câncer é quase certa, requer outros exames

Fonte: Adaptada (MEHTA; VASANTH; BALACHANDRAN, 2009).

Em sua pesquisa, Norup (2005) utilizou características de células individuais coletadas através do exame, como tamanho, formato, área e brilho, para identificar os diferentes graus de lesões precursoras e diferenciar células saudáveis de células pré-malignas. A proporção Núcleo/Citoplasma (N/C) também foi um elemento descritivo importante, pois ela tende a aumentar em células pré-malignas.

A proporção N/C é definida da seguinte forma:

$$\frac{N}{C} = \frac{Área\ do\ Núcleo}{Área\ do\ Núcleo + Área\ do\ Citoplasma} \quad (1)$$

As características de acordo com o tipo de célula, bem como imagens obtidas através de análise microscópica, podem ser observadas na Figura 7.

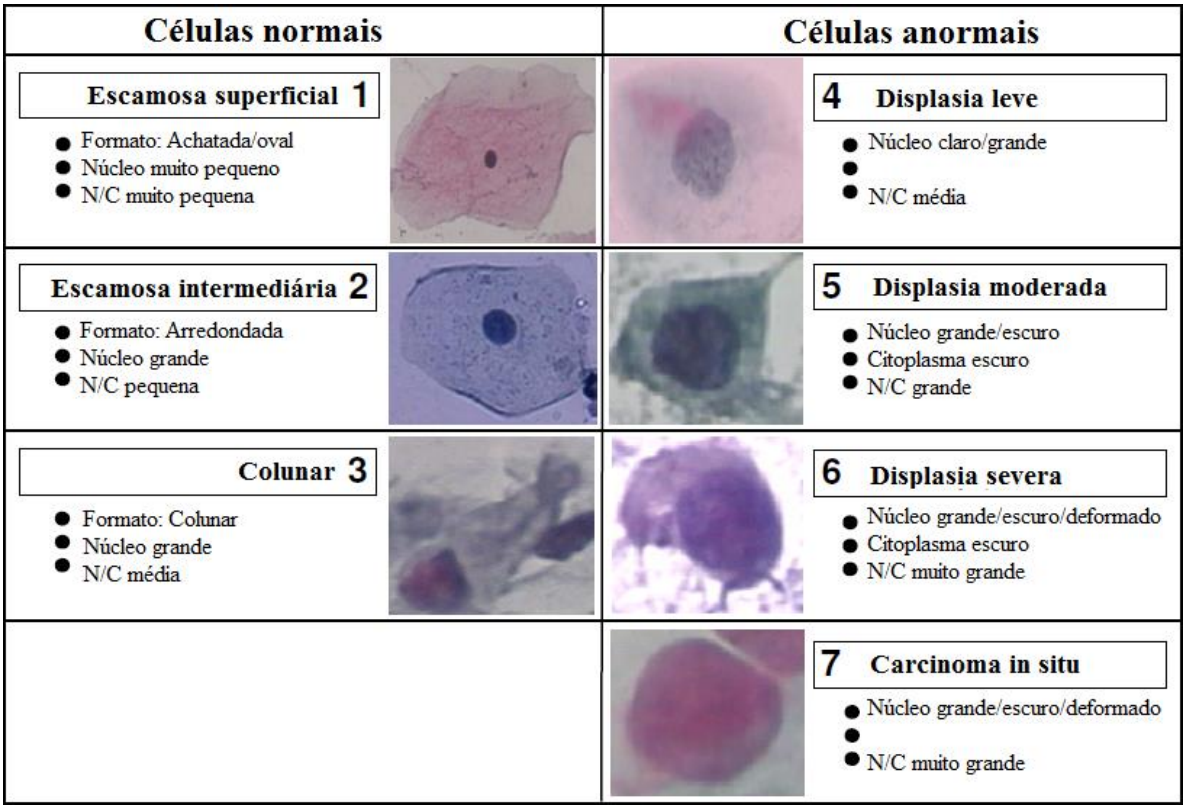

Figura 7 – Características de tipos de células coletadas através do exame de Papanicolaou.
Fonte: Adaptada de NORUP, 2005.

Na Figura 7, as células de numerações 1 e 2 correspondem, respectivamente, às células escamosas superficiais e intermediárias, e são encontradas na ectocérvice. Já a célula 3 corresponde às células colunares, presentes na endocérvice.

Já as numerações 4, 5 e 6 representam, respectivamente, células em displasia leve, moderada e severa, onde cada tipo descreve o risco que a célula possui de se tornar uma célula cancerosa maligna. Na displasia leve, as células possuem um núcleo largo e grande. Já a displasia moderada é caracterizada por um núcleo escuro e ainda maior, o qual já está em processo de deterioração. Na displasia severa, o núcleo é grande, escuro e normalmente deformado, e o citoplasma é escuro e pequeno, quando comparado ao núcleo (NORUP, 2005).

A célula de numeração 7 representa uma célula cancerosa, que possui um núcleo muito grande. Possui características muito similares às de células em displasia severa.

3 PROCESSO DE KDD E O APRENDIZADO DE MÁQUINA NA SAÚDE

A Inteligência Artificial tem se destacado entra as diversas de técnicas de computação presentes na literatura como ferramenta inovadora, sendo cada vez mais difundida e se mostrando extremamente eficaz quando comparada à programação convencional. Na área da medicina, técnicas de Inteligência Artificial têm sido amplamente utilizadas para automatizar os processos de diagnóstico. Dentre estas técnicas, destacam-se os Sistemas Especialistas e os Algoritmos de Classificação, sendo que estes se utilizam de estratégias de Aprendizado de Máquina (BALUZ; SANTOS, 2011; VERAS, 2015).

Ferro e Lee (2001) reconheceram a importância da aplicação de métodos capazes de auxiliar no processo de tomada de decisão e na descoberta de novos conhecimentos médicos específicos (diagnósticos, prognósticos, monitoramento, etc.). Acreditam também que métodos eficientes para a análise de dados médicos, com o auxílio de técnicas computacionais, se tornaram indispensáveis, assim como a extração das informações e padrões que existem nesses dados.

3.1 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

O processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* – KDD) possui aplicações nas mais diversas áreas do conhecimento, especialmente na saúde. O KDD consiste no processo de identificar padrões válidos, novos, úteis e compreensíveis em dados, com o objetivo de melhorar a compreensão de um problema ou um procedimento de tomada de decisão (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; SOUZA, 2015).

Graças à sua habilidade de encontrar padrões e relações em grandes volumes de dados, o processo de KDD pode ser utilizado em dados da área da saúde para identificar pacientes com risco de possuírem certas doenças. Dessa forma, as organizações de saúde podem mantê-los saudáveis e diminuir custos de tratamento (DEGRUY, 2000).

O KDD é um processo amplo composto pelas seguintes etapas: Seleção dos dados; Pré-processamento; Transformação; Mineração de Dados e Avaliação (SOUZA, 2015). As técnicas específicas podem variar dependendo do projeto onde o KDD está sendo aplicado, porém o processo básico permanece o mesmo na maioria das situações. As etapas do KDD estão ilustradas na Figura 8 e são descritas com mais detalhes a seguir.

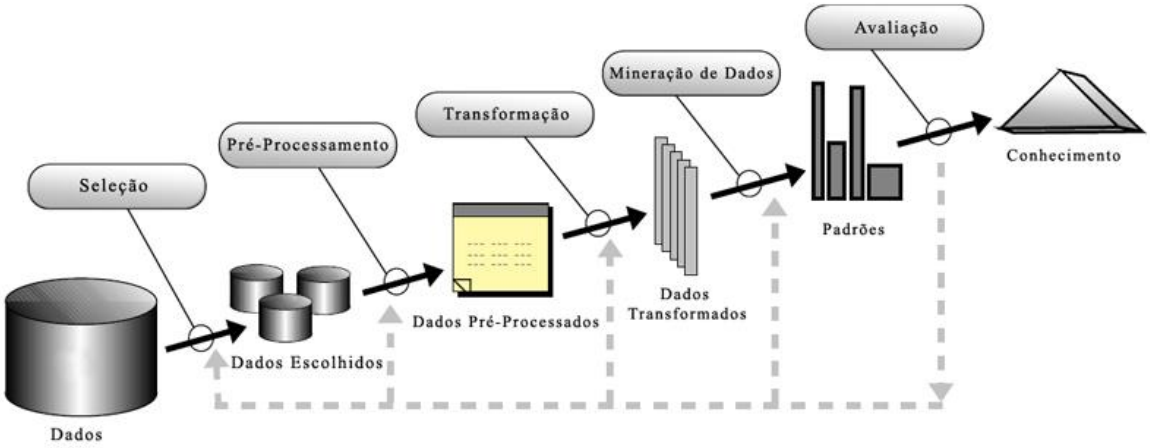

Figura 8 – O processo de KDD.
Fonte: CAMILO; SILVA, 2009.

- **Seleção dos dados:** Consiste em selecionar um conjunto de dados, podendo focar em um subconjunto de atributos que interessem ao usuário ou em uma amostra de dados, no qual será aplicada a descoberta do conhecimento (AMO, 2004; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

- **Pré-processamento:** Fase onde é realizada a limpeza dos dados relevantes, através da eliminação de ruídos e dados inconsistentes. Nessa etapa também são determinadas as estratégias para lidar com dados faltantes. Pode haver também a integração de dados, onde diferentes fontes de dados podem ser combinadas em uma única base (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; SOUZA, 2016).

- **Transformação:** Nessa fase, os dados são formatados de maneira adequada para a aplicação de algoritmos de aprendizado, responsáveis pela mineração de dados (AMO, 2004; SOUZA, 2015).

- **Mineração de Dados:** É uma etapa essencial do processo, apoiada pelo Aprendizado de Máquina. Consiste na aplicação de algoritmos de aprendizado de máquina, com o objetivo de extrair padrões relevantes nos dados (AMO, 2004; FERRO; LEE, 2001).

