RENORBIO – Rede Nordeste de Biotecnologia

Programa de Pós-Graduação em Biotecnologia

VANESSA EDILENE DUARTE MARTINS

DESENVOLVIMENTO DE SISTEMA COMPUTACIONAL PARA PREDIÇÃO DA DOENÇA RENAL CRÔNICA

São Luís -MA

2020

RENORBIO – Rede Nordeste de Biotecnologia

Programa de Pós-Graduação em Biotecnologia

VANESSA EDILENE DUARTE MARTINS

DESENVOLVIMENTO DE SISTEMA COMPUTACIONAL PARA PREDIÇÃO DA DOENÇA RENAL CRÔNICA

Defesa de tese apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal do Maranhão (UFMA) como requisito parcial para obtenção do título de Doutor em Biotecnologia.

Orientador: Allan Kardec Duailibe Barros Filho

Orientanda: Vanessa Edilene Duarte Martins

São Luís- MA

2020

Martins, Vanessa Edilene Duarte.
DESENVOLVIMENTO DE SISTEMA COMPUTACIONAL PARA PREDIÇÃO DA DOENÇA RENAL CRÔNICA / Vanessa Edilene Duarte Martins. - 2020.
87 f.

Orientador(a): Allan Kardec Duailibe Barros Filho.
Tese (Doutorado) - Programa de Pós-graduação em Biotecnologia - Renorbio/ccbs, Universidade Federal do Maranhão, UFMA, 2020.

1. Aprendizado de máquina. 2. Autocuidado. 3. Diagnóstico precoce. 4. Inteligência artificial. 5. Medicina computacional. I. Barros Filho, Allan Kardec Duailibe. II. Título.

renorbio
rede nordeste de biotecnologia

FOLHA DE APROVAÇÃO DE DEFESA DE TESE

ALUNA: Vanessa Edilene Duarte Martins

TÍTULO DO PROJETO: Desenvolvimento de Sistema Computacional para Predição da Doença Renal Crônica

PROFESSOR ORIENTADOR: Allan Kardec Duailibe Barros Filho

BANCA EXAMINADORA:	**CONCEITO**	**ASSINATURA**
Prof. Dr. Allan Kardec Duailibe Barros Filho- UFMA (Presidente)	_____	_____
Profª. Drª. Audirene Amorim Santana – UFMA (Titular)	_____	_____
Profª. Drª. Joicy Cortez de Sá Sousa – UNICEUMA (Titular)	_____	_____
Prof. Dr. Daniel Praseres Chaves – UEMA (Titular)	_____	_____
Prof. Dr. Antônio Carlos Romão Borges – UFMA (Titular)	_____	_____

DATA DA APROVAÇÃO: 12 de março de 2020.
HORÁRIO: 14:00h.
LOCAL: Prédio da Pós-Graduação do CCBS-UFMA.

"Tudo posso Naquele que me fortalece"

(Filipenses: 4.13)

"A melhor de todas as coisas é aprender.
O dinheiro pode ser perdido ou roubado,
a saúde e força podem falhar,
mas o que você dedicou à sua mente
é seu para sempre".
(Louis L. Amour)

"A satisfação vem quando causamos um impacto positivo e eterno
na vida dos outros"

(Janet Bly)

AGRADECIMENTOS

A Deus, meu protetor e guia, pela dádiva de poder concluir mais um sonho em minha vida.

A Universidade Federal do Maranhão, ao Programa de Doutorado em Biotecnologia (RENORBIO) pelo incentivo da realização da pesquisa científica.

Ao Laboratório de Processamento de Informação Biológica (PIB) e a todos envolvidos nesse ambiente de trabalho pelo conhecimento na área de aprendizado de máquina a qual fiquei imensamente apaixonada.

Ao prof. Allan Kardec, meu orientador, por me aceitar em sua equipe, sua confiança, ensinamentos, paciência e "puxões de orelha".

Ao prof. André e aos meninos Antonino e Jonnilson, que foram fundamentais na minha pesquisa, me ajudaram bastante, sem eles não teria conseguido.

Aos amigos que o doutorado me proporcionou, em especial a Viviane, Marta, Ilka, Lourival, Carlos, pela companhia, auxilio e trocas de experiencias.

A minha família, meus pais Valmir e Marineide, por todo amor, educação, incentivo em todos os momentos da minha vida profissional e pessoal. A minha irmã Ediele, enfermeira, que me auxiliou muito em relação a área da saúde.

A meu namorado Junior Lima, pelo amor, compreensão, companheirismo, paciência e por me transmitir paz nos momentos difíceis e estressantes, além de ser meu patrocinador oficial na conclusão da Tese.

À todos que, de uma forma ou de outra, fizeram companhia dentro do laboratório, e que quando era preciso alguém me auxiliar, nunca me negaram ajuda. Meu agradecimento ao Daniel, Luís, Gomes, Cláudia, Gisele, Ilmar e todos do PIB.

E a todos aqueles que contribuíram de forma direta ou indiretamente para a realização deste trabalho.

Minha eterna gratidão!

Resumo geral

A doença renal crônica (DRC) não apresenta sinais e/ou sintomas em seus estágios iniciais, sendo de suma importância o estudo e desenvolvimento de métodos de diagnóstico e/ou triagem alternativos que tenham alta sensibilidade. Assim, objetivou-se desenvolver um sistema computacional para predição da doença renal crônica. O trabalho de tese está dividido em três capítulos, além da fundamentação teórica os quais são apresentados os principais temas que fundamentam o presente projeto. O Capitulo I representa um artigo intitulado "Artificial Intelligence in Predicting Chronic Kidney Disease" publicado na Revista *International Journal of Development Research* que objetivou-se realizar uma revisão da literatura sobre o uso da Inteligência Artificial na predição de Doença Renal Crônica. De acordo com as pesquisas, foi observado que a DRC pode ser prevista usando vários classificadores em mineração de dados, bem como prever o estágio da doença com uso da IA e que as diferentes experiências observadas mostraram que a maioria dos classificadores fornece alto valor de acurácia, acima de 90%. O Capítulo II, artigo de pesquisa intitulado "Development of a computer system to screenin patients with chronic kidney disease" publicado na Revista *International Journal of Development Research* visa construir um sistema computacional para auxiliar no diagnóstico precoce da Doença Renal Crônica (DRC) usando dados clínicos não invasivos, explorando técnicas de aprendizado de máquina. E por fim, o Capítulo III artigo de pesquisa intulado "Support System for Chronic Kidney Disease Prediction using Machine Learning" submetido na Revista *PeerJ,* objetivou-se construir e validar um software preditor da doença renal crônica baseado em um algoritmo classificador para triagem de pacientes. Dentre os 3 classificadores utilizados nos experimentos, o SVM foi o que obteve melhores resultados e usado para obtenção do software preditor da DRC, demostrou bom desempenho na validação o qual pode ser usado na pratica clínica como forma de triagem de pacientes com a doença e para a população em geral, apresentando uma alternativa de baixo custo e fácil execução.

Palavras-chave: Aprendizado de máquina, autocuidado, classificadores, diagnóstico precoce, inteligência artificial, medicina computacional.

Abstract

Chronic kidney disease (CKD) does not show signs and / or symptoms in its recent symptoms, and it is important to study and develop alternative methods of diagnosis and / or screening with high sensitivity. Thus, the objective was to develop a computer system for predicting chronic kidney disease. The thesis work is divided into three chapters, in addition to the theoretical foundation and what are the main themes that underlie the present project. Chapter I represents an article entitled "Artificial Intelligence in the Prediction of Chronic Kidney Disease" published in the International Journal of Development Research which aimed to conduct a review of the literature on the use of Artificial Intelligence in the prediction of Chronic Kidney Disease. According to the research, it was observed that CKD can be considered using various classifiers in data mining, as well as predicting the stage of the disease with the use of AI and that the different observed experiences that most classifiers use with high accuracy value, above 90%. Chapter II, a research article entitled "Development of a computer system for tracking patients with chronic kidney disease" published in the International Journal of Development Research aims to build a computer system for the early diagnosis of Chronic Kidney Disease (CKD) using clinical data non-invasive, exploring machine learning techniques. And finally, the Chapter III research article "Support System for Chronic Kidney Disease Prediction using Machine Learning" submitted in PeerJ Magazine, aimed to build and validate a chronic kidney disease predictor software related to a classifier algorithm for patients with kidney disease. Among the 3 classifiers used in the experiments, the SVM was the one that obtained the best results and used to test the RDC software predator, demonstrated a good validation performance or can be used in clinical practice as a way of screening patients with disease and for a general population, presenting a low cost and easy execution alternative.

Key words: Machine learning, self-care, classifiers, early diagnosis, artificial intelligence, computational medicine.

LISTA DE FIGURAS

Capitulo II

Capitulo III

LISTA DE TABELAS

Capitulo I

Capitulo II

Capitulo III

LISTA DE ABREVIATURAS E SIGLAS

AINE Anti-Inflamatórios Não-Esteroides

AP Aprendizado de Máquina

BPA Algoritmo de Propagação Traseira

BN Bayesian Network

BPN Redes Neurais Artificiais Incluindo Redes de Propagação Reversa

CKD Chronic Kidney Disease

DCV Doença Cardiovascular

DCBD Descoberta de Conhecimento em Banco de Dados

DM Diabetes mellitus

DDS Serviço de Distribuição de Dados

DT Árvore de Decisão

DRC Doença Renal Crônica

DTPA-Tc99m Ácido Dietilenotriaminopentácetico Marcado Com Tecnécio99m

EDTA Ácido etilenodiaminotetraacético

FFR Falência Funcional Renal

FG Filtração Glomerular

FP Falso Positivo

FN Falso Negativo

GRNN Redes Neurais De Alimentação Generalizada

HDL Lipoproteínas de Alta Densidade

IA Inteligência Artificial

IV Intravenoso

KNN K-Nearest Neighbor

LDL Lipoproteínas de Baixa Densidade

NB Naive Bayes

MLPC Multilayer Perceptron Classifier

MNN Redes Neurais Modulares

PA	Pressão Arterial
PAD	Pressão Arterial Diastólica
PAS	Pressão Arterial Sistólica
RF	Random Forest
RBF	Radial Basis Function
RS	Reed-Solomon
SMO	Otimização Mínima Seqüencial
SLG	Simple Logistic
SVM	Máquinas de Vectores Suporte
TFG	Taxa Filtração Glomerular
TPV	Taxa Positiva Verdadeira
TNV	Taxa Negativa Verdadeira
UCI	Universidade da Califórnia de Irvine
UIHC	Hospitais da Universidade de Iowa e Clínicas
VLDL	Lipoproteína de Muito Baixa Densidade
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

Introdução geral

Atualmente, os maiores problemas de saúde pública são os casos de doenças crônicas não transmissíveis (DCNT), sendo responsáveis por 63% de um total de 36 milhões de mortes ocorridas no mundo, segundo as estimativas da Organização Mundial da Saúde (OMS) (Who, 2011; Ministério da Saúde, 2015). No Brasil, as DCNT são igualmente relevantes, tendo sido responsáveis, em 2011, por 72,7% do total de mortes, com destaque para as doenças do aparelho circulatório (30,4% dos óbitos), as neoplasias (16,4%), o diabetes (5,3%) e as doenças respiratórias (6,0%) (Malta et al., 2014; Ministério da Saúde, 2015)

Dentre as DCNT destaca-se a doença renal crônica (DRC) que é caracterizada pela alteração da função renal, definida como anormalidades da estrutura ou função dos rins presentes por mais de três meses e com implicações para a saúde do indivíduo (National Kidney Foundation, 2013; Draibe, 2014). Considerada um problema de saúde global que está aumentando, principalmente, como resultado da crescente incidência da obesidade, diabetes e hipertensão (Baumgarten e Gehr, 2011), além de estar associada a mudanças nos padrões de consumo, alteração no estilo de vida e transição demográfica (Neto e Malik, 2012; Silva et al., 2016).

Por ser uma doença assintomática em seus estágios iniciais é de suma importância o estudo e desenvolvimento de métodos de diagnóstico e/ou triagem alternativos que tenham alta sensibilidade. O impacto econômico dessa patologia também é outra preocupação das autoridades em saúde pública, já que, além de muito dispendioso, o tratamento medicamentoso e dialítico praticamente alija os indivíduos em idade produtiva de sua capacidade laborativa, afetando o sistema de previdência pública e seguridade social (Sodré et al., 2007). Logo, medidas com foco na detecção precoce DRC, especialmente em pacientes com risco aumentado de desenvolver a doença, incluindo-se nesse grupo hipertensos, diabéticos, pacientes portadores de doença cardiovascular e pessoas com história familiar de insuficiência renal crônica (IRC) são de grande auxilio para os profissionais de saúde (Sodré et al., 2007).

O tratamento ideal da DRC é baseado em três pilares de apoio incluindo o diagnóstico precoce da doença, encaminhamento imediato para tratamento nefrológico e implementação de medidas para preservar a função renal (Bastos e Kirsztajn, 2011). Portanto, é de grande importância trabalhos que auxiliem na prevenção e diagnóstico precoce da DRC. Pois, dessa forma podem auxiliar no retardamento ou interrompimento

da progressão da DRC para os estágios mais avançados, bem como diminuir a morbidade e mortalidade iniciais.

Na área de aprendizado de máquina (AM), vários algoritmos classificadores são usados em estudos na previsão de diversas doenças , como doença cardíaca (Xing et al., 2007; Lee et al., 2008; Srinivas et al., 2010; Pal et al., 2011), câncer, epilepsia, doença de Parkinson, diabetes, doença de Parkinson (Su et al., 2001; Rajan e Chelvan, 2013; Ilayaraja e Meyyappan, 2013; Ghannad-Rezaie e Soltanain-Zadeh, 2008; BONATO et al., 2004) incluindo estudos recentes para a detecção DRC (Lakshmi et al., 2014; Xun et al., 2010; Kunwar et al., 2016; Chiu et al., 2012). Assim, verifica-se que métodos de AM são uma solução para problemas de classificação como a triagem de pacientes com DRC, pois oferecerem uma previsão mais exata sobre a saúde do indivíduo (Lenart, 2016).

Objetivos

Geral

Desenvolver um sistema computacional capaz de predizer a DRC com base nos fatores de risco na doença.

Específicos

- Avaliar indicadores antropométricos, hemodinâmico e bioquímico dos pacientes;
- Analisar e comparar o desempenho dos diferentes algoritmosclassificadores;
- Obter uma interface gráfica de variáveis de entrada com dados não invasivos para classificação da DRC;
- Desenvolver e validar um software para rastreamento de pacientes com DRC.

Fundamentação Teórica

Nesta revisão bibliográfica, são apresentados os principais conceitos que fundamentam o presente projeto. Na Seção 1 são abordados os conceitos introdutórios da Doença Renal Crônica (DRC), bem como epidemiologia, classificação e tratamento. O cenário inicial do trabalho refere-se à DRC, englobando os principais fatores de risco relacionados à doença e diagnóstico. Como o objetivo da pesquisa de tese é desenvolver um sistema computacional para predição da DRC, a Seção 2 conceitua sobre Mineração de dados e Aprendizado de Máquina (AM). Desta forma, a Seção 3 trata dos algoritmos de AM utilizados no trabalho. Por fim, na Seção 4 é abordado a ferramenta Weka que auxiliou na tomada de decisão do algoritmo a ser implementado e na Seção 5 os cálculos utilizados para a validação dos resultados.

1. Doença Renal Crônica: Definição, Epidemiologia, Classificação e Tratamento

A DRC é caracterizada pela alteração da função renal (Ministério da Saúde, 2015), sendo definida como anormalidades da estrutura ou função dos rins presente por mais de três meses e com implicações para a saúde, considerada problema de saúde pública em todo o mundo (National Kidney Foundation, 2004; Draibe, 2014).

A prevalência mundial da DRC tem aumentado nas últimas décadas, cerca de 2,5 milhões de pacientes estavam em diálise no mundo em 2013, e este número pode chegar a 6,5 milhões em 2030 (Care, 2014). Nos Estados Unidos a DRC afeta aproximadamente 27 milhões de adultos e está associada ao aumento da mortalidade, morbidade e custos dos cuidados de saúde (Baumgarten e Gehr, 2011). No ano de 2014, o número total estimado de pacientes em diálise no Brasil foi de 112.004, com 91,4% em hemodiálise, e 8,6% em diálise peritoneal (Sesso et al., 2016).

No Brasil, a incidência e a prevalência de falência da função renal estão aumentando, o prognóstico ainda é ruim e os custos do tratamento da doença são altíssimos. O número projetado para pacientes em tratamento dialítico e com transplante renal está próximo dos 120.000, a um custo de 1,4 bilhão de reais (Sesso et al., 2008).

A DRC pode ser classificada em cinco estágios, de acordo com o grau de redução da filtração glomerular, indo da condição normal/elevada até diálise ou transplante (Draibe, 2014). As complicações da doença renal crônica podem afetar todo o organismo e apresentar-se em qualquer estágio de evolução da enfermidade, frequentemente levando à morte, sem progressão para insuficiência renal estágio 5 (Polito, 2014).

Tabela 1. Estágios da doença renal crônica de acordo com os valores da taxa de filtração glomerular, conforme as recomendações da National Kidney Foundation

Estágio	Descrição	FG*
1	Lesão renal com FG normal ou aumentada	≥90
2	Lesão renal com FG levemente diminuída	60-89
3	Lesão renal com FG moderadamente diminuída	30-59
4	Lesão renal com FG severamente diminuída	15-29
5	FFR** estando ou não em terapia renal substitutiva	<15

*FG= Filtração Glomerular em mL/min/1,73m² **FFR= Falência Funcional Renal

Na Tabela 1 está a classificação dos estágios da DRC, em que o estágio 1 representa lesão renal normal ou aumentada com filtração glomerular preservada e ritmo de igual ou superior a 90 ml/min/1,73m^2. Do ponto de vista epidemiológico, é importante verificar, pois inclui pessoas integrantes dos chamados grupo de risco para o desenvolvimento da DRC, que são os hipertensos, diabéticos, histórico familiar, entre outros, mas que ainda não desenvolveram lesão renal (Romão, 2004).

O estágio 2 apresenta insuficiência renal levemente diminuída, onde já ocorre no início da perda de função dos rins. Nesta fase, os níveis de uréia e creatinina plasmáticos ainda são normais, não há sinais ou sintomas clínicos importantes e somente métodos acurados de avaliação da função do rim (métodos de depuração, por exemplo) irão detectar estas anormalidades. Os rins conseguem manter razoável controle do meio interno. Compreende a um ritmo de filtração glomerular entre 60 e 89 ml/min/1,73m^2 (Romão, 2004).

O estágio 3 compreende lesão renal moderada, nesta fase, embora os sinais e sintomas da uremia possam estar presentes de maneira discreta, o paciente mantém-se clinicamente bem. Na maioria das vezes, apresenta somente sinais e sintomas ligados à causa básica (lúpus, hipertensão arterial, diabetes mellitus, infecções urinárias, etc.). Avaliação laboratorial simples apresenta, quase sempre, níveis elevados de ureia e de creatinina plasmáticos. O ritmo de filtração glomerular apresenta entre 30 e 59 ml/min/1,73m^2 (Romão, 2004).

O estágio 4 é representado pela insuficiência renal clínica ou severa, nesta fase o paciente já se ressente de disfunção dos rins. Apresenta sinais e sintomas marcados de uremia. Dentre estes a anemia, a hipertensão arterial, o edema, a fraqueza, o mal-estar e os

sintomas digestivos são os mais precoces e comuns. Corresponde à faixa de ritmo de filtração glomerular entre 15 a 29 ml/min/1,73m^2 (Romão, 2004).

E por fim, o estágio 5 que apresenta a fase terminal de insuficiência renal crônica, como o próprio nome indica, falência da função renal, em que os rins perderam o controle do meio interno, tornando-se este bastante alterado para ser incompatível com a vida. Nesta fase, o paciente encontra-se intensamente sintomático. Suas opções terapêuticas são os métodos de depuração artificial do sangue (diálise peritoneal ou hemodiálise) ou o transplante renal. Compreende a um ritmo de filtração glomerular inferior a 15 ml/min/1,73m^2 (Romão, 2004).

O tratamento ideal da DRC é baseado em três pilares de apoio: 1) diagnóstico precoce da doença, 2) encaminhamento imediato para tratamento nefrológico e 3) implementação de medidas para preservar a função renal (Bastos e Kirsztajn, 2011). Portanto, é de grande importância trabalhos que auxiliem na prevenção e diagnóstico precoce da DRC. Pois, dessa forma podem auxiliar no retardamento ou interrompimento da progressão da DRC para os estágios mais avançados, bem como diminuir a morbidade e mortalidade iniciais.

1.1 Fatores de risco da Doença Renal Crônica (DRC)

Geralmente os pacientes em estágio precoces da DRC não são diagnosticados nem tratados oportunamente e, com frequência, apresentam múltiplos fatores de risco que concorrem para aumentar o risco de perda da função renal, desenvolvimento de complicações e morte cardiovascular precoce. As estratégias para melhorar o panorama da DRC requer a identificação dos fatores de risco para lesão renal, permitindo orientar os esforços para o diagnóstico precoce, em populações com alto risco de desenvolver esta doença, e, seguidamente, a aplicação oportuna de intervenções nefroprotetoras para prevenir ou retardar a progressão da lesão renal (Levin, 2001).

Existem vários fatores de risco para a DRC, que pode ser divididos em fatores *predisponentes* ou de *suscetibilidade*, fatores *iniciadores* e fatores *perpetuadores* da lesão renal e de sua progressão, com algumas combinações entre eles (Tabela 2). Os fatores *predisponentes* são aquelas características dos indivíduos que aumentam a probabilidade de desenvolverem DRC. Os fatores *iniciadores* são aqueles que, de forma independente,

podem associar-se ao desenvolvimento de DRC*, e os de progressão ou perpetuadores são* aqueles que podem associar-se à progressão da lesão renal.

Esses fatores, normalmente, interagem como um círculo vicioso sobre a função do rim e provocam a perda progressiva da função renal. A identificação dos fatores de *suscetibilidade* e de *início* são importante para reconhecer as pessoas com maior risco de desenvolver DRC, enquanto que a identificação dos fatores de *progressão* é útil para definir quais pacientes com DRC têm maior risco de progredir até estágios terminais da doença (K/DOQI, 2002; KDIGO, 2012).

Tabela 2. Fatores de risco para DRC.

Predisponentes	**Indicadores**	**Perpetuadores**
Idade avançada (>60 anos)	Doenças renais primárias	Proteinúria
Histórico familiar de DRC	Diabetes mellitus (DM)	PA sistólica >130 mmHg
Grupo étnico (origem hispano)	Hipertensão arterial sistêmica.	Elevada ingestão de proteínas
Gênero masculino	Doenças autoimunes	Mau controle glicêmico
Síndrome metabólica	Nefrotoxinas	Obesidade
Redução da massa renal	AINE	Anemia
Baixo nível socioeconômico e educacional	Aminoglicosídeos	Dislipidemia
Estágios de hiperfiltração	Meios de contraste IV	Tabagismo
Diminuição do número de néfrons	Outros	Hiperuricemia
PA >125/75 mmHg.	Patologias urológicas	Nefrotoxinas
Obesidade	Obstrução urinária	Doença cardiovascular
Ingestão excessiva de proteínas	Litíase urinária	
Anemia	Infecção urinária recorrente	
Aumento da excreção urinária de proteínas	Doenças hereditárias	
Dislipidemia		

AINE: anti-inflamatórios não-esteroides; **IV:** intravenoso; **PA:** pressão arterial.

Fonte: K/DOQI, 2002.

A população geral deveria ser avaliada para determinar se apresenta ou não lesão renal; entretanto, isso nem sempre é factível, principalmente porque não se sabe se a

avaliação da população total teria uma boa relação custo-benefício. Assim, a detecção deve estar orientada aos grupos de pacientes com maior risco de desenvolver lesão renal, por exemplo: hipertensos, diabéticos, idosos, pacientes com doença cardiovascular (DCV) e antecedentes familiares (Metsarinne et al., 2015).

Para o grupo de hipertensos, a hipertensão arterial é comum na DRC, podendo ocorrer em mais de 75% dos pacientes de qualquer idade. Os pacientes diabéticos, por sua vez, apresentam risco aumentado para DRC e doença cardiovascular e devem ser monitorizados frequentemente para a ocorrência da lesão renal. Nos idosos, a diminuição fisiológica da filtração glomerular (FG) e, as lesões renais que ocorrem com a idade, secundárias a doenças crônicas comuns em pacientes de idade avançada, tornam os idosos susceptíveis a DRC. Em pacientes com doença cardiovascular (DCV), a DRC é considerada fator de risco para DCV, alguns estudos demonstram que a DCV se associa independentemente com diminuição da FG e com a ocorrência de DRC. E, familiares de pacientes portadores de DRC, apresentam prevalência aumentada de hipertensão arterial, *Diabetes Mellitus,* proteinúria e doença renal (Barret et al., 1997; Levin et al., 1996 e K/DOQI, 2002).

1.2 Diagnóstico

A DRC é assintomática nos pacientes que se encontram nos estágios iniciais, exigindo que os médicos mantenham um nível adequado de suspeição, especialmente naqueles pacientes com fatores de risco médico ou sociodemográfico para DRC. Alterações funcionais, principalmente na taxa de filtração glomerular (TFG), são um importante componente no diagnóstico e classificação da DRC (Bastos e Kirsztajn 2011).

A TFG é a medida padrão da função renal e a mais facilmente compreendida pelos médicos e pacientes. Definida como a capacidade dos rins de eliminar uma substância do sangue e expressa como o volume de sangue é completamente depurado em uma unidade de tempo. Normalmente, o rim filtra o sangue e elimina os produtos finais do metabolismo proteico, enquanto preserva solutos específicos, proteínas (particularmente albumina) e componentes celulares. Na maioria das doenças renais progressivas, a TFG diminui com o tempo como resultado da diminuição no número total de néfrons ou redução na TFG por néfron, decorrentes de alterações fisiológicas e farmacológicas na hemodinâmica glomerular. A TFG pode estar reduzida bem antes do início dos sintomas e se correlaciona com a gravidade da DRC (K/DOQI, 2002; Levey, 1990; Praxedes, 2004).

A maneira para medir corretamente a TFG é determinar o clearance de substâncias exógenas como a inulina, iotalamato-I125, EDTA (Ácido etilenodiaminotetraacético), DTPA-Tc99m (ácido dietilenotriaminopentácetico marcado com tecnécio99m) ou iohexol. Esses agentes preenchem o critério de marcador ideal de filtração, uma vez que são excretados do corpo via filtração glomerular e não estão sujeitos à secreção e/ou reabsorção quando passam através dos túbulos renais (Steinman, 1989). Como essas substâncias não estão presentes na circulação e, consequentemente, precisam ser infundidas, a medida desses clearances é difícil, requer tempo do paciente e da equipe clínica e tem sido utilizada em geral de forma restrita, para fins de pesquisa ou em condições patológicas específicas nas quais as técnicas de clearance mais simples não oferecem informações suficientes para guiar as decisões médicas (Bastos e Kirsztajn, 2011).

Clinicamente, a TFG tem sido avaliada por meio da mensuração de níveis de substâncias que são normalmente produzidas pelo corpo. A ureia é um marcador endógeno utilizado, porém não é completamente confiável devido seus níveis serem mais vulneráveis a mudanças por razões não relacionadas com a TFG. Uma dieta com alto consumo de proteínas, destruição tecidual, hemorragia gastrointestinal e terapia com corticosteróides podem determinar um aumento nos níveis de ureia plasmática, enquanto uma dieta pobre em proteínas e doença hepática podem levar a uma redução. Em adição, 40-50% da uréia filtrada pode ser reabsorvida pelos túbulos, embora a proporção esteja reduzida na insuficiência renal avançada (Levey, 1990).

A creatinina plasmática era considerada o marcador endógeno cujo perfil mais se assemelhava àquele de uma substância endógena ideal para medir a TFG. A creatinina é quase exclusivamente um marcador ideal, embora a ingestão de carne também possa contribuir levemente para os níveis dessa substância no sangue. Sua geração é relativamente constante durante o dia e diretamente proporcional à massa muscular (Steven, 2006).

A TFG é determinada, nas rotinas clínicas, pela dosagem da creatinina sérica e/ou pela depuração desta pelo rim. A depuração da creatinina pode ser realizada em urina coletada no período de 24 horas, porém a coleta urinária inadequada, seja por falta de compreensão do procedimento ou tipo de atividade do paciente, é um limitador do método. Além disso, as fórmulas ou equações mais utilizadas no Brasil também apresentam limitações, já que foram derivadas e validadas em populações específicas, e apresentam desempenho variável (Sodré e Oliveira, 2014).

Muitas vezes, é difícil determinar a condição de um paciente com base em um único valor laboratorial. A avaliação pode se tornar mais complicada e requerer julgamentos subjetivos quando envolve muitos valores de teste laboratoriais que podem levar a conclusões diferentes. Neste caso, algumas técnicas de análise de dados surgem como uma boa solução que oferecem uma previsão mais exata sobre a saúde do indivíduo (Lenart et al., 2016).

2. Mineração de Dados

A mineração de dados consiste em encontrar padrões úteis em grandes volumes de dados, ou seja, consiste em técnicas de extração de dados, também conhecido como extração de conhecimento, arqueologia de dados, processamento padrão de dados e descoberta de informações. O termo mineração de dados é o mais usado por profissionais da computação, estatísticos e analistas de dados (Melo, 2010).

Atualmente, a mineração de dados vem crescendo grandemente, aumentando significativamente a quantidade de dados legíveis por máquinas na forma de arquivos e bancos de dados (Gregory e Pretto, 2016). Pode-se definir mineração de dados como descoberta de novas informações, que busca, através de métodos automáticos baseados em estatística, a extração de conhecimento de alto nível, partindo de bases de dados reais (Benicasa e Paixão, 2006).

A fase de mineração de dados é uma fase do processo de Descoberta de Conhecimento em Banco de Dados (DCBD). Esta etapa é responsável pela aplicação dos algoritmos que são capazes de identificar e extrair padrões relevantes presente nos dados. O processo DCBD é interdisciplinar, tanto em sua aplicação, quanto das suas fundamentações teóricas. O processo pode ser aplicado a qualquer problema de identificação de padrões em dados e contém fundamentação de diversas áreas como a banco de dados, inteligência artificial, estatística, probabilidade e visualização de dados (Han, 2001; Witten, 2000).

2.1 Aprendizado de máquina

A aprendizagem de máquina (AM) é uma área multidisciplinar que refere-se a capacidade dos computadores em aprender. É considerada uma área multidisciplinar por

utilizar resultados da Inteligência Artificial, estatística, probabilidade, teoria da complexidade computacional, teoria da informação, psicologia, neurobiologia e outros campos. É dito que um computador aprende se o uso de experiências na resolução de conjunto de tarefas melhora com as experiências apresentadas, isto é, ele tem a capacidade de melhorar seu desempenho com o uso das experiências. Sendo assim, a aprendizagem dos computadores acontece quando os mesmos são capazes de formar ações generalizadas a partir de experiências pré-apresentadas. A experiência é apresentada na forma de dados e através de algoritmos é possível aprender, ou seja, construir, identificar padrões gerais presente nos dados. Os testes feitos para medir a capacidade do computador em aprender são feitos através da observação de uma nova experiência e a comparação deste resultado com o uso de experiências anteriores (Michalski et al., 1986).

De acordo com Melo (2010), toda técnica de mineração passa por um processo chamado de treinamento. A fase de treinamento tem este nome por ser um processo de apresentação dos dados processados para o algoritmo de mineração, cujo objetivo é identificar, ou seja, "aprender" as características ou padrões úteis ao objetivo do processo de descoberta de conhecimento. Os dados são processados para a realização do aprendizado, sendo que, após o aprendizado ter sido realizado, é aplicada uma avaliação, onde se podem verificar medidas estatísticas dos resultados alcançados. A avaliação do algoritmo treinado deve ser realizada utilizando dados não vistos pelo algoritmo, os dados devem ser inéditos.

A utilização de dados inéditos fornecerá medidas realistas sobre o desempenho do algoritmo, pois os mesmos serão feitos a partir de dados não vistos na fase de treinamento. A fase de avaliação será realizada de forma correta caso a divisão do conjunto de dados seja realizada. O conjunto deve ser dividido em dados de treinamento e de teste. Às vezes, é necessário dividir o conjunto de dados em 3 diferentes conjuntos: treinamento, validação e teste. O conjunto de validação é utilizado para ajustar valores dos parâmetros de alguns algoritmos e ao mesmo tempo uma boa generalização. Quando o conjunto de dados é dividido em dois, geralmente a divisão é de 70% para o conjunto de treinamento e 30% para o conjunto de testes. E, quando o conjunto será divido em 3 (três), usa-se a proporção 70% para treinamento, 20% para validação e 10% para testes (Melo, 2010).

Existem várias aplicações para Aprendizado de Máquina, o mais significativo deles é a mineração de dados preditivos. Muitos exemplos em qualquer conjunto de dados usado por algoritmos de aprendizado de máquina são representados usando o mesmo conjunto de dados. Os recursos podem ser contínuos, categoriais ou binários. Se as instâncias são dadas

com rótulos conhecidos, então o aprendizado é chamado supervisionado, em contraste com o aprendizado não supervisionado, onde as instâncias são sem rótulo (Jain et al., 1999).

Algoritmos de aprendizado de máquina supervisionados podem ser aplicados para prever os eventos futuros com a ajuda do que foi aprendido no passado para novos dados usando exemplos rotulados. Primeiro, o conjunto de dados de treinamento conhecido é analisado, com o qual o algoritmo de aprendizado produz uma função inferida para fazer previsões sobre os valores de saída. Após o treinamento suficiente, o sistema é capaz de fornecer alvos para quaisquer novas entradas.

Algoritmos de aprendizado supervisionado são baseados em padrões de pares de entrada-saída. Esses algoritmos têm como objetivo prever valores de saída com base em determinados valores de entrada e focam principalmente em problemas de classificação e a regressão (Madhura Rambhajani et al., 2015).

2.2 Classificação binária

Na área de aprendizagem automática, a classificação consiste na atribuição de um valor (classe) a uma determinada instância. Se o domínio dos valores a atribuir for constituído apenas por dois elementos, trata-se então de um problema binário. Se tiver mais do que dois elementos se têm então um problema multi-classe.

De acordo com Almeida (2010), as classificações binárias são frequentemente realizadas pelo uso de funções g: x C $R^m \rightarrow$R com a seguinte estratégia: as amostras são designadas para a classe positiva, se $g(x) \geq 0$, e caso contrário, para a classe negativa.

A superfície de decisão será representada por um hiperplano na forma: g(x)=(wx)+b=0, onde w e R^m é o valor de pesos, b e R é o bias.

Assim pode-se aplicar a seguinte estratégia de decisão (Burges e Christopher 1998):

$(w^t x_i) + b \geq 0$ para y= +1

$(w^t x_i) + b < 0$ paray=- 1

Para descrever o lugar geométrico dos hiperplanos separados utiliza-se a seguinte forma canónica onde o vector de peso w e o viés b são novamente escalados de tal maneira a atender as desigualdades (France et al., 2004):

$(w^t x_i) + b \geq +1$ para y= +1

$(w^t x_i) + b \leq +1$ para= -1

Para um dado vector de pesos *w* e viés *b*, a separação entre g(x)=(wtx)+b=0, e o dado de entrada mais perto é chamado de margem de separação denotada por ρ. Sempre que for positivo um ρ> 0 existirão infinitos hiperplanos, dentre os quais se busca um hiperplano particular em que a margem de separação ρ é maximizada (Almeida, 2010).

3. Algoritmos de aprendizado de máquina

3.1 KNN

Muito usado no método de aprendizado supervisionado, o algoritmo K-Nearest Neighbor (KNN) utiliza uma metodologia bem simples de classificação que consiste na identificação de grupos de indivíduos com características similares e seu posterior agrupamento (clustering) (Rosa, 2003). A principal ideia do KNN é encontrar o número (K) de exemplares rotulados mais próximos do exemplo não classificado e, de acordo com a base no rótulo desses exemplares mais próximos, é tomada a decisão relativa à classe do exemplo não rotulado (Ferrero, 2009).

Lemos (1999) afirma que o método KNN usa como métrica de classificação a similaridade entre amostras. Esse método utiliza o conceito de distância entre amostras, considerando-se uma amostra um vetor (linha) contendo várias colunas (variáveis), ou seja, uma medida de similaridade entre pontos poderia ser baseada nas várias distâncias entre um ponto a ser classificado em uma dada classe e pontos de classes conhecidas.

O valor de K (o número de vizinhos mais próximos que serão considerados pelo algoritmo) é definido pelo usuário, sendo recomendada a escolha de um número ímpar para evitar um empate na classificação. Se duas classes A e B possuem vários pontos em seus domínios, dado um ponto desconhecido X, este ponto será classificado em função da quantidade de pontos cujas distâncias forem as menores possíveis em relação às classes A e B (Rosa, 2003).

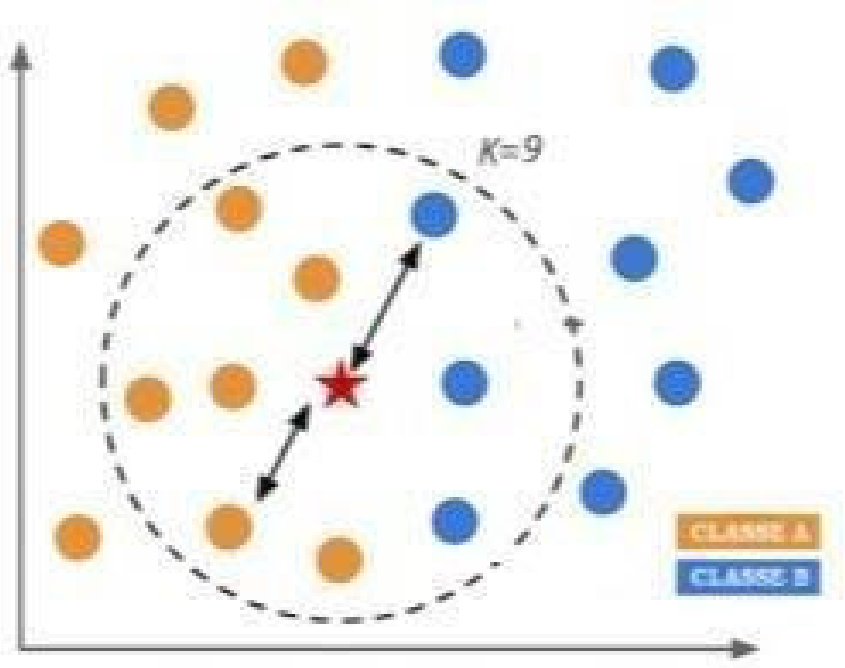


Figura 1: Dinâmica do KNN.
Fonte: Rosa, 2003.

O cálculo da distância da amostra desconhecida em relação às amostras conhecidas é realizado pela raiz quadrada do somatório do quadrado da diferença de cada uma das *m* variáveis (colunas) da amostra desconhecida *(vd)* em relação às *m* variáveis (colunas) de cada uma das *i* (i = 1... n) amostras conhecidas *(vc):*

$$dist_i = \sqrt{\sum_{j=1}^{j=m} (vdj - vc_ij)^2}$$

Considerando-se *i* amostras conhecidas e duas classes 0 e 1, o algoritmo de classificação pelo método KNN pode ser descrito da seguinte forma (Boente; Lemos; Rosa, 2009):

Algoritmo KNN
Início
Ler n amostras de classes conhecidas *(ac)*
Ler amostra de classe desconhecida *(ad)*
Iniciar *i = 1*
Repita
Computar a distância de ad para *aci*
Atribuir a distância e a classe de *aci* ao vetor de distâncias

```
        Incrementar i = i + 1
Até (computar todas as distâncias, isto é, i > n)
Ordenar o vetor de distâncias por ordem de distância
Selecionar as K primeiras posições do vetor de distâncias
Iniciar i = 1
Iniciar classe 0 = 0
Iniciar classe 1 = 0
Repita
    Comparar a classe
    Se (classe igual a 0)
        Incrementar classe 0 = classe 0 + 1
    Senão
        Incrementar classe 1 = classe 1 + 1
    Fim Se
    Incrementar i = i + 1
Até (computar todas as K amostras, isto é, i > K)
Fim do Algoritmo
```

3.2 SVM

Desenvolvida por Vapnik, em 1995, com intuito de resolver problemas de classificação, as Máquinas de Vectores Suporte, do inglês *Support Vector Machines* – SVMs, é uma técnica de Aprendizagem de Máquina que vem recebendo grande atenção nos últimos anos. As SVMs vêm sendo utilizadas em diversas tarefas de reconhecimento de padrões, obtendo resultados superiores aos alcançados por técnicas similares em várias aplicações.

As SVMs apresentam algumas características principais que tornam seu uso atrativo, como uma boa capacidade de generalização, significa a medida por sua eficiência na classificação de dados que não pertençam ao conjunto utilizado em seu treino. Os classificadores gerados por uma SVM em geral alcançam bons resultados de generalização; Robustez em grandes dimensões, geralmente as SVMs são robustas diante de objetos de grandes dimensões (imagens); Convexidade da função objetiva, em que a aplicação das SVMs implica na optimização de uma função quadrática, que possui apenas um mínimo global. Esta é uma vantagem sobre, por exemplo, as Redes Neurais Artificiais, em que há a presença de mínimos locais na função objetiva a ser minimizadas. E for fim, uma teoria

bem definida, as SVMs possuem uma base teórica bem estabelecida dentro da Matemática e Estatística (Smola et al., 1999b).

As SVMs são originalmente utilizadas para classificação dos dados em duas classes distintas. Para o problema de classificação binária, o algoritmo SVM funciona atribuindo duas classes e um conjunto de pontos que pertencem a essas classes, uma SVM determina o hiperplano óptimo que separa os pontos de forma a colocar o maior número de pontos da mesma classe do mesmo lado, enquanto maximiza a distância de cada classe a esse hiperplano. A distância de uma classe a um hiperplano é a menor distância entre ele e os pontos dessa classe são chamados de margem de separação. O hiperplano gerado pela SVM é determinado por um subconjunto dos pontos das duas classes, a que se dá o nome vectores suporte (Chaves, 2004).

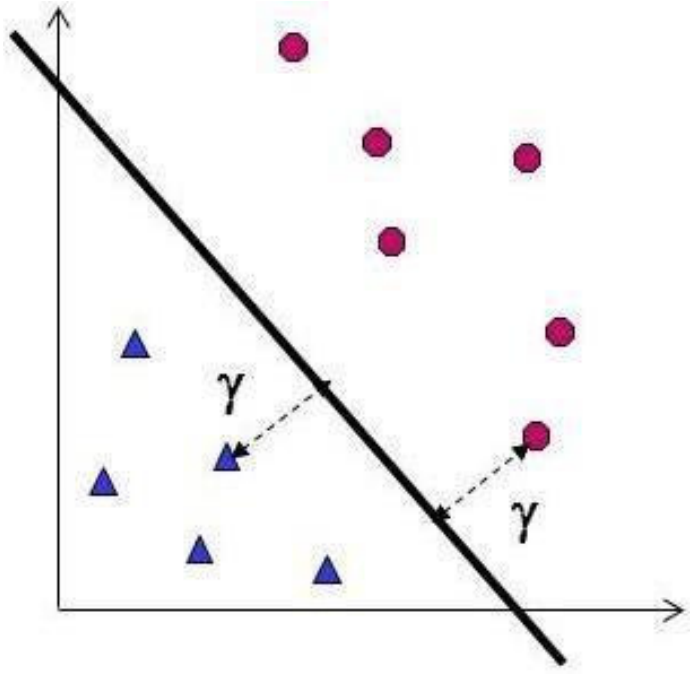


Figura 2: Hiperplano gerado pela SVM

Fonte: Chaves, 2004.

3.3 Árvore de decisão

As Árvores de Decisão utilizam a estratégia *dividir-e-conquistar* ("*divide-and-conquer*"), onde as árvores são construídas utilizando-se de apenas alguns atributos. As Árvores de Decisão são uma das técnicas de aprendizado de máquina ("*machine learning*"), onde um problema complexo é decomposto em subproblemas mais simples, sendo que a mesma estratégia é aplicada a cada sub-problema (Gama, 2002).

Quinlan, da Universidade de Sidney, é considerado o "pai das Árvores de Decisão". A sua contribuição foi a elaboração de um novo algoritmo chamado *ID3*, desenvolvido em 1983. O *ID3* e suas evoluções (*ID4*, *ID6*, *C4.5*, *See 5*) são algoritmos muito utilizados para gerar Árvores de Decisão. O atributo mais importante é apresentado na árvore como o primeiro nó, e os atributos menos importantes, segundo o critério utilizado, são mostrados nos nós subsequentes. As vantagens principais das Árvores de Decisão são que elas "tomam decisões" levando em consideração os atributos que são considerados mais relevantes, segundo a métrica escolhida, além de serem compreensíveis para as pessoas.

Os classificadores baseados em árvores de decisão procuram encontrar formas de dividir sucessivamente o universo em vários subconjuntos (criando para tal nós contendo os testes respectivos) até que cada um deles contemple apenas uma classe ou até que uma das classes demonstre uma clara maioria não justificando posteriores divisões (gerando nessa situação uma folha contendo a classe maioritária). Como é evidente, a classificação consiste apenas em seguir o caminho ditado pelos sucessivos testes colocados ao longo da árvore até que seja encontrada uma folha que conterá a classe a atribuir ao novo exemplo (Fonseca, 1994).

O algoritmo de Random Forest (RF) introduzido por Breiman (2001) é um termo geral para métodos de ensaio utilizando classificadores do tipo árvore. A RF constrói uma grande quantidade de árvores de decisão para fora do sub-conjunto de dados a partir de um único treinamento definido. Tal treinamento é realizado usando bagging (um meta-algoritmo para melhorar a classificação e a regressão de modelos de acordo com a estabilidade e a precisão da classificação) (Breiman, 2001).

3.4 Naive Bayes

Naive Bayes é um método de aprendizado de máquina supervisionado usado para classificação que considera as variáveis como independentes, por esse motivo é tido como ingênuo (naive em inglês). É um bom método, simples de compreender e de fácil implementação, frequentemente aplicado em processamento de linguagem natural e diagnósticos médicos. Esse método pode ser usado quando os atributos que descrevem as instâncias forem condicionalmente independentes dada à classificação (Chakrabarti, 2002).

O classificador Naive-Bayes funciona de acordo com o teorema de Bayes, criado por Thomas Bayes no século XVIII, considerado o mais eficiente na precisão e rotulação de novas amostras. Trata sobre probabilidade condicional, isto é, a probabilidade de o

evento A ocorrer, dado o evento B. O teorema de Bayes pode ser expresso pela seguinte fórmula (Bayes, T. e Price, R., 1763):

$$P(A \mid B) = P(B \mid A) \times P(A) / P(B)$$

A probabilidade do evento A ocorrer dado o evento B é igual probabilidade do evento B ocorrer dado o evento A vezes a probabilidade do evento A sobre a probabilidade do evento B.

Essa mesma lógica pode ser utilizada para o cálculo das probabilidades necessárias para problemas de classificação, em que basta substituir um dos argumentos da fórmula pela classe a ser calculada.

$$P(classe \mid B) = P(B \mid classe) \times P(classe) / P(B)$$

A probabilidade de pertencer a classe escolhida dado o atributo B é igual probabilidade do atributo B ocorrer dado que ele pertence a classe escolhida vezes a probabilidade de ocorrer a classe sobre a probabilidade do evento B.

De acordo com Pardo e Nunes (2002), para calcular a classe mais provável da nova instância, calcula-se a probabilidade de todas as possíveis classes e, no fim, escolhe-se a classe com a maior probabilidade como rótulo da nova instância. Em termos estatísticos, isso é o mesmo do que maximizar a P (classe | $a_1 \ldots a_n$). Para tanto, deve-se maximizar o valor do numerador P ($a_1 \ldots a_n$ | classe) × P(classe) e minimizar o valor do denominador P ($a \ldots a_n$). Como o denominador P ($a \ldots a_n$) é uma constante, pois não depende da variável classe que se está procurando, pode-se anulá-lo no Teorema de Bayes, na qual se procura a classe que maximize o valor do termo P (classe | $a_1 \ldots a_n$) = P ($a_1 \ldots a_n$ | classe) × P (classe), resultando na formula:

$$argmax\ P(classe \mid a_1 \ldots a_1) = argmax\ P(a_1 \ldots a_1 \mid classe) \times P(classe)$$

A suposição "ingênua" que o classificador Naive-Bayes realiza, é que todos os atributos $a_1 \ldots a_n$ da instância que se quer classificar sejam independentes. Sendo assim, o cálculo do valor do termo P ($a_1 \ldots a_n$ | classe) reduz-se ao simples cálculo de P (a | classe) × … ×P (a_n | classe). Assim, a fórmula final utilizada pelo classificador é:

$$argmax\ P(classe \mid a_i \ldots a_1) = argmax \prod_i P(a_i \mid classe) \times P(classe)$$

Entretanto, mesmo conhecendo que a suposição de independência dos atributos de uma instancia seja falsa, na maioria dos casos o classificador Naive Bayes consegue

resultados bem satisfatórios, sendo que os atributos devem ser realmente independentes para que o classificador possa fornecer uma solução favorável. O cálculo da classe da nova instancia será a probabilidade de todas as possíveis classes e a classe com maior probabilidade será escolhida (Pardo e Nunes, 2002).

4. WEKA

O nome Weka é de um pássaro que não voa e tem natureza inquisitiva. O pacote de software **Weka** (*Waikato Environment for Knowledge Analysis*) começou a ser escrito em 1993, usando Java, na Universidade de *Wakato*, Nova Zelândia sendo adquirido posteriormente por uma empresa no final de 2006. O Weka encontra-se licenciado ao abrigo da General *Public License* sendo portanto possível estudar e alterar o respectivo código fonte (Witten et al., 1999).

O Weka é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Contém ferramentas para preparação de dados, classificação, regressão, agrupamento, mineração de regras e visualização. Está sendo comumente usado na academia para pesquisas e aplicações deestudos. Atualmente, é pertencente ao pacote Pentaho (Bouman e Dongen, 2009).

O principal objetivo do Weka é agregar algoritmos provenientes de diferentes abordagens/paradigmas na subárea da inteligência artificial dedicada ao estudo da aprendizagem automática. Essa subárea pretende desenvolver algoritmos e técnicas que permitam a um computador "aprender" (no sentido de obter novo conhecimento) quer indutiva quer dedutivamente (Almeida, 2010).

Os algoritmos de *Machine Learning* podem ser aplicados diretamente a partir do seu próprio código Java. A ferramenta trabalha com dados no formato .csv, mas possui um formato próprio, em que é possível, através de meta-dados, delimitar informações e realizar um pré-processamento destes. A ferramenta possui uma interface fácil e conta com um conjunto de dados para testes, pesquisas e estudos. Além de ser uma ferramenta gratuita, pode ser usada para fins comerciais (Viana, 2012).

Figura 3. Logotipo do WEKA
Fonte: Viana, 2012

5. Validação dos Resultados

Existem várias opções para avaliar o desempenho dos algoritmos classificadores. Nos casos de aprendizado supervisionado, é muito útil a construção de uma matriz de confusão que permite a visualização do desempenho de um classificador. Uma matriz de confusão ordena todos os casos do modelo em categorias, determinando se o valor previsto corresponde ao valor real.

Na matriz de confusão, cada linha da matriz representa os valores previstos do modelo (o que o algoritmo classificou), enquanto as colunas representam os valores reais. O número de acertos de cada classe se localiza na diagonal principal da matriz, enquanto os demais elementos da matriz representam erros de classificação. A matriz de confusão de um classificador ideal possui todos os elementos fora da diagonal principal iguais a zero, uma vez que ele não comete erros.

Na Tabela 3 mostra uma matriz de confusão binária (2 classes), onde os diagnósticos corretos e incorretos são armazenados. Os conceitos das siglas VP, FP, VN e FN são definidos da seguinte maneira (Cerri, 2010):

- VP (verdadeiros positivos): define os casos em que o indivíduo foi corretamente classificado como doente.
- VN (verdadeiros negativos): define os casos em que o indivíduo estava saudável e que foram corretamente classificados.
- FP (falsos positivos): denota os casos em que o indivíduo foi classificado erroneamente como doente, mas estava saudável.
- FN (falsos negativos): denota os casos em que o indivíduo foi classificado erroneamente como saudável, mas estava doente.

Tabela 3. Matriz de confusão de 2 classes em relação ao classificador e a realidade.

	Realidade		
Classificador	Saudável	Doente	
Saudável	(VN)	(FN)	Total de Saudáveis Classificados= VN + FN
Doente	(FP)	(VP)	Total de Doentes Classificados=FP+VP
	Total de Saudáveis Reais= VN + FP	Total de Doentes Reais= FN + VP	Total de amostras= VN+FN+FP+VP

Partindo da matriz de confusão podem-se definir as equações de sensibilidade, especificidade e exatidão ou acurácia, que quanto mais próximo de 1, melhor é o desempenho.

A sensibilidade ou taxa positiva verdadeira (TPV) definida como a probabilidade de classificar corretamente um paciente, com o valor definido como positivo. Indica a capacidade do estimador para classificar como positivos aqueles casos que realmente são positivos. Ou seja, a sensibilidade caracteriza a capacidade do estimador para detectar a doença em pacientes doentes (proporção de pacientes doentes corretamente identificados). O valor dessa classificação concorda com o estado real do paciente, expressa da seguinte forma (Asucion e Newman, 2007):

$$\text{Sensibilidade} = \frac{VP}{VP+FN}$$

A especificidade ou taxa negativa verdadeira (TNV) expressa a capacidade do estimador para classificar como negativos aqueles casos que realmente são negativos. É a capacidade do estimador para detectar a ausência da doença em pacientes saudáveis (proporção de pacientes saudáveis corretamente identificados) expressa como (Quilan, 1993):

$$\text{Especificidade} = \frac{VN}{VN+FP}$$

A acurácia avalia a efetividade geral do algoritmo classificador, expressa como (Chiu et al., 2012):

$$\text{Acurácia} = \frac{VP+VN}{VN+FN+VP+FP}$$

6. Referências

Almeida, E.D. Algoritmos de Classificação Com a Opção de Rejeição. (Dissertação de Mestrado) Porto: Faculdade De Engenharia Da Universidade Do Porto, 19 p. 2010.

Asuncion, A. and Newman, D. J. UCI Machine Learning Repository [Online]. Disponível: http://www.ics.uci.edu/mlearn/ MLRepository.html. Acessado em 15/08/2018. 2007.

Barrett, B.J., Parfrey, P.S., Morgan J.B.P, Fine, A., Goldstein, M.B., Handa, S.P., Jindal K.K., Kjellstrand, C.M., Levin, A., Mandin, H., Muirhead, N., Richardson, R.M. Prediction of early death in end-stage renal disease patients starting dialysis. *Am J Kidney Dis* 29:214-22. 1997.

Bastos., M.G e Kirsztajn, G.M. Doença renal crônica: importância do diagnóstico precoce, encaminhamento imediato e abordagem interdisciplinar estruturada para melhora do desfecho em pacientes ainda não submetidos à diálise. *J. Bras. Nefrol.* [online]. vol.33, n.1, pp.93-108. 2011.

Baumgarten, M., Gehr, T. Chronic Kidney Disease: Detection and Evaluation American Family Physician, Number 10. 2011.

Bayes, T. e Price, R. An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S., Philosophical Transactions of the Royal Society of London 53 (0): 370-418. doi:10.1098/rstl.1763.0053. 1763.

Benicasa, A. X., Paixao, R.S. Mineração de dados como ferramenta para descoberta de conhecimento. Macapá: Faculdade Seama/Ministério Público do Estado do Amapá. 2006.

Boente, A.N.P., Lemos C.A.A., Rosa, J.L.D.A. Metodologia knn-fuzzy: uma abordagem da classificação de dados por similaridade. *XIV SEGeT - Simpósio de Excelência em Gestão e tecnologia,* Resende, Rio de Janeiro. 2009.

Bonato, P., Sherrill, D.M., Standaert, D.G., Salles, S.S., Akay, M. Data mining techniques to detect motor fluctuations in Parkinson's disease. In Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE, Vol. 2, pp. 4766-4769. IEEE. 2004.

Bouman, R., Dongen, J.V. Pentaho solutions: business intelligence and data warehousing with Pentaho and MySQL. Hoboken: Wiley. 2009.

Breiman, L. Random forests, Machine Learning 45(1): 5–32. 2001.

Burges, Christopher J.C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2, 121- 167. 1998.

Care, F.M. Fresenius Medical Care Annual Report 2013. Fresenius Med Care, 294. 2014.
Cerri, Ricardo. Técnicas de classificação hierárquica multirrótulo. 241f. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional) - Universidade de São Paulo, São Carlos. 2010.

Chakrabarti, S. Mining the Web: Discovering knowledge from hypertext data. [S.l.]: Elsevier. 2002.

Chaves A. Extração de Regras de Fuzzy para Máquinas de Vector Suporte (SVM) para Classificação em Múltiplas Classes, p. 13-16. 2004.

Chiu, R.K., Chen, R.Y., Wang, S.A., Jian, S.J. Intelligent systems on the cloud for the early detection of chronic kidney disease. In Machine Learning and Cybernetics (ICMLC), 2012 International Conference on Vol. 5, pp. 1737-1742. IEEE. 2012.

Draibe, S.A. Panorama da Doença Renal Crônica no Brasil e no mundo. UNASUS/UFMA - São Luís. 2014.

Ferrero, C. A. Algoritmo KNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia.

Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, ICMC-USP, São Carlos, SP. 2009.

Fonseca, J.M.M.R. Indução de árvores de decisão: HistClass - proposta de um algoritmo não paramétrico. 140p. Dissertação (Mestrado em Engenharia Informática) - Departamento de Informática, Universidade Nova de Lisboa, Lisboa. 1994.

Franc, V., Hlavac, V. Statistical Pattern Recognition Toolbox User's Guide", 24 Junho. 2004.

Gama, J. Árvores de Decisão, 2000. Disponível em: http://www.liacc.up.pt/~jgama/Masters/ECD1/Trees.html. Acesso em: 14 ago. 2019.

Ghannad-Rezaie, M. e Soltanian-Zadeh, H. Interactive knowledge discovery for temporal lobe epilepsy. INTECH Open Access Publisher. 2008.
Gregory, G.; Pretto, F. Mineração de Dados para Descoberta de Conhecimento em Dados de Promoção à Saúde. *Revista Destaques Acadêmicos*. Vol. 8, n. 4, p. 51-65. 2016.

Han, J.; Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann. 2001.

Ilayaraja, M., Meyyappan, T. Mining medical data to identify frequent diseases using Apriori algorithm. In Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on pp. 194-199. IEEE. 2013.

Jain, A. K.; Murty, M. N.; Flynn, P. J. Data clustering: a review. *ACM computing surveys* (CSUR), v. 31, p. 264-323. 1999.

K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification and stratification. *Am J Kidney Dis.* (Suppl 2):S1-S246. 2002.

Kidney Disease Improving Global Outcomes: KDIGO. Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. (Suppl 3): 1-150. 2012.

Kunwar, V et al. Chronic Kidney Disease analysis using data mining classification techniques. 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence). 2016.

Lakshmi, K.R., Nagesh, Y., Veerakrishna, M. Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability. International Journal of Advances in Engineering & Technology (IJAET), 7 (1), 242-254. 2014.

Lee, H.G., Noh, K.Y., RYU, K.H. A data mining approach for coronary heart disease prediction using HRV features and carotid arterial wall thickness. In BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on , Vol. 1, pp. 200-206. IEEE. 2008.

Lemos, C.A.A. Uma Implementação Fuzzy de Algoritmo para Classificação de Dados. Rio de Janeiro, Dissertação de Mestrado - Departamento de Engenharia Civil, UFRJ/COPPE/PEC, Rio de Janeiro. 1999.

Lenart, M., Mascarenhas, N., Xiong, R., Flower, A. Identifying Risk of Progression for Patients with Chronic Kidney Disease Using Clustering Models. *IEEE Systems and Information Engineering Design Conference* (SIEDS '16) p. 221-226. 2016.

Levey, A.S. Measurement of renal function in chronic renal disease. Kidney Int; 38:167 - 84. 1990.

Levin, A. Identification of patients and risk factors in chronic kidney disease-evaluating risk factors and therapeutic strategies. *Nephrol Dial Transplant*. (Suppl 7): 57-60. 2001.

Levin, A., Singer, J., Thompson, C.R., Ross, H., Lewis, M. Prevalent left ventricular hypertrophy in the predialysis population: identifying opportunities for intervention. Am J Kidney Dis. Vol. 27, p. 347-54. 1996.

Melo, D.M. Desenvolvimento de uma Metodologia Para Criação de Sistemas de Previsão Criminal em Regiões Metropolitanas Brasileiras – Caso De Uso: Região Metropolitana De Fortaleza. Dissertação (Mestrado) - Universidade Estadual do Ceara, UFC. 2010.

Metsarinne K., Broijersen A., Kantola, Niskanen L., Rissanen A., Appelroth T., Pöntynen N., Poussa T., Koivisto V., Virkamäki A. STages of NEphropathy inType 2 Diabetes Study Investigators. High prevalence of chronic kidney disease in Finnish patients with type 2 diabetes treated in primary care. Prim. Care Diabetes. Vol. 9, p.31–38. 2015.

Michalski, R., Carbonell, J., Mitchell, T. Machine learning: An artificial intelligence approