ANÁLISE DO DESEMPENHO DO ALGORTIMO K-NEAREST NEIGHBORS NA CLASSIFICAÇÃO DE PATOLOGIAS DE COLUNA VERTEBRAL

Natasha Seleidy Ramos de Medeiros [1]
Ingrid Bergmam do Nascimento Silva [2]
Larissa Duarte de Britto Lira[3]
Ingrid Rafaella dos Santos Melo [4]
Katia Suely Queiroz Silva Ribeiro [5]

RESUMO

A análise de grandes quantidades de dados torna-se, uma tarefa desafiadora. Para essa tarefa é imprescindível a utilização e desenvolvimento de ferramentas computacionais que, de forma inteligente, processem as informações dos bancos de dados para que auxiliem na tomada de decisão. Assim, o objetivo desse trabalho é buscar o melhor resultado do desempenho do algoritmo K-Nearest Neighbor (KNN) na classificação de patologias da Coluna Vertebral, para auxílio na tomada de decisão. Trata-se de uma pesquisa científica, experimental do tipo exploratória e prescritiva de abordagem quantitativa a partir da aplicação do algoritmo KNN, utilizando atributos biomecânicos para efetuar a categorização de um paciente em três classes: pacientes normais, pacientes com espondilolistese e pacientes com hérnia de disco. Os dados do banco foram treinados por várias simulações em busca do melhor modelo, e os melhores valores de predição apresentados pelo algoritmo KNN ocorreram no modo Percentage Split de 66%, Random seed for xval/% Split de 2, com *k* igual a 9 e a distância ponderada inversa. Obtendo como resultado um ótimo coeficiente Kappa capaz de classificar corretamente 88,57% dos indivíduos. A matriz de confusão mostrou os acertos e erros encontrados no modelo, correspondendo ao número total de instâncias para teste de 105, desses 93 foram classificados corretamente e 12 classificados incorretamente. Mostrando o modelo como eficaz para o proposto.

Palavras-chave: Tomada de Decisões, Coluna Vertebral, K-Nearest Neighbors.

INTRODUÇÃO

De acordo com Tessarolo e Magalhães (2015), existe um aumento de dados coletados da sociedade, empresas, ciência, engenharia, medicina e outras áreas da vida diária. A análise de grandes quantidades de dados se torna então, uma tarefa desafiadora. Para essa tarefa torna-se

[1] Fisioterapeuta, Mestranda em Modelos de Decisão e Saúde da Universidade Federal da Paraíba – UFPB, natashaseleidy@gmail.com;
[2] Enfermeira, Mestranda em Modelos de Decisão e Saúde da Universidade Federal da Paraíba – UFPB, ingridgba2006@hotmail.com;
[3] Fisioterapeuta, Mestranda em Modelos de Decisão e Saúde da Universidade Federal da Paraíba – UFPB, larissadblira@hotmail.com;
[4] Mestranda em Modelos de Decisão e Saúde da Universidade Federal da Paraíba – UFPB, ingridmeello@gmail.com;
[5] Professora orientadora: doutora em Educação, Universidade Federal da Paraíba - UFPB, katiaribeiro.ufpb@gmail.com.

imprescindível a utilização e desenvolvimento de ferramentas computacionais que, de forma inteligente, processem as informações dos bancos de dados para que auxiliem na tomada de decisão.

A Inteligência Artificial (IA) é uma das áreas mais importantes do ramo da computação atualmente. Essa é subdividida em diversos campos, sendo um dos principais: a Aprendizagem de Máquina (AM) - definida como a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência (MITCHELL, 1997). Ou seja, técnicas de aprendizagem de máquina têm como objetivo a classificação de padrões. Esse por sua vez, significa classificar um padrão desconhecido dentre várias classes possíveis (ALEEN; CAPREZ; AHMED, 2015).

Têm-se dois modelos de aprendizagem supervisionada: os modelos de regressão, onde a variável alvo é contínua e os modelos de classificação, onde a predição é feita para um atributo classificador que assume valores discretos (MESQUITA, 2015; WITTEN; FRANK; HALL, 2012).

Existem dois tipos de classificadores: os paramétricos, onde o desempenho e a exatidão dos classificadores estão ligados à distribuição normal dos dados, e os não paramétricos que podem ser usados para distribuições que não obedecem aos parâmetros de normalidade da curva (TANG; HAIBO, 2015; OLIVEIRA, 2016).

Para a tarefa de classificação e predição de desempenho são utilizados modelos matemáticos denominados algoritmos, sendo necessário que esse algoritmo utilizado na construção do modelo de aprendizado seja treinado. Independente, do tipo de problema ou de algoritmo a utilizar, a maioria das aplicações de aprendizagem computacional exige um passo inicial de pré-processamento dos dados que tem o objetivo de tornar os dados válidos e consistentes, aumentando a sua qualidade e também muitas vezes de os colocar em um formato em que o algoritmo possa ter um melhor desempenho (GOMES, 2015; LIMA, 2016).

A etapa de pós-processamento pode ser definida pelos processos de filtragem, estruturação e classificação dos resultados obtidos na mineração. Somente após esta fase, o conhecimento descoberto é apresentado ao usuário. O conhecimento gerado pelo processo de KDD é, via de regra, utilizado para dar suporte à tomada de decisões humana na resolução de problema em domínios específicos (SINGH; BANSAL, 2013; FARIA, 2016).

Nessa perspectiva, o objetivo desse trabalho é buscar o melhor resultado do desempenho do algoritmo K-Nearest Neighbor (KNN) na classificação de patologias da Coluna Vertebral, para auxílio na tomada de decisão.

METODOLOGIA

Quanto ao tipo do estudo, trata-se de uma pesquisa científica, experimental do tipo exploratória e prescritiva de abordagem quantitativa a partir da aplicação do algoritmo KNN para a construção de modelos preditivos na classificação.

O conjunto de dados correspondente a este artigo está disponibilizado no repositório internacional do site Uci machine learning repository (UCI, 2017), por meio do link: http://archive.ics.uci.edu/ml, autoria de Guilherme de Alencar Barreto, Ajalmar Rêgo da Rocha Neto e Henrique Antônio Fonseca da Mota Filho. Nesse repositório há duas versões do banco de dados, a versão utilizada nessa pesquisa trata-se da versão muticlasse, com 3 possíveis saídas (hérnia de disco, espondilolistese e normal), onde os dados foram extraídos de 310 pacientes. Sendo, que 100 indivíduos não possuem patologias na coluna, chamados de normais, 60 indivíduos apresentam hérnias de disco e 150 indivíduos espondilolistese. Cada um dos 310 pacientes é descrito por seis atributos biomecânicos, referentes a dores e deformidades na Coluna Vertebra, a saber: 1) ângulo de incidência pélvica; 2) ângulo de versão pélvica; 3) declive sacral; 4) ângulo de lordose, 5) raio pélvico e 6) grau de deslizamento. A fim de auxiliar o médico ortopedista na tomada de decisão no diagnóstico de patologias da coluna.

Existe uma grande diversidade de ferramentas que podem ser utilizadas para tarefas de mineração de dados. Nessa pesquisa será utilizada ferramenta Waikato Environment for Knowledge Analysis (WEKA), foi escrito em Java e desenvolvido na Universidade de Waikato (Nova Zelândia), na versão 3.8.2, essa ferramenta tem como objetivo agregar algoritmos provenientes de diferentes abordagens da área da inteligência artificial dedicada ao estudo de aprendizado de máquina. Além de possuir uma vasta coleção de algoritmos, desde algoritmos mais clássicos aos algoritmos mais atuais, possibilita a inclusão de novos algoritmos desde que atendam aos requisitos descritos na documentação. De acordo com Witten, Frank e Hall (2012) o WEKA foi projetado para que seja possível a experimentação rápida de métodos existentes em novos conjuntos de dados de maneira flexível.

Na tela inicial da ferramenta WEKA na sua versão 3.8.2, versão utilizada nos experimentos dessa pesquisa, utilizou-se a opção WEKA *Explorer* que é um ambiente que permite ao usuário executar apenas um algoritmo por vez apresentando diversas estatísticas a respeito do seu desempenho. É neste ambiente que são executadas as tarefas de pré-processamento de dados, classificação, regressão, clustering, regras de associação, seleção de atributos (BOUCKAERT et al., 2017; LIMA, 2016).

Após a seleção do conjunto de dados é feita a escolha do tipo de algoritmo que será utilizado para processar o conjunto de treinamento. Os algoritmos estão distribuídos em abas conforme o tipo de processamento que realizam, no caso do algoritmo KNN ele é listado na aba *Classify*, indicando que é um algoritmo classificador. O algoritmo KNN está dentro de um grupo denominado *Lazy* (aprendizado preguiçoso), com a denominação de *IBK*. Nessa tela também é feita a escolha do tipo de processamento que será empregado no algoritmo, usetraining set, Supllied test set, Cross-Validation e Percentage-Split, esse último consiste em uma dada porcentagem definida pelo usuário para treinamento e o restante para testes, sendo que os dados são previamente reordenados aleatoriamente. Outro parâmetro que foi utilizado para fazer os treinamentos foi obtido a partir da janela de mais opções (*more options*), o *random seed*, esse parâmetro é utilizado para a reordenação aleatória dos dados Bouckaert et al. (2017) e precisa ser diferente para cada treinamento caso se deseje que a reordenação seja varie entre experimentos.

Ao selecionar o algoritmo é possível configurá-lo para execução sobre o conjunto de treinamento. Ao clicar sobre o algoritmo é possível modificar seu padrão, tais como número de *k* (vizinhos mais próximos) e a medida de distância dentro do algoritmo de busca. Para o artigo em questão foi selecionado o número de *k* vizinhos mais próximos progressivamente de 1 em 1 e o *nearestNeighbourSearchAlgorithm* (Algoritmo de Procura do Vizinho mais próximo) que assume o valor com a distância Euclidiana, para cada *k* inserido, o parâmetro *distanceWeighting* (ponderação de distância) é treinado em 3 tentativas – distância não ponderada, distância ponderada inversa e distância ponderada linear. Os demais parâmetros como debug, meanSquared, entre outros permaneceram com os valores padrões do ambiente WEKA.

Dentro desse contexto, se faz necessário conhecimento sobre as métricas de avaliação, que são utilizadas para identificar o desempenho do algoritmo aplicado sobre o problema em questão. Ao clicar em *start* o software WEKA apresenta a saída dos resultados com as principais métricas: 1) no Sumário apresentam – instâncias classificadas corretamente (Correctly Classified Instances - CCI) e incorretamente (Incorrectly Classified Instances - ICI), estatística Kappa (Kappa statistic), Erro médio absoluto (Mean Absolute Error - MAE), erro quadrado médio da raiz (Root Mean Squared Error - RMSE), erro relativo absoluto (Relative Absolute Error- RAE), raiz quadrada do erro relativo (Root Relative Squared Error - RRSE) e número total de instâncias (Total Number of Instances); 2) Precisão detalhada por classe (Detailed Accuracy By Class); 3) matriz de confusão (Confusion Matrix).

Por fim, o modelo treinado é salvo em arquivo com a extensão *.model* e com nome indicando qual o algoritmo foi utilizado, sendo esse o padrão para que o módulo de classificação consiga utilizar esse modelo para classificar novas instâncias.

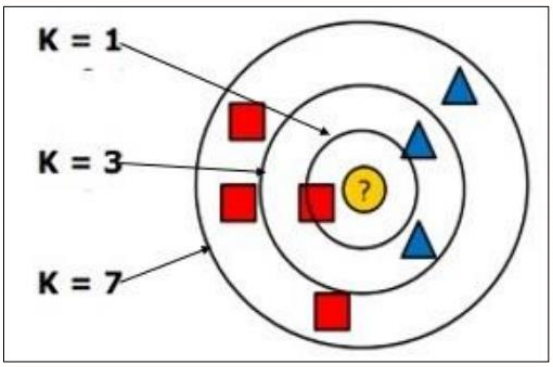
DESENVOLVIMENTO

O algoritmo K-Nearest Neighbor (KNN), do português K-Vizinho mais próximo, pertence à família de algoritmos Instance-based Learning (IBL), ou seja, aprendizagem baseada em instâncias. Os algoritmos desta família armazenam todas as instâncias de treinamento e, quando uma nova instância é apresentada ao algoritmo para ser classificada, um conjunto de instâncias similares à essa nova instância é recuperada do conjunto de treinamento e utilizada para classificá-la (FARIA, 2016).

O KNN, é um método de aprendizagem supervisionada, do tipo classificador, não-paramétrico, que utiliza *Lazy Learning* e possui três elementos principais: um conjunto de exemplos rotulados (por exemplo, um conjunto de registros armazenados), uma métrica de distância, e o valor de *k* (o número de vizinhos mais próximos) (OLIVEIRA, 2016).

A Figura 1 apresenta um exemplo para ilustrar o funcionamento do algoritmo KNN. Nessa figura existe uma instância a ser classificada representada pela interrogação, e instâncias de treinamento já classificadas associadas à classe triângulo e à classe quadrado.

Figura 1: Funcionamento do KNN



Nesse contexto, para o valor de $k = 1$, pelo funcionamento do algoritmo KNN, a nova instância será classificada como pertencente à classe quadrado, uma vez que a classe do vizinho mais próximo é a classe quadrado. Para um caso de valor $k = 3$, a classe da nova instância será

triângulo, dado que duas instâncias dos três vizinhos mais próximos têm classe triângulo e uma tem classe quadrado. Já no caso de $k = 7$, a classe da nova instância será a classe quadrado. Neste algoritmo, o que determina a classificação é a maior frequência das classes dos *k* vizinhos mais próximos da instância a ser classificada (FARIA, 2016; AKBARI; OVERLOOP; AFSHAR, 2011).

Para determinar os vizinhos mais próximos ou similares é utilizado o conceito de distância entre a instância a ser classificada e as instâncias do conjunto de treinamento mais próximas a ela. A medida mais usada é a distância Euclidiana, que é normalmente definida como a distância mais curta (TANG; HAIBO, 2015).

Sabendo-se que existe dois dados associados com os pontos E_i e E_j pertencentes a um espaço m-dimensões, notados por $E_i = (x_{i1}, x_{i2}, ..., x_{im})$ e $E_j = (x_{j1}, x_{j2}, ..., x_{jm})$, a distância Euclidiana (*dist*) entre esses dois pontos é dada pela Equação [1]. A distância Euclidiana entre os pontos E_i e E_j representa o comprimento do segmento de reta que os conecta.

$$dist(E_i, E_j) = \sqrt{\sum_{l=1}^{M} (x_{il} - x_{jl})^2} \quad [1]$$

As vantagens do KNN, estão na sua simplicidade de implementação, o seu desempenho é bastante eficaz em diversas situações e áreas (engenharia, saúde, educação, entre outras), possui fácil interpretação, e é ideal para bancos de dados pequenas ou médios. Além disso, constrói diretamente a regra de decisão sem estimar as densidades condicionadas às classes, sendo uma boa escolha para problemas de classificação em que padrões próximos no espaço de características possivelmente pertencem à mesma classe (OLIVEIRA, 2016).

Infelizmente, algumas desvantagens são inerentes ao algoritmo, onde nem sempre é fácil determinar o melhor valor para o parâmetro *k*. Sendo, o valor de *k* obtido a partir de tentativas e erros. Outra desvantagem é que na busca pelos *k* vizinhos mais próximos o algoritmo necessita realizar uma passagem completa por todas as instâncias de treinamento, torna-se um problema quando se trata de grandes volumes de dados. Ainda, quando o número de amostras não é balanceado, como uma classe com um grande número de amostras, enquanto as outras classes são pequenas, pode acontecer erros de classificação, porque na previsão de novas amostras, a maioria

dos vizinhos podem pertencer às classes de grandes dimensões (OLIVEIRA, 2016; TANG; HAIBO, 2015).

Para resolver essas limitações é atribuído um peso à distância do KNN (WANG; WANG, 2016; TANG; HAIBO, 2015). Com isso, temos a distância Euclidiana ponderada [2]:

$$d(\mathbf{X}_q, \mathbf{X}_i) = \sqrt{\sum_{j=1}^{n} w_j^A (x_{ij} - x_{qj})^2} \quad [2]$$

Em que, w^A_j é o peso do j-ésimo atributo. Com os valores de saída Y_i dos *k* vizinhos mais próximos, o valor previsto para Y_q é calculado pela equação [3]:

$$y_q = \frac{\sum_{i=1}^{k} w_i^V y_i}{\sum_{i=1}^{k} w_i^V} \quad [3]$$

Onde, w^V_i é o peso do i-ésimo vizinho. Na sua forma mais simples, o modelo KNN usa $w^V_i = 1$ e a estimativa é o valor médio dos *k* vizinhos mais próximos. Na versão melhorada, entretanto, cada vizinho possui um peso baseado na distância d (X_q, X_i). Assim, um vizinho mais longe recebe um peso menor, o que reduz seu efeito sobre a predição em comparação com outros vizinhos mais próximos. Desta forma, algumas funções kernel, que decrescem, à medida que a distância aumenta, têm sido utilizadas [4] (AKBARI et al., 2011; SANTOS; CELESTE, 2014; WANG; WANG, 2016):

$$\text{Linear:} \quad w_i^V = 1 - d(\mathbf{X}_q, \mathbf{X}_i)$$

$$\text{Inversa:} \quad w_i^V = [d(\mathbf{X}_q, \mathbf{X}_i)]^{-1} \quad [4]$$

RESULTADOS E DISCUSSÃO

Dessa forma a utilização do banco de dados para verificar e determinar qual modelo do algoritmo KNN consegue classificar melhor as instâncias do conjunto de dados como normal, Hérnia de Disco ou Espondilolistese, perpassa por duas fases, a saber: treinamento e teste. Cada experimento foi experimentado por diversas simulações, dentro do Use training set (uso do conjunto

de treinamento), Percentage Split (Porcentagem de Divisão) ou Cross-Validation (validação cruzada) nesses foram ajustados os parâmetros parâmetros de acordo com a necessidade de cada modelo, número de *k* vizinhos e o cálculo da distância – distância não ponderada, distância ponderada inversa e distância ponderada linear e demais.

Assim, após diversas tentativas o modelo se adequou melhor a Percentage Split – onde, a percentagem de treinamento é definida pelo usuário (BOUCKAERT et al. 2017). Nesse contexto, a percentagem de treinamento que obteve melhor resultado foi de 66%, com resultados obtidos em teste para 34% restantes (total de instâncias para teste de 105).

Tabela 1: Resultados das simulações para identificar o melhor modelo.

Percetage splint 66% random2	**Número de vizinhos próximos**	**Distâncias**	**Instâncias Classificadas Corretamente**	**%**	**Kappa**	**Erro médio absoluto**
	8	Não ponderada	89	84.7619	0.751	0.1814
		Ponderada inversa	91	86.6667	0.7813	0.1767
		Ponderada linear	90	85.7143	0.7659	0.1813
	9	Não ponderada	91	86.6667	0.7827	0.1817
		Ponderada inversa	**93**	**88.5714**	**0.8125**	**0.1776**
		Ponderada linear	93	88.5714	0.8125	0.1816
	10	Não ponderada	91	86.6667	0.7815	0.1864
		Ponderada inversa	92	87.619	0.7955	0.1817
		Ponderada linear	91	86.6667	0.7801	0.1862

Fonte: Dados da pesquisa, 2018.

Quando a amostra foi subdividida com o Percentage Split de 66%, Random seed for xval/% Split de 2, com *k*=9 e a distância ponderada inversa, ela obteve percentuais de acertos bastante consideráveis. Desse modo, os parâmetros foram executados 3 vezes, no software WEKA, para confirmar os valores, determinando os melhores parâmetros para ajuste do modelo. Ao ponderar as distâncias obteve-se melhor acertos na classificação, especialmente na ponderada inversa.

Para avaliar o desempenho dos algoritmos classificadores se faz necessário o uso de métricas de avaliação para verificar se o resultado da predição é satisfatório. A Matriz de Confusão é uma ferramenta útil para analisar o quão bem o classificador pode reconhecer exemplos de diferentes classes. Ela contém o número de elementos que foram corretamente ou incorretamente classificadas para cada classe, na sua diagonal principal apresenta o número de exemplos que foram corretamente classificados, enquanto que os elementos fora da diagonal indicam o número de exemplos que foram classificados incorretamente.

As linhas desta Matriz de Confusão correspondem às classes conhecidas através dos dados nas amostras, enquanto as colunas correspondem à de classificação do modelo. Sendo assim, o modelo apresentou 88.57% de instâncias classificadas corretamente.

Tabela2: Matriz de Confusão apresentada pelo modelo KNN.

	Hérnia de Disco	Espondilolistese	Normal
Hérnia de Disco	**17**	0	3
Espondilolistese	2	**52**	2
Normal	3	2	**24**

Fonte: Dados da pesquisa, 2018.

Ainda podemos observar que a Matriz de Confusão é uma ótima ferramenta para tomada de decisão. Assim, temos os resultados de Falsos Positivos (FP): que referem-se aos exemplos negativos que foram incorretamente rotulados como exemplos positivos pelo classificador. Logo, um diagnóstico é denominado **FP** quando um indivíduo sadio é diagnosticado como doente, ou ainda, quando o indivíduo apresenta uma determinada doença e é classificado com outra doença.

Já os Falsos Negativos (FN): referem-se aos exemplos positivos que foram incorretamente rotulados como exemplos negativos pelo classificador. Logo, um diagnóstico é denominado **FN** quando um indivíduo doente (com espondilolistese ou hérnia de disco) é diagnosticado como saudável.

Verifica-se, portanto, a ocorrência de 7 Falsos Positivo: 2 indivíduos com espondilolistese classificados como possuindo hérnia de disco, 3 indivíduos normais classificados como possuindo hérnia de disco e 2 indivíduos normais classificados como possuindo espondilolistese.

Com relação aos Falsos Negativo observa-se um total de 5: sendo, 3 indivíduos com hérnia de disco e 2 com espondilolistese classificados como normais (saudáveis). Esse torna-se um caso mais grave de erro na classificação do diagnóstico, pois esses indivíduos ficarão sem tratamento adequado, o que agravará seu quadro clínico e posterior qualidade de vida.

Outra métrica de avaliação importante na análise de desempenho do modelo de decisão é o Erro Médio Absoluto (Mean Absolute Error – MAE), trata-se da diferença entre a previsão e a observação, ou seja, mede o quanto próximo (distância em valor absoluto) os valores preditivos estão dos valores corretos. Os valores mais próximos de zero indicam que ocorreu melhor classificação. O modelo proposto apresenta MAE 0.1776.

O Kappa é uma medida de concordância entre as predições de um classificador com a classe correta, tem como valor máximo 1, onde este valor representa total concordância e os valores próximos e até abaixo de 0, indicam nenhuma concordância. Quando um classificador tem uma alta porcentagem de instâncias classificadas corretamente, o Kappa terá um valor maior. Esse coeficiente é uma medida de exatidão mais robusta, pois representa inteiramente a matriz de confusão (LANTZ, 2013).

A partir desses conceitos, podemos afirmar que os resultados obtidos a partir dos dados dessa pesquisa, apresentaram-se satisfatórios para o MAE com 0.1776 e Kappa de 0.8125. Ou seja, o Erro Absoluto Médio próximo de zero e uma concordância forte do Kappa que segundo Lantz (2013), é aquela acima de 0,80. Essas evidências indicam que ocorreu uma boa classificação e que o modelo é adequado para o diagnóstico de patologias da Coluna Vertebral, auxiliando o médico ortopedista na Tomada de Decisão.

Utilizando o mesmo banco de dados para implementação de um software denominado SINPATCO, que utiliza o algoritmo KNN como um de seus algoritmos para classificação de patologias de coluna, e utilizando a distância euclidiana como métrica de proximidade e em um *k* igual a 7, Neto et. al. (2008), encontrou uma porcentagem de 85,24% das instâncias classificadas corretamente, 7 Falsos Negativos e 11 Falsos Positivos. O que mostra uma melhor adequação do modelo KNN para esse presente artigo, pois obteve melhores resultado de classificação.

CONSIDERAÇÕES FINAIS

Neste contexto, esta pesquisa apresentou uma avaliação e descrição do desempenho do algoritmo KNN na construção de um modelo de tomada de decisão com base em um banco de

dados referente a patologias da coluna vertebral. Demonstrando assim, que é possível a aplicação e utilização de um modelo de tomada de decisão, a partir de ferramentas computacionais, direcionado para a o apoio diagnóstico dessas patologias por médicos especialistas em ortopedia.

Os dados do banco foram treinados por várias simulações em busca do melhor modelo, e os melhores valores de predição apresentados pelo algoritmo KNN ocorreram no modo Percentage Split de 66%, Random seed for xval/% Split de 2, com *k* igual a 9 e a distância ponderada inversa. Obtendo como resultado um ótimo coeficiente Kappa capaz de classificar corretamente 88,57% dos indivíduos. A matriz de confusão mostrou os acertos e erros encontrados no modelo, correspondendo ao número total de instâncias para teste de 105, desses 93 foram classificados corretamente e 12 classificados incorretamente.

Assim, o modelo KNN avaliado possui potencialidades e fragilidades no banco de dados utilizado, no entanto para o objetivo proposto nessa pesquisa, o mesmo atendeu a necessidade e pode servir como para o problema de classificação de patologias da coluna vertebral.

REFERÊNCIAS

AKBARI, M.; OVERLOOP, P.; AFSHAR, A. Clustered k nearest neighbor algorithm for daily inflow forecasting. **Water Resources Management**, v. 25, n. 5, p. 1341–1357, 2011.

ALEEN, Saiqa; CAPREZ, Luiz Fernando; AHMED, Faheen. Benchmarking machine learning techniques for software defect detection. **International Journal of Software Engineering & Applications (IJSEA)**, (3):11–23, May 2015.

FARIA, Mauricio Mendes. **Detecção de intrusões em redes de computadores com base nos algoritmos KNN, K-Means++ e J48**. Dissertação (Programa de Mestrado em Ciência da Computação) – Faculdade Campo Limpo Paulista – FACCAMP. São Paulo, 2016.

LANTZ, B. **Machine Learning with R**. Packt Publishing Ltd., out. 2013.

LIMA, André Accioly. **Estudo Experimental de Aprendizado de Máquina para Desenvolvimento de um Classificador de Texto de Incidentes de Grandes Eventos**. Monografia (Graduação) — Universidade de Brasília, Brasília. 2016.

MESQUITA, D. P. P.; GOMES, J. P. P.; SOUZA JUNIOR, A. H. **Ensemble of minimal learning machines for pattern classification**. International Work Conference on Artificial Neural Networks, IWANN, 2015.

MITCHELL, Tom. **Machine Learning**. McGraw-Hill, 1997.

NETO, Ajalmar R. R.; CORTEZ, Paulo C.; MOTA; Henrique da; BARRETO, Guilherme A. SINPATCO: Sistema Inteligente para Diagnóstico de Patologias da Coluna Vertebral. XVI Congresso Brasileiro de Automática. Salvador-BA, 2008.

OLIVEIRA, Adonias Caetano de. **Máquina de Aprendizagem Mínima com Opção de Rejeição**. Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2016.

OLIVEIRA, A. R; ROESLER, V; IOCHPE, C; SCHMIDT, M. I; VIGO, A; BARRETO, S. M; DUNCAN, B. B. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes – ELSA-Brasil: accuracy study. **Sao Paulo Med J,** v.135, n. 3, p.234-46, 2017.

OMAR, Sakima; NGADI, Asru; JEBUR, Hamid H. Machine learning techniques for anomaly detection: An overview. **International Journal of Computer Applications**, Outubro, 2013.

SANTOS, J.R.S; CELESTE, A.B. Avaliação de estratégias de modelagem guiada por dados para previsão de vazão em rio sergipano. **Rev. Ambient. Água, v**. 9, n. 3, 2014.

SANTOS, R.M.M. **Técnicas de Aprendizagem de Máquina Utilizadas na Previsão de Desempenho Acadêmico** [dissertação]. Diamantina: Universidade Federal dos Vales do Jequitinhonha e Mucuri; 2016.

SINGH, S.; BANSAL, M., A Survey on Intrusion Detection System in Data Mining. **International Journal of Advanced Research in Computer Engineering & Technology**, 2013.

TANG, B; HAIBO, H. ENN: Extended Nearest Neighbor Method for Pattern Recognition. **IEEE Computational intelligence magazine**, v.10 (Issue: 3), 2015.

TESSAROLO, Pedro Henrique; MAGALHÃES, William Barbosa. A era do big data no conteúdo digital: os dados estruturados e não estruturados. **XVII SEINPAR - Semana de Informática de Paranavaí**, 2015.

WANG, X; LI, H; ZHANG, Q; WANG, R. Predicting Subcellular Localization of Apoptosis Proteins Combining GO Features of Homologous Proteins and Distance Weighted KNN Classifier. **BioMed Research International**, 2016.