- **Avaliação:** Consiste em interpretar os padrões obtidos através da mineração, gerando conhecimento que pode ser utilizado pelo usuário e incorporado a outro sistema para tomada de decisões (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

3.2 APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina é um ramo da Inteligência Artificial que se utiliza de métodos estatísticos e probabilísticos para permitir ao computador "aprender" e detectar padrões em dados complexos e volumosos (CRUZ; WISHART, 2006). Os principais tipos de aprendizado são o Aprendizado Supervisionado e o Aprendizado Não Supervisionado, sendo o primeiro o tipo mais utilizado (SANTANA, 2018).

3.2.1 Aprendizado Supervisionado

Os algoritmos de Aprendizado Supervisionado são voltados para tarefas de previsão. Utilizam dados cujos registros são categorizados por um atributo classe, possibilitando o "aprendizado", ou seja, a detecção de padrões, e a classificação de novos dados com base no que foi aprendido. Suas principais tarefas são a Classificação (*Classification*) e a Regressão (*Regression*) ou Estimação (*Estimation*) (SANTANA, 2018; SOUZA, 2015).

- **Classificação**: A Classificação, método utilizado neste trabalho, é considerada uma das tarefas de aprendizado mais populares. Os algoritmos de classificação analisam um conjunto de dados, onde seus registros contém uma classe definida, para aprender a classificar novos registros. A Classificação busca a correlação entre um registro e uma de suas possíveis classes (CAMILO; SILVA, 2009; SOUZA, 2016). Por exemplo, os registros da base de dados utilizada neste trabalho estão categorizados em 7 classes distintas, onde cada classe corresponde à uma possível classificação para uma célula do colo do útero. Os algoritmos classificadores analisam os registros e identificam padrões, e assim se tornam capazes de classificar novas células, de acordo com o que foi previamente aprendido. A Classificação também pode ser usada para: Determinar quando uma transação de cartão de crédito pode ser uma fraude; Identificar em uma escola, qual a turma mais indicada para um determinado aluno; Identificar quando uma pessoa pode ser uma ameaça para a segurança.

- **Regressão ou Estimação**: Tarefa bastante similar à Classificação. É usada quando o registro é identificado por um valor numérico ao invés de um valor categórico. Pode ser usada, por exemplo, para estimar a pressão ideal de um paciente de acordo com fatores como idade, sexo e massa corporal (CAMILO; SILVA, 2009; CHAGAS, 2016).

3.2.2 Aprendizado Não Supervisionado

Os algoritmos de Aprendizado Não Supervisionado são voltados para tarefas de descrição, sendo utilizados para detectar padrões e tendências nos dados. Ao contrário do Aprendizado Supervisionado, os registros dos dados utilizados não necessitam serem categorizados por um atributo classe. São utilizados em tarefas de Agrupamento (*Clustering*) e Associação (*Association*) (CAMILO; SILVA, 2009; SANTANA, 2018; SOUZA, 2015).

- **Agrupamento**: É uma tarefa que busca identificar e aproximar registros que apresentam similaridades entre si, formando grupos (clusters) com características semelhantes e afastando-os de outros grupos ou registros diferentes (CHAGAS, 2016; SOUZA, 2016). A tarefa não é utilizada com o objetivo de classificar, estimar ou predizer o valor de uma variável, servindo apenas para identificar grupos de dados similares, conforme mostrado na Figura 9 (CAMILO; SILVA, 2009).

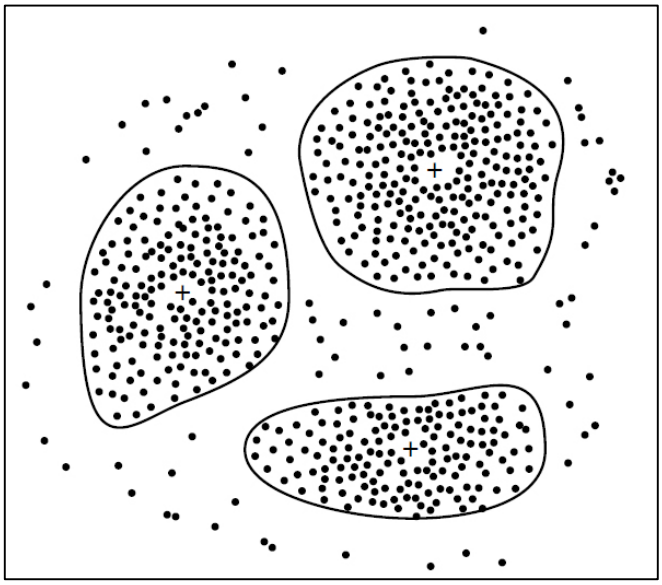

Figura 9 – Registros agrupados em 3 clusters.
Fonte: CAMILO; SILVA, 2009.

- **Associação**: A tarefa de associação consiste em identificar relações entre os atributos dos registros, determinando o quanto a presença de um certo conjunto de atributos implica na presença de outro grupo de atributos em um mesmo registro. Considerando, por exemplo, uma lista de compras, as regras de associação podem identificar se a compra de um produto, como bolo, influencia na compra de outro produto, como refrigerante (CAMILO; SILVA, 2009; SOUZA, I., 2016; SOUZA, M., 2015).

3.3 ALGORITMOS DE APRENDIZADO DE MÁQUINA NA SAÚDE

Atualmente, muitos trabalhos vêm sendo desenvolvidos com o objetivo de solucionar problemas na área da medicina através do Aprendizado de Máquina, pois esta é uma área da Inteligência Artificial capaz de analisar conjuntos complexos de dados médicos. Além disso, possui grande potencial para diagnóstico e predições em diversos cenários clínicos (RAMESH et al., 2004).

Baluz e Santos (2011) fizeram uso de diferentes algoritmos de aprendizado de máquina para classificar exames de cardiotocografia, um método que permite o estudo da frequência cardíaca fetal, bem como suas variações de acordo com a atividade uterina. Eles realizaram duas tarefas de classificação, onde a primeira caracterizou o exame de acordo com o estado fetal, e a segunda de acordo com o padrão morfológico. Os resultados obtidos foram bastante promissores, com o algoritmo *Random Forest* obtendo os melhores resultados: 94,9% de acerto para a primeira tarefa, e 87,3% de acerto para a segunda tarefa.

Já Veras (2015) se utilizou de uma Rede Neural Artificial (RNA), do tipo *Multilayer Perceptron* (MLP), para classificar calcificações mamárias observadas em exames mamográficos de acordo com o sistema *Breast Image Reporting and Data System* (BI-RADS). A RNA MLP foi treinada utilizando o algoritmo *Backpropagation*, obtendo taxas de acerto de até 91% durante a fase de treinamento.

Chagas (2016) realizou um trabalho similar ao de Veras (2015), e inovou ao utilizar um algoritmo de Árvore de Decisão (J48) para realizar a categorização do parâmetro que possui maior influência na descoberta do padrão BI-RADS em exames de mamografia. O treinamento da RNA MLP utilizando o conjunto de dados existentes, com a duplicação do parâmetro obtido por meio da Árvore de Decisão, resultou em uma taxa de treinamento e validação com convergência de 100% de acerto.

Sharma e Gupta (2016) utilizaram algoritmos de Árvores de Decisão, presentes na ferramenta WEKA, para analisar dados extraídos de tumores cervicais e identificar os estágios do câncer do colo do útero. Eles obtiveram uma taxa de acerto máxima de aproximadamente 38%. Este resultado provavelmente se deve a grande complexidade da tarefa a ser realizada.

Vlahou et al. (2003) realizaram um estudo baseado em Árvores de Decisão para discriminar câncer de ovário, doenças benignas e normalidade em mulheres. Para isso eles analisaram os perfis de 139 pacientes, incluindo mulheres saudáveis, ou com câncer de ovário ou com doenças pélvicas benignas, atráves do software *Biomarker Patterns Software* (BPS).

Os melhores resultados foram atingidos através do método de validação cruzada, com uma acurácia de 81,5% nos resultados.

Keerthana (2017) criou um sistema de predição de doença cardíaca ainda em estágio inicial. Para isso, utilizou algoritmos de Aprendizado de Máquina presentes na ferramenta WEKA, como *Naive Bayes,* J48, e *Random Forest*. A ferramenta foi integrada ao sistema através de uma Interface de Programação de Aplicações (API). A base de dados utilizada foi a *Cleveland Heart Disease Database,* que é uma base disponível publicamente e amplamente utilizada em trabalhos na área de predição de doenças cardíacas. O melhor resultado foi alcançado com a utilização do algoritmo *Naive Bayes*, que apresentou uma acurácia de 83,33%.

Por fim, Dangare e Apte (2012) realizaram uma análise sobre a mesma base de dados utilizada por Keerthana (2017), porém, eles incluíram dois atributos (Obesidade e Hábito de fumar) com o objetivo de conseguirem melhores resultados nos experimentos. As técnicas de classificação selecionadas por eles foram as Redes Neurais, Árvores de Decisão e a Classificação Probabilística, que apresentaram acurácias de 100%, 99,62% e 90,74% respectivamente.

4 MATERIAIS E MÉTODOS

A pesquisa consistiu em analisar o desempenho de algoritmos de aprendizado de máquina, aplicados a uma base de dados pré-existente obtida por meio do exame citológico preventivo para a tarefa de detecção de anomalias em células do colo do útero. Para atingir o objetivo principal do trabalho, foi necessária uma pesquisa bibliográfica, onde artigos, livros, monografias e teses foram consultadas a fim de compreender as áreas presentes no estudo.

Para a realização da pesquisa, foi também preciso selecionar uma base de dados adequada ao contexto dos experimentos e realizar o pré-processamento desses dados com o auxílio da ferramenta WEKA. Para a realização dos experimentos, foram selecionados 4 algoritmos de classificação e foram definidas algumas métricas para avaliação dos resultados.

O trabalho, ao longo do desenvolvimento, seguiu de forma sequencial as etapas descritas visualmente na Figura 10.

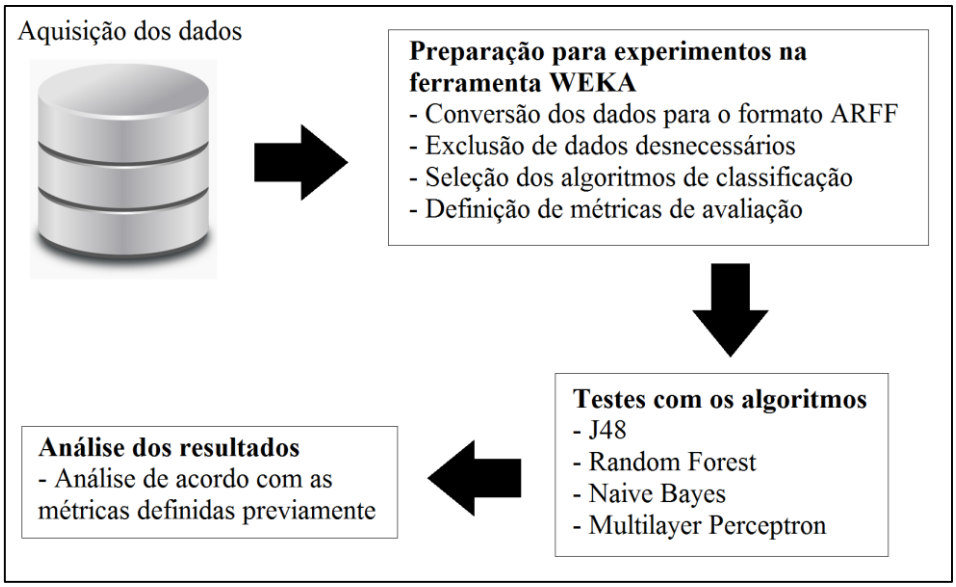

Figura 10 – Etapas aplicadas na metodologia.
Fonte: O autor.

4.1 BASE DE DADOS

A base de dados selecionada para este experimento foi desenvolvida pelo Hospital da Universidade de Herlev, na Dinamarca, no ano de 2005, e se encontra disponível para uso público no domínio <http://mde-lab.aegean.gr/index.php/downloads> (Acesso em: 28/02/2018). Os dados estão dispostos originalmente em forma de tabela no formato XLS, que é um formato implementado em versões mais antigas do software Microsoft Excel.

A base possui 917 registros (instâncias), onde cada um corresponde a informações de uma única célula colhida através do exame de Papanicolaou. As colunas de cada registro representam diferentes atributos das células, como área, posição e claridade, tanto do núcleo quanto do citoplasma. A última coluna representa o atributo classe, responsável por categorizar os registros. A base e suas respectivas classes estão organizadas conforme mostra a Tabela 3.

Tabela 3 – Distribuição dos 917 registros na base de dados.

Classe	Categoria	Tipo de célula	Quantidade	Subtotal
1	Normal	Escamosa superficial	74	
2	Normal	Escamosa intermediária	70	
3	Normal	Colunar	98	242 normais
4	Anormal	Displasia leve	182	
5	Anormal	Displasia moderada	146	
6	Anormal	Displasia severa	197	
7	Anormal	Carcinoma in situ	150	675 anormais

Fonte: Adaptada de JANTZEN et al., 2005.

Visto que as classes de 1 a 3 são categorizadas como células normais e as classes de 4 a 7 como anormais, uma base secundária foi preparada, com a quantidade de classes reduzida a apenas duas: Normal, representada pelo número “1”; e Anormal, representada pelo número “2”. A redução do número de classes foi feita com o intuito de alcançar melhores resultados nos experimentos, visto que a base conta com uma quantidade relativamente baixa de instâncias. A quantidade de instâncias por classe sofreu um aumento considerável, o que deve melhorar a qualidade da fase de treinamento e aumentar a precisão das classificações.

O primeiro atributo presente na base é do tipo nominal e serve apenas para referenciar as imagens das células, sendo assim desconsiderado para os experimentos. Os demais atributos, referentes às características do Núcleo (N) e Citoplasma (C) de cada célula, são numéricos e permanecem idênticos entre as duas bases, conforme listado na Tabela 4.

Tabela 4 – Atributos utilizados para a descrição completa de cada célula.

Atributos descritivos das células		
1. Área (N)	8. Alongamento (N)	15. Perímetro (C)
2. Área (C)	9. Redondeza (N)	16. Posição relativa (N)
3. Proporção N/C	10. Menor diâmetro (C)	17. Maxima (N)
4. Claridade (N)	11. Maior diâmetro (C)	18. Minima (N)
5. Claridade (C)	12. Alongamento (C)	19. Maxima (C)
6. Menor diâmetro (N)	13. Redondeza (C)	20. Minima (C)
7. Maior diâmetro (N)	14. Perímetro (N)	

Fonte: Adaptada de NORUP, 2005.

4.2 FERRAMENTA WEKA

Para a realização dos experimentos, foi utilizada a ferramenta WEKA na versão 3.8.1. A ferramenta foi desenvolvida pela Universidade de Waikato, na Nova Zelândia. Consiste em um software livre, escrito na linguagem Java, que compila diversos algoritmos de aprendizado de máquina em sua interface, podendo realizar tarefas de classificação, regressão, agrupamento e associação. É considerada uma das melhores ferramentas livres, sendo amplamente utilizada na literatura. Seus algoritmos podem ser utilizados tanto na ferramenta como em programas Java, através de uma API. (BALUZ; SANTOS, 2011; CAMILO; SILVA, 2009; SOUZA, 2015).

O software conta com uma interface simples de compreender, ferramentas de visualização e análise dos dados e uma grande lista de algoritmos de aprendizado com parâmetros ajustáveis. A interface da ferramenta, na aba onde os experimentos com tarefas de classificação e regressão são realizados, pode ser observada na Figura 11.

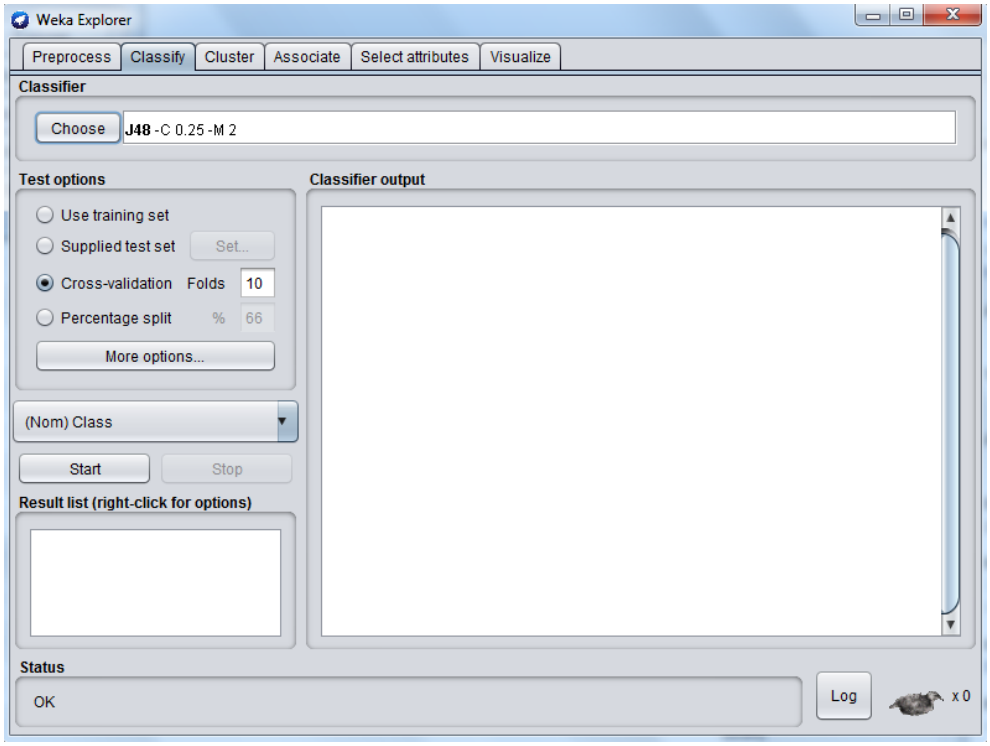

Figura 11 – Interface da ferramenta WEKA 3.8.1.
Fonte: O autor.

A ferramenta trabalha com um formato de arquivo de dados específico, chamado de *Attribute-Relation File Format* (ARFF), porém consegue ler dados em formatos como *Comma-Separated Values* (CSV) e *JavaScript Object Notation* (JSON).

A fim de adaptar a base de dados original para o formato ARFF, primeiramente foi realizada a conversão do formato original XLS para o formato CSV, para que pudesse ser feita a leitura dos dados na ferramenta. Para isso foi utilizado o conversor online Convertio (Disponível em: <https://convertio.co/pt/xls-csv/>. Acesso em: 18 jun. 2018). Após a conversão, o arquivo em CSV foi importado para a ferramenta WEKA e salvo no formato ARFF, para que pudesse ser manipulado.

A base de dados já se apresentava limpa e consistente, sem dados faltantes e com todos os atributos relevantes já expressos, sendo apenas necessária a remoção do atributo nominal que identificava cada linha.

4.3 ALGORITMOS DE CLASSIFICAÇÃO

Para a realização dos experimentos, a tarefa de Classificação, pertencente ao Aprendizado Supervisionado, foi escolhida por ser a tarefa mais adequada à estrutura em que os dados se encontram e também ao propósito da análise, visto que essa é a tarefa mais utilizada para predição de diagnósticos na área médica.

A ferramenta WEKA conta com um grande leque de algoritmos de classificação, dentre os quais foram escolhidos: J48, *Random Forest*, *Naive Bayes* e *Multilayer Perceptron*. Os algoritmos foram selecionados de acordo com sua popularidade na literatura, e também por apresentarem naturezas distintas entre si, com relação aos processos de treinamento e teste.

4.3.1 J48

As Árvores de Decisão são umas das ferramentas mais importantes para descoberta de conhecimento em conjuntos de dados grandes e complexos. O J48 é uma implementação na ferramenta WEKA do algoritmo C4.5, proposto por Ross Quinlan em 1993, que é considerado um dos algoritmos de árvore de decisão mais tradicionais utilizados para classificação. O C4.5 é inspirado no algoritmo *Iterative Dichotomizer* (ID3), também desenvolvido por Quinlan, e ambos utilizam a estratégia “dividir para conquistar”. A estratégia consiste em dividir um problema maior recursivamente em problemas menores. (BALUZ; SANTOS, 2011; JUNIOR, 2016; VIDYA; NASIRA, 2015).

O J48 surgiu da necessidade de recodificar o C4.5, com o propósito de gerar uma árvore de decisão a partir de um conjunto de dados de treino rotulados, a fim de permitir a classificação de instâncias no conjunto de teste (BALUZ; SANTOS, 2011; SOUZA, 2015).

Dado um conjunto de treinamento, cada atributo do conjunto de dados pode ser usado na tomada de decisão, dividindo os dados em subconjuntos menores pertencentes a mesma categoria. O J48 escolhe o atributo com a maior entropia (ganho de informação) e o utiliza para criar um nó de decisão, a fim de quebrá-lo em sub-árvores menores, então o algoritmo repete o mesmo procedimento nessas novas sub-árvores (BALUZ; SANTOS, 2011; SOUZA, 2015).

O atributo com maior ‘‘número” de informação (ocorrências) é escolhido como nó raiz na construção da árvore. As folhas são criadas para cada valor desse atributo e cada uma delas possui um subconjunto de vetores associados pertencente a mesma categoria (SOUZA, 2015).

O algoritmo J48 se mostra adequado para trabalhar com atributos numéricos, nominais e textuais, mesmo possuindo valores ausentes em atributos do conjunto de treinamento (VIDYA; NASIRA, 2015).

Um exemplo de árvore de decisão, gerada pelo algoritmo J48 presente na ferramenta WEKA, pode ser observado na Figura 12. Neste exemplo, a árvore analisa as condições meteorológicas para indicar se um indivíduo deve ou não sair de casa.

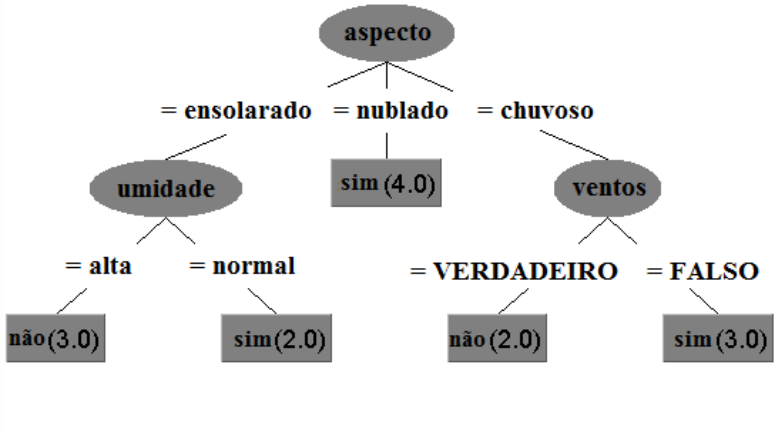

Figura 12 – Exemplo de árvore de decisão gerada pelo algoritmo J48.
Fonte: O autor.

4.3.2 Random Forest

O algoritmo Random Forest, criado por Leo Breiman, funciona através da criação de uma “floresta” de árvores de decisão, onde cada árvore é treinada a partir de um subconjunto, gerado através da divisão aleatória do conjunto de dados de treino (BALUZ; SANTOS, 2011; SOUZA, 2015).

Para a classificação, cada instância passa por todas as árvores da floresta, até chegar em suas classes finais. O mesmo processo é repetido para todas as instâncias. Para eleger uma classe é realizada uma votação, onde cada uma das árvores sinaliza seu voto. A classificação final é obtida através da classe que recebeu o maior número de votos entre todas as árvores da floresta (SARWAR; SHARMA; GUPTA, 2015; SOUZA, 2015).

4.3.3 Naive Bayes

A Classificação Bayesiana utiliza uma técnica estatística (probabilidade condicional) baseada no teorema de Thomas Bayes. O algoritmo Bayesiano, chamado de *Naive Bayes*, é considerado um dos classificadores probabilísticos mais simples, e ainda assim consegue atingir resultados compatíveis com os métodos de árvores de decisão e redes neurais (BALUZ; SANTOS, 2011; CAMILO; SILVA, 2009). Devido a sua simplicidade e eficácia, o algoritmo tem sido amplamente utilizado em diversas tarefas de classificação e seu uso vem sendo considerado uma tendência geral (SANTANA; OLIVEIRA; PACCA, 2014).

O algoritmo *Naive Bayes* trabalha a partir do princípio de que não existem relações de dependência entre os atributos, o que acaba sendo um ponto vantajoso, pois permite ao algoritmo obter boas estimativas de variáveis úteis, como médias e variâncias, mesmo quando se utiliza de pequenos volumes de dados. A classificação das instâncias se dá através da estimação da probabilidade que essa instância possui de pertencer a uma de suas possíveis classes, baseada no produto das probabilidades condicionais individuais para os valores característicos da instância (BALUZ; SANTOS, 2011; CAMILO; SILVA, 2009).

4.3.4 Multilayer Perceptron

As Redes Neurais Artificiais (RNAs) têm sido amplamente utilizadas em trabalhos voltados para a área médica nas últimas duas décadas. Atualmente, a RNA é considerada a técnica de Inteligência Artificial mais popular na medicina (RAMESH et al., 2004).

São inspiradas no funcionamento do sistema nervoso biológico, e são constituídas unidades de processamento interconectadas, denominadas neurônios, capazes de realizar processamento paralelo e distribuído para processamento de dados e representação de conhecimento. Os neurônios estão dispostos em uma ou mais camadas, interligadas por conexões (sinapses) geralmente unidirecionais. Cada conexão possui um peso (peso sináptico), onde esses pesos representam o conhecimento da rede. O processo de aprendizado se dá através

da atualização iterativa dos pesos sinápticos, chegando ao seu término quando atinge um critério pré-estabelecido (ROCHA et al., 2007).

As camadas das RNAs são divididas em três partes (VERAS, 2015):

1. **Camada de entrada:** Responsável pelo recebimento de informações ou características predefinidas do meio externo.
2. **Camadas escondidas, intermediárias, ocultas ou invisíveis:** Composta por neurônios, responsáveis por extrair as características associadas ao processo ou sistema a ser analisado. A maioria dos processos são feitos nessas camadas.
3. **Camada de saída:** Constituída de neurônios, é responsável por apresentar os resultados processados pela rede neural de acordo com os processos derivados das camadas anteriores.

Na Figura 13 pode ser observado um esquema de uma RNA de topologia 4:3:2 (4 neurônios na primeira camada escondida, 3 na segunda camada escondida e 2 na camada de saída), com a característica *feedforward*, onde o fluxo de informações segue em uma única direção. A topologia da rede em cada aplicação depende de diversos fatores, pois cada classe de aplicação possui características e volumes de informações particulares (VERAS, 2015).

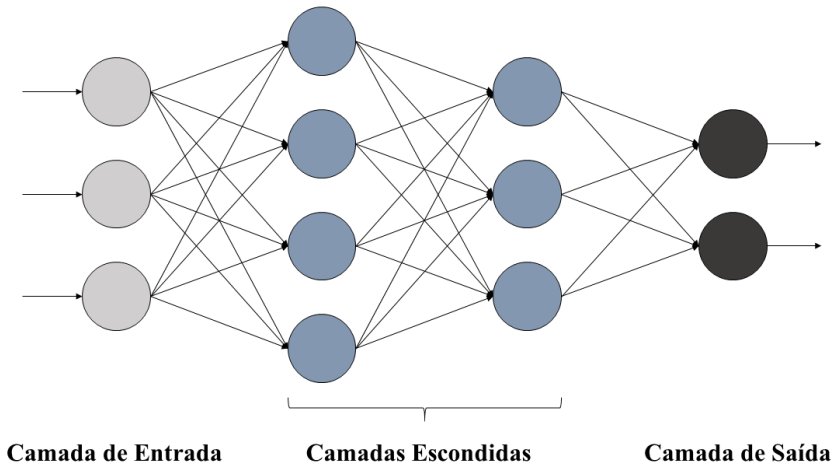

Figura 13 – Exemplo de rede neural artificial *feedforward*.
Fonte: O autor.

Um modelo de RNA bastante utilizado para tarefas de predição é o *Multilayer Perceptron* (MLP) ou Perceptron de múltiplas camadas, cuja arquitetura é composta por uma camada de entrada, uma ou mais camadas intermediárias e uma camada de saída. A ferramenta WEKA utiliza o algoritmo *backpropagation* durante o treinamento da rede, que consiste em um

algoritmo supervisionado que utiliza pares de entrada e saída desejada para atualizar os pesos sinápticos através de um mecanismo de correção de erros (BALUZ; SANTOS, 2011; ROCHA et al., 2007; SOUZA, 2016).

O treinamento se dá em duas etapas: *forward*, onde a rede é percorrida da camada de entrada para a camada de saída, que determina as saídas para um dado padrão de entrada; e *backward*, onde a rede é percorrida da camada de saída para a camada de entrada, que compara as saídas e entradas e atualiza os pesos sinápticos. Propõe também que a atualização dos pesos seja realizada através do método do gradiente descendente, garantindo que a rede caminhe no sentido de reduzir de erros. (BALUZ; SANTOS, 2011; ROCHA et al., 2007).

As RNAs são capazes de aprender padrões complexos de dados, analisar dados não-lineares, lidar com informações imprecisas e generalizar a informação aprendida. Essas capacidades tornam as RNAs ferramentas analíticas muito atrativas para a medicina (RAMESH et al., 2004; VERAS, 2015).

4.4 MÉTRICAS DE AVALIAÇÃO

A fim de avaliar os desempenhos dos algoritmos, foi necessária a definição de métricas de avaliação. As seguintes métricas foram escolhidas com base nos resultados apresentados pela ferramenta WEKA:

- **Taxa de Acerto (Acurácia):** Informa a porcentagem de instâncias classificadas corretamente após a execução do algoritmo no conjunto de dados.
- **Estatística Kappa (K):** Avalia o grau de concordância de uma tarefa de classificação, ou seja, determina em termos de porcentagem o quão correto as instâncias foram classificadas. O índice Kappa estabelece intervalos de valores que resumem o grau de concordância, os quais estão listados na Tabela 5 (SOUZA, 2015).

Tabela 5 – Intervalos de valores do Kappa.

Valor do Kappa	Concordância
0	Ruim
0 – 0,20	Ligeira
0,21 – 0,40	Considerável
0,41 – 0,60	Moderada
0,61 – 0,80	Substancial
0,81 – 1	Excelente

Fonte: SOUZA, 2015.

- **Matriz de Confusão (Tabela de Contingência):** Permite visualizar como as instâncias foram classificadas pelos algoritmos (SOUZA, 2015). Utilizando a matriz da Figura 14 como exemplo, os resultados são apresentados em uma matriz, onde os valores da diagonal principal (cor verde) indicam as instâncias classificadas corretamente. No caso, 8 instâncias de classe "a" e 2 instâncias de classe "b" foram corretamente classificadas. Os valores que não estão contidos na diagonal principal (cor vermelha) indicam as instâncias que foram incorretamente classificadas como pertencentes a outra classe. Nessa matriz, 1 instância de classe "a" foi incorretamente classificada como "b", enquanto 3 instâncias de classe "b" foram classificadas como pertencentes à classe "a".

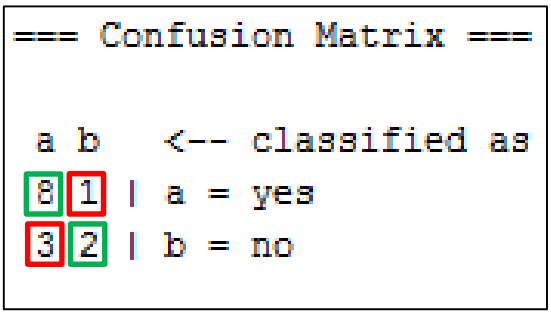

Figura 14 – Exemplo de uma matriz de confusão.
Fonte: O autor.

- **Percentual de Falsos Positivos e Falsos Negativos:** Essa métrica será usada apenas na base de dados secundária, que contém apenas 2 classes. A métrica permite observar a porcentagem de instâncias classificadas como determinada classe, quando na verdade não pertencem a ela. O percentual de falsos positivos na classe de células normais necessita atenção especial, pois, do ponto de vista médico, é pior classificar resultados anormais como normais do que o contrário.

4.5 PREPARAÇÃO DOS EXPERIMENTOS

Dentre as opções de teste disponíveis na ferramenta WEKA, foi selecionada a técnica de validação cruzada *k-fold cross-validation*. Nesse método, a base de dados é dividida em k subconjuntos, ou partições, mutuamente exclusivos (*folds*) e são executadas k sessões de treinamento e teste (BALUZ; SANTOS, 2011; KOHAVI, 1995).

A cada iteração um subconjunto diferente é utilizado para testar o sistema e todos os outros k – 1 subconjuntos são utilizados para treinar o sistema. O resultado final é a média

de classificações corretas nos k subconjuntos de teste. Por exemplo, no caso de *3-fold cross-validation* todo o conjunto de dados é dividido em 3 subconjuntos distintos, aleatórios e mutuamente exclusivos. São realizadas 3 iterações, a cada iteração um subconjunto diferente é utilizado para testar o sistema e os outros 2 são utilizados para treinar o sistema. O resultado final é dado pela média de classificações corretas nos 3 subconjuntos de teste (BALUZ; SANTOS, 2011; FARIAS, 2015). A Figura 15 mostra um esquema para o método *3-fold cross-validation*.

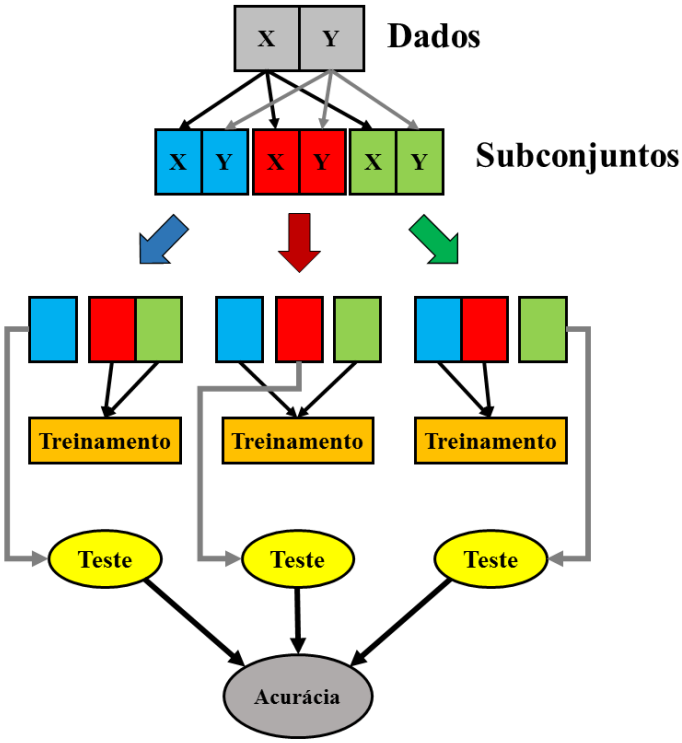

Figura 15 – Esquema para o método *3-fold cross-validation*.
Fonte: Adaptada de SOUZA, 2016.

A execução dos algoritmos nos experimentos se deu com 10 folds, por ser um número bastante utilizado na literatura, e comprovadamente eficiente (JANTZEN et al., 2005; KOHAVI, 1995; NORUP, 2005; SOUZA, 2015). As fases de treinamento individual das 10 partições se utilizam de 90% da base de dados completa, ou seja, uma quantidade relativamente próxima da totalidade dos dados.

As configurações da máquina utilizada para realização dos testes foram: processador Intel® Core™ i5-2410M de 2.30 GHz, sistema operacional Windows 7 Ultimate em sua versão de 32 bits, e memória física (RAM) de 3 GB.

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Os experimentos consistiram na aplicação dos algoritmos de classificação selecionados nas duas bases de dados. O Experimento 1 é caracterizado pelos testes realizados utilizando a base original com 7 classes. Já o Experimento 2 ficou definido como os testes realizados na base secundária, prepara pelo autor, com a quantidade de classes reduzida para 2.

A ferramenta WEKA permite a alteração dos parâmetros dos algoritmos de classificação. Nos experimentos, foram utilizados os parâmetros padrões da versão 3.8.1 da ferramenta. Os resultados foram analisados de acordo com as métricas definidas na seção 4.4.

5.1 COMPARAÇÃO DAS TAXAS DE ACERTO

Após a aplicação dos algoritmos em ambas as bases de dados, os resultados relativos às suas taxas de acerto, listados na Tabela 6, foram extraídos e analisados.

Tabela 6 – Taxas de acerto dos algoritmos selecionados em ambos os experimentos.

Algoritmos	**Taxas de Acerto**	
	Experimento 1	Experimento 2
J48	55,18 %	92,15 %
Random Forest	60,96 %	93,67 %
Naive Bayes	55,40 %	91,71 %
Multilayer Perceptron	63,14 %	93,78 %

Fonte: O autor.

As taxas de acerto no Experimento 1 variam de aproximadamente 55,18% a 63,14%, com o algoritmo *Multilayer Perceptron* obtendo o melhor resultado. Já no Experimento 2, as taxas variam de aproximadamente 91,71% a 93,78%, com o *Random Forest* e o *Multilayer Perceptron* obtendo as melhores taxas. Os melhores resultados já eram esperados no Experimento 2, devido a quantidade reduzida de classes.

Dentre os algoritmos testados, o *Multilayer Perceptron* foi o que obteve o maior tempo de execução. Baluz e Santos (2011) observaram que, em aplicações pesadas, a rede neural apresenta um grande gasto de tempo na etapa de aprendizado, pois o *backpropagation*, combinado ao método de validação cruzada *k-fold cross-validation* muitas vezes demora para alcançar convergência para as saídas desejadas.

5.2 COMPARAÇÃO DE ACORDO COM A ESTATÍSTICA KAPPA

Os valores do índice Kappa (K) indicam o grau de concordância das tarefas de classificação, e serviu como métrica de avaliação das classificações feitas pelos algoritmos. Os índices mais próximos de 1 indicam que os algoritmos obtiveram ótimas concordâncias em suas classificações. Já os índices mais próximos de 0 indicam que a confiabilidade das classificações é baixa. Os valores do K para cada algoritmo, em ambos os experimentos, estão listados na Tabela 7.

Tabela 7 – Valores do Kappa para cada algoritmo em ambos os experimentos.

Algoritmos	**Valor do Kappa**	
	Experimento 1	Experimento 2
J48	0,4644	0,7946
Random Forest	0,5337	0,8328
Naive Bayes	0,4757	0,7715
Multilayer Perceptron	0,5596	0,8406

Fonte: O autor.

Os valores do K no Experimento 1 variam de aproximadamente 0,46 a 0,56. Isso quer dizer que os algoritmos obtiveram concordâncias moderadas em suas classificações, de acordo com os intervalos definidos pela Estatística Kappa. Os valores do Experimento 2 variam de aproximadamente 0,77 a 0,84, com os algoritmos J48 e *Naive Bayes* alcançando concordâncias substanciais, enquanto os algoritmos *Random Forest* e *Multilayer Perceptron* obtiveram concordâncias excelentes.

Os resultados indicam que as classificações realizadas nos experimentos, principalmente no Experimento 2, possuem alta confiabilidade, pois as taxas de acerto dos algoritmos, além de serem relativamente altas, são compatíveis com seus índices Kappa.

5.3 ANÁLISE DAS MATRIZES DE CONFUSÃO

As matrizes de confusão permitem identificar como as instâncias dos dados foram classificadas durante os testes, o que torna essa métrica suficiente para a extração de todas as outras. A análise dessas classificações se torna de grande importância, especialmente para o

Experimento 1, pois permite a compreensão das taxas de acerto e da razão de seus valores. Permite também analisar como as classes interagem entre si, e se há confusão entre elas no decorrer dos testes. Na Figura 16, estão presentes as 4 matrizes de confusão geradas no Experimento 1.

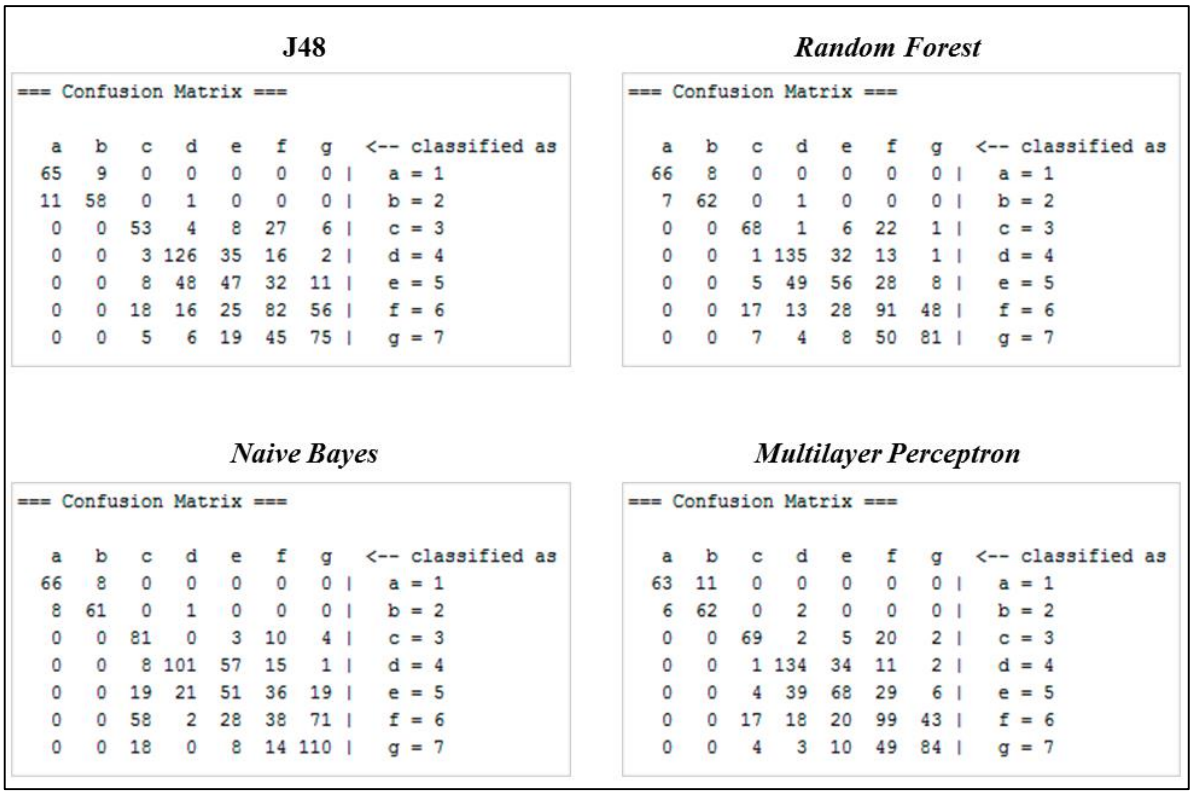

Figura 16 – Matrizes de confusão geradas no Experimento 1.
Fonte: O autor.

A distribuição das instâncias para os 4 algoritmos se deu de forma bastante similar. Os algoritmos conseguiram classificar as classes 1 (células escamosas superficiais) e 2 (células escamosas intermediárias) com precisões elevadas, visto que a maioria das instâncias dessas duas classes se encontra na diagonal principal. Esses resultados indicam que essas células possuem diferenças claras entre si, como a proporção N/C por exemplo, que apresentam valores bem distintos entre os dois tipos de células.

As classificações mais dispersas no intervalo de classes 4, 5, 6 e 7 indicam que ocorre uma transição suave entre esses tipos de células (displasias leve, moderada e severa, e carcinoma in situ). Visto que as características são similares entre as classes vizinhas desse intervalo, os algoritmos apresentaram certa dificuldade em diferenciá-las com clareza.

A instâncias de classe 3 (células colunares), no geral, foram classificadas de maneira satisfatória. Porém, em diversas ocasiões, essas instâncias foram confundidas como

pertencentes à classe 6 (displasia severa). Esse comportamento provavelmente se deve às diferenças existentes entre as estruturas das células colunares da endocérvice e das células escamosas da ectocérvice. Os atributos das células colunares apresentam características similares aos de células escamosas displásicas, gerando assim a confusão nas classificações.

Apesar de as classes 4, 5, 6 e 7 terem obtido classificações um tanto inconsistentes, sendo algumas vezes confundidas entre si no Experimento 1, elas pertencem à mesma categoria (anormal) e são unidas em uma só classe na base utilizada durante o Experimento 2. Essa união implicou na eliminação da maior parte dessas inconsistências observadas nas matrizes de confusão do Experimento 1, o que acabou sendo responsável pelo aumento considerável das taxas de acerto. As matrizes de confusão do Experimento 2 podem ser observadas na Figura 17.

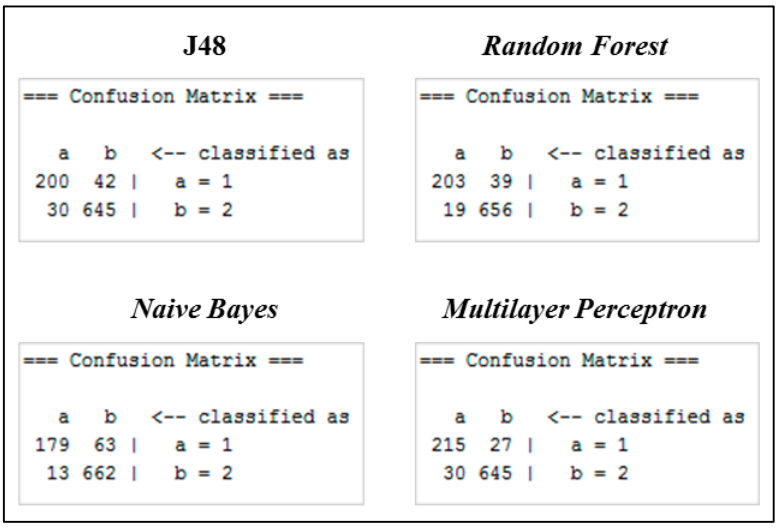

Figura 17 – Matrizes de confusão geradas no Experimento 2**.**
Fonte: O autor.

De acordo com as análises realizadas nas matrizes de confusão do Experimento 1, pôde-se concluir que as classificações incorretas no Experimento 2 estão fortemente ligadas ao problema observado na classificação das células colunares (classe 3), visto que é nessa classe onde ocorrem as maiores confusões entre células normais (1, 2 e 3) e anormais (4, 5, 6 e 7).

5.4 ANÁLISE DAS TAXAS DE FALSOS POSITIVOS

Visto que as classes da base de dados do Experimento 2 estão definidas como normal e anormal, foi possível analisar as taxas de falsos positivos. Essas taxas são de grande importância, visto que o experimento realizado neste trabalho está lidando diretamente com a

classificação de exames médicos e, consequentemente, com a saúde de indivíduos. Dito isso, é de grande importância que os algoritmos classificadores atinjam as menores taxas possíveis.

Os falsos positivos são definidos como os dados classificados como determinada classe, quando na verdade pertencem à outra. Um diagnóstico falso positivo para anormalidade, apesar de futuramente causar incômodo ao paciente, não põe em risco sua vida, ao contrário do diagnóstico falso positivo para normalidade, que indica ao paciente que sua saúde está em estado normal, quando na verdade há presença de alguma patologia.

Na matriz de confusão, os valores presentes na diagonal principal, ou seja, as instâncias classificadas corretamente, são considerados os verdadeiros positivos (VP). Já os valores que não estão contidos na diagonal principal são chamados de falsos positivos (FP), indicando as classificações incorretas. Essas relações estão demonstradas na Tabela 8.

Tabela 8 – Definições dos valores nas matrizes de confusão.

Classe Estimada		**Classes**
A	**B**	
VP (A)	FP (B)	**A**
FP (A)	VP (B)	**B**

Fonte: O autor.

As taxas de falsos positivos dos algoritmos aplicados no Experimento 2, para ambas as classes, podem ser observadas na Tabela 9.

Tabela 9 – Taxas de falsos positivos no Experimento 2.

Algoritmos	**Taxas de Falsos Positivos (Experimento 2)**	
	Classe 1 (Normal)	Classe 2 (Anormal)
J48	4,4%	17,4%
Random Forest	2,8%	16,1%
Naive Bayes	1,9%	26,0%
Multilayer Perceptron	4,4%	11,2%

Fonte: O autor.

No geral, as taxas de falsos positivos para normalidade (TFPN) foram relativamente baixas. Essa é uma característica bastante desejada, pois indica que poucas células anormais foram classificadas como normais. Já as taxas de falsos positivos para anormalidade (TFPA) foram mais elevadas, o que mostra que os algoritmos classificaram uma quantidade considerável de células normais como sendo anormais. O ideal é que as TFPAs sejam baixas, porém, seus valores não interferem de forma nociva na saúde de indivíduos.

O algoritmo *Naive Bayes* obteve a menor TFPN, com 1,9% de classificações incorretas em normais, porém, obteve a maior TFPA, com 26% de classificações incorretas de células anormais. Os algoritmos *Random Forest* e *Multilayer Perceptron* foram considerados os algoritmos com melhor desempenho em relação às taxas, pois ambos obtiveram taxas baixas e equilibradas entre si.

5.5 SELEÇÃO DE ATRIBUTOS E COMPARAÇÃO DE RESULTADOS

Em tarefas de classificação, os dados utilizados muitas vezes apresentam uma grande quantidade de atributos, dos quais apenas uma pequena quantidade pode estar relacionada ao atributo classe. A seleção de atributos relevantes nas bases de dados é importante pois permite melhorar a performance das classificações, facilitando a identificação dos padrões e reduzindo o tempo utilizado para execução das tarefas (GUYON; ELISSEEFF, 2003; KIRA; RENDELL, 1992).

Chandrashekar e Sahin (2014) observaram que, no geral, a seleção de atributos sempre traz benefícios, promovendo um melhor entendimento dos dados, aprimorando o modelo de classificação e identificando atributos irrelevantes.

Partindo desses princípios, foi realizada uma análise do desempenho de um dos algoritmos testados anteriormente, o *Multilayer Perceptron*. Esse algoritmo foi selecionado por ter obtido os melhores resultados nos testes, junto ao *Random Forest*, e por ser amplamente utilizado em problemas de natureza médica.

O algoritmo foi testado através do meta-algoritmo *Attribute Selected Classifier*, presente na ferramenta WEKA. O meta-algoritmo realiza a diminuição da quantidade de dados de treino e teste, através da seleção de atributos. Após a seleção, os dados são passados para o algoritmo classificador.

A avaliação dos atributos foi realizada pelo avaliador *Correlation Attribute Eval*, que avalia a correlação de Pearson entre cada atributo e a classe. O coeficiente de Pearson assume valores que variam de 1 a -1. Os atributos cujos coeficientes estejam próximos de 1 ou

-1 possuem alta correlação positiva ou negativa com a classe, enquanto os atributos com valores próximos de 0 possuem baixa correlação (BROWNLEE, 2016). Os atributos foram selecionados através do método *Ranker*, que escolhe os atributos mais relevantes de acordo com os resultados obtidos através do avaliador.

O teste foi realizado com a base de 2 classes (Experimento 2). A comparação dos resultados, com e sem seleção de atributos, pode ser observada na Tabela 10.

Tabela 10 – Resultados do teste com e sem seleção de atributos.

Métricas		**Resultados Multilayer Perceptron (Experimento 2)**	
		Sem Seleção	Com Seleção
Taxa de Acerto		93,78 %	94,44 %
Índice Kappa		0,8406	0,8563
Taxa de Falsos Positivos	Normal	4,4%	3,6%
	Anormal	11,2%	11,2%

Fonte: O autor.

Após a realização do experimento, foi observado um leve aumento da taxa de acerto e do índice Kappa, porém, o principal resultado foi a diminuição da taxa de falsos positivos para normalidade, que ocorreu sem que a taxa de falsos positivos para anormalidade sofresse aumento. A diminuição foi de apenas 0,8%, mas qualquer redução dessa taxa possui grande valor, devido à sua importância.

Os resultados finais obtidos nesse último experimento foram satisfatórios, entretanto não superaram os obtidos no trabalho de NORUP (2005). A Tabela 11 demonstra um comparativo entre os melhores resultados obtidos pelos dois trabalhos.

Tabela 11 – Comparação entre os resultados obtidos e os de NORUP (2005).

Métricas		**Melhores Resultados**	
		Autor	NORUP (2005)
Taxa de Acerto		94,44 %	94,87 %
Taxa de Falsos Positivos	Normal	3,6%	4,3%
	Anormal	11,2%	7,4%

Fonte: O autor.

Os melhores resultados no trabalho de NORUP (2005) foram obtidos com a utilização de um algoritmo de classificação autoral, entitulado *Nearest Class gravity Center* (NCC). O autor o descreve como sendo um método simples e sem sensibilidade à predominância de uma das classes, o que justifica o bom desempenho do algoritmo nessa base de dados onde a classe Anormal apresenta certa predominância.

Neste trabalho, foi obtida uma taxa de acerto muito próxima à taxa obtida por NORUP (2005), e também foi alcançada uma taxa de falsos positivos para normalidade levemente menor. No trabalho de NORUP (2005) as taxas de falsos positivos possuem um equilíbrio maior entre seus valores, o que é uma característica desejada nesses tipos de experimentos.

De acordo com essa comparação, fica evidente o potencial da ferramenta WEKA que, apesar de ser uma ferramenta simples de operar, é capaz de obter ótimos resultados em tarefas de classificação.

6 CONSIDERAÇÕES FINAIS

O Exame de Papanicolaou é o método mais simples e eficaz para prevenção do câncer do colo do útero, permitindo que a doença seja detectada precocemente e que as chances de cura sejam muito elevadas. A classificação correta desse exame é de extrema importância para todas as mulheres que possuem risco de contraírem a doença. Dessa forma, um método para reduzir o erro humano no momento da interpretação dos exames é a utilização da análise computacional e de técnicas de aprendizado de máquina.

O objetivo desta pesquisa foi analisar o desempenho de algoritmos de aprendizado de máquina, em uma base de dados pré-existente, para a tarefa de detecção de anomalias em células do colo do útero. Para isso, foi utilizada uma base de dados que contém dados descritivos de células normais e anormais colhidas através do Exame de Papanicolaou.

A base foi analisada através da ferramenta de mineração de dados WEKA, com a utilização da técnica de classificação e dos algoritmos: J48; *Random Forest*; *Naive Bayes*; e *Multilayer Perceptron*. Os testes foram realizados através da validação cruzada *k-fold cross-validation*. Os desempenhos dos algoritmos foram analisados por meio da verificação das taxas de acerto, da estatística Kappa, das matrizes de confusão e das taxas de falsos positivos.

Foram realizados dois experimentos. No primeiro experimento, os algoritmos foram aplicados na base de dados original, que possui 7 classes distintas, onde cada classe corresponde à uma possível classificação do Exame de Papanicolaou. Para o segundo experimento, foi preparada uma base de dados que contém apenas duas classes, onde as classes da base original foram resumidas à dois possíveis resultados: Normal; e Anormal.

O algoritmo de rede neural *Multilayer Perceptron* obteve a melhor taxa de acerto e índice Kappa no segundo experimento, de 93,78 % e 0,8406, respectivamente, indicando que as classificações obtiveram uma excelente concordância. O *Multilayer Perceptron* também obteve a menor taxa de falsos positivos para anormalidade (11,2%), enquanto que o algoritmo *Naive Bayes* obteve a menor taxa de falsos positivos para normalidade (1,9%).

Afim de alcançar resultados ainda melhores, foi realizado um teste com algoritmo *Multilayer Perceptron*, onde a base de dados passou previamente por um processo de seleção de atributos, onde os atributos com maior correlação com a classe alvo foram analisados. As novas taxas de acerto e índice Kappa do algoritmo foram de 94,44 % e 0,8563, respectivamente. O principal resultado desse novo teste foi a redução da taxa de falsos positivos para normalidade, a qual passou de 4,4% para 3,6%.

Os resultados apresentaram um grande potencial dos algoritmos de aprendizado de máquina na análise de dados médicos de células do colo do útero. Apesar de não terem sido superiores aos que foram obtidos em trabalhos anteriores, eles atingiram desempenhos similares com a utilização de métodos simples. Sendo assim, apresentam-se as seguintes sugestões para possíveis trabalhos futuros:

- *Realizar os experimentos em uma base de dados maior* – A base de dados utilizada no experimento não conta com o número ideal de registros, principalmente quando se leva em consideração a quantidade de células obtidas através de um único exame preventivo.
- *Testar outras configurações de parâmetros nos algoritmos* – Neste trabalho foram usadas as configurações padrão da ferramenta utilizada. A adaptação desses parâmetros, de acordo com o problema, pode trazer melhores resultados.
- *Avaliar mais a fundo a seleção de atributos* – A seleção de atributos na base de dados mostrou grande potencial para a melhoria dos resultados. A utilização de outros avaliadores e seletores pode se mostrar mais eficaz em diversas situações.
- *Avaliar o desempenho dos algoritmos em outras bases de dados da área da saúde* – As possibilidades de aplicação do Aprendizado de Máquina na saúde são inúmeras. Dessa forma, outras bases de dados médicos podem ser utilizadas como materiais de estudo.

REFERÊNCIAS

ACS. American Cancer Society. **What is cervical cancer?** Disponível em: <https://www.cancer.org/cancer/cervical-cancer/about/what-is-cervical-cancer.html#written_by>. Acesso em: 13 jun. 2018a.

ACS. American Cancer Society. **Signs and Symptoms of Cervical Cancer**. Disponível em: <https://www.cancer.org/cancer/cervical-cancer/detection-diagnosis-staging/signs-symptoms.html>. Acesso em: 13 jun. 2018b.

AMO, S. DE. **Técnicas de mineração de dados**. Disponível em: <http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf>. Acesso em: 6 jul. 2018.

ANDRADE, J. M. et al. Rastreamento, Diagnóstico e Tratamento do Carcinoma do Colo do Útero. **Projeto Diretrizes Associação Médica Brasileira e Conselho Federal de Medicina**, p. 1–18, 2001.

ARAÚJO, F. L. A. DE. **A prevenção e controle do câncer do colo de útero: relato de experiência**. [s.l.] Universidade Estadual da Paraíba, 2017.

ARBYN, M. et al. European guidelines for quality assurance in cervical cancer screening. Second edition-summary document. **Annals of Oncology**, v. 21, n. 3, p. 448–458, 2010.

BALUZ, R. A. R. S.; SANTOS, C. N. DOS. Applying machine learning approaches to assess cardiotocography exams. **6th Iberian Conference on Information Systems and Technologies (CISTI 2011)**, 2011.

BEZERRA, S. J. S. et al. Perfil de mulheres portadoras de lesões cervicais por HPV quanto aos fatores de risco para câncer de colo uterino. **DST – J bras Doenças Sex Transm**, v. 17, n. 2, p. 143–148, 2005.

BRASIL. Ministério da Saúde. **Falando sobre câncer do colo do útero**. Rio de Janeiro - RJ: Instituto Nacional do Câncer, 2002.

BRENNA, S. M. F. et al. Conhecimento, atitude e prática do exame de Papanicolaou em mulheres com câncer de colo uterino. **Cad. Saúde Pública, Rio de Janeiro**, v. 17, n. 4, p. 909–914, 2001.

BROWNLEE, J. **How to Perform Feature Selection With Machine Learning Data in Weka**. Disponível em: <https://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/>. Acesso em: 27 jun. 2018.

CAMILO, C. O.; SILVA, J. C. DA. **Mineração de Dados: Conceitos, tarefas, métodos e ferramentasUniversidade Federal de Goiás (UFC)**. [s.l: s.n.].

CHAGAS, P. V. C. DAS. **Aplicação de técnicas de mineração de dados para otimização de classificação de exames mamográficos**. [s.l.] Universidade Estadual do Piauí, 2016.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers and Electrical Engineering**, v. 40, n. 1, p. 16–28, 2014.

CRUZ, J. A.; WISHART, D. S. Applications of machine learning in cancer prediction and prognosis. **Cancer Informatics**, v. 2, p. 59–77, 2006.

DANGARE, C. S.; APTE, S. S. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. **International Journal of Computer Applications**, v. 47, n. 10, p. 44–48, 2012.

DEGRUY, K. B. Healthcare Applications of Knowledge Discovery in Databases. **Journal of Healthcare Information Management**, v. 14, n. 2, p. 59–69, 2000.

FARIAS, A. M. DE O. **Utilização de redes neurais artificiais para otimização do processo de produção de filmes finos de ftalocianina (FeTSPc)**. [s.l.] Universidade Estadual do Piauí, 2015.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37, 1996.

FERNANDES, K.; CARDOSO, J. S.; FERNANDES, J. Transfer learning with partial observability applied to cervical cancer screening. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 10255 LNCS, p. 243–250, 2017.

FERRO, M.; LEE, H. D. O Processo de KDD – Knowledge Discovery in Database para Aplicações na Medicina. **Semana de Informática de Cascavel (SEMINC)**, p. 1–6, 2001.

FILHO, L. DE A. F. **O exame papanicolau e o diagnostico das lesões invasoras do colo de utero**. [s.l.] Universidade Paulista, 2011.

GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. **Journal of Machine Learning Research (JMLR)**, v. 3, n. 3, p. 1157–1182, 2003.

INCA. Instituto Nacional do Câncer. **Tipos de Câncer: Colo do Útero**. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/colo_utero/definicao>. Acesso em: 1 jan. 2018.

JANTZEN, J. et al. **Pap-smear Benchmark Data For Pattern Classification**. Proceedings of NiSIS 2005: Nature inspired Smart Information Systems (NiSIS). **Anais**...2005

JUNIOR, S. B. **Tutorial : Arvore de Decisão com Weka para a classificação de carne suína**, 2016.

KEERTHANA, T. K. Heart Disease Prediction System using Data Mining Method. **International Journal of Engineering Trends and Technology**, v. 47, n. 6, p. 361–363, 25 maio 2017.

KIRA, K.; RENDELL, L. A. **A Practical Approach to Feature Selection**. [s.l.] Morgan Kaufmann Publishers, Inc., 1992.

KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. **Appears in the International Joint Conference on Articial Intelligence (IJCAI)**, v. 5, p. 1–7, 1995.

KURNIAWATI, Y. E.; PERMANASARI, A. E.; FAUZIATI, S. **Comparative study on data mining classification methods for cervical cancer prediction using pap smear results**. 2016 1st International Conference on Biomedical Engineering (IBIOMED). **Anais**...IEEE, out. 2016.

LEAL, E. A. S. et al. Lesões precursoras do câncer de colo em mulheres adolescentes e adultas jovens do município de Rio Branco - Acre. **Revista Brasileira de Ginecologia e Obstetrícia**, v. 25, n. 2, p. 81–86, 2003.

LINARD, A. G.; SILVA, F. A. D. E; SILVA, R. M. DA. Mulheres submetidas a tratamento para câncer de colo uterino -percepção de como enfrentam a realidade*. **Revista Brasileira de Cancerologia**, v. 48, n. 4, p. 493–498, 2002.

MEHTA, V.; VASANTH, V.; BALACHANDRAN, C. Pap smear. **Indian Journal of Dermatology, Venereology and Leprology**, v. 75, n. 2, p. 214–216, 2009.

MELO, S. C. C. S. DE et al. Alterações citopatológicas e fatores de risco para a ocorrência do câncer de colo uterino. **Revista Gaúcha de Enfermagem (Online)**, v. 30, n. 4, p. 602–608, 2009.

NORUP, J. **Classification of Pap-smear data by transductive neuro-fuzzy methods**. [s.l.] Technical University of Denmark, 2005.

PEREIRA, R. T.; CHAMORRO, M. C. Y.; ROMERO, A. C. Detecting Survival Patterns in Women with Invasive Cervical Cancer with Decision Trees. In: **Ibero-American Conference on Artificial Intelligence**. [s.l: s.n.]. p. 130–139.

PINHEIRO, P. **Exame Papanicolau**. Disponível em: <https://www.mdsaude.com/2014/09/exame-papanicolau.html>. Acesso em: 18 maio. 2018.

RAMESH, A. N. et al. Artificial intelligence in medicine. **Annals of the Royal College of Surgeons of England**, v. 86, n. 5, p. 334–338, 2004.

ROCHA, E. L. et al. Diagnóstico Automatizado De Doenças No Colo Do Útero Baseado Em Redes Neurais Artificiais E Processamento De Imagens Digitais. p. 10, 2007.

SANTANA, J. E.; OLIVEIRA, M. C.; PACCA, H. L. L. **Aprendizagem De Redes Bayesianas Para Prevenção Do Câncer Cervical**. XVI Congresso Brasileiro em Informática em Saúde - CBIS. **Anais**...2014

SANTANA, R. **Tipos de Aprendizado de Máquina e a Historia dos Criadores de Vacas**. Disponível em: <http://minerandodados.com.br/index.php/2018/03/20/tipos-de-aprendizado-de-maquina/>. Acesso em: 17 jun. 2018.

SANTOS, M. A. DOS et al. A importância da prevenção do câncer do colo uterino: em pauta o exame de papanicolaou. **Revista Científica de Enfermagem**, v. 4, n. 12, p. 15–20, 2014.

SARWAR, A. et al. Novel benchmark database of digitized and calibrated cervical cells for artificial intelligence based screening of cervical cancer. **Journal of Ambient Intelligence and Humanized Computing**, v. 7, n. 4, p. 593–606, 2016.

SARWAR, A.; SHARMA, V.; GUPTA, R. Hybrid ensemble learning technique for screening of cervical cancer using Papanicolaou smear image analysis. **Personalized Medicine Universe**, v. 4, p. 54–62, 2015.

SELLORS, J. W.; SANKARANARAYANAN, R. **Colposcopia e tratamento da neoplasia intra-epitelial cervical : Manual para principiantes**. Washington, D.C.: Centro Internacional de Pesquisa sobre o Câncer, 2004.

SHARMA, S.; GUPTA, S. Decision tree approach in machine learning for prediction of cervical cancer stages using WEKA. **International Journal of Recent Trends in Engineering & Research**, v. 2, n. 8, p. 74–83, 2016.

SILVA, D. W. DA et al. Cobertura e fatores associados com a realização do exame Papanicolaou em município do Sul do Brasil. **Rev bras ginecol obstet**, v. 28, n. 1, p. 24–31, 2006.

SOUZA, I. T. DE. **Aplicação de uma técnica de mineração de dados para elaboração de um modelo de previsão do índice pluviométrico no litoral do estado do piauí**. [s.l.] Universidade Estadual do Piauí, 2016.

SOUZA, M. N. V. DE. **Comparação de algoritmos de aprendizado de máquina aplicados na mineração de dados educacionais**. [s.l.] Universidade Federal Rural de Pernambuco, 2015.

THULER, L. C. S. Câncer do Colo do Útero no Brasil : Estado da Arte. **Revista Brasileira de Cancerologia**, v. 58, n. 3, p. 321–337, 2012.

VALDESPINO, V. M.; VALDESPINO, V. E. Cervical cancer screening: state of the art. **Current opinion in obstetrics & gynecology**, v. 18, n. 1, p. 35–40, 2006.

VERAS, E. DE F. **Redes neurais artificais na classificação de microcalcificações em exames de mamografia**. [s.l.] Universidade Estadual do Piauí, 2015.

VIDYA, R.; NASIRA, G. M. Knowledge Extraction in Medical Data Mining : a Case Based Reasoning for Gynecological Cancer - an Expert Diagnostic Method. **ARPN Journal of Engineering and Applied Sciences**, v. 10, n. 9, p. 3997–4001, 2015.

VLAHOU, A. et al. Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data. **Journal of biomedicine & biotechnology**, v. 2003, n. 5, p. 308–314, 2003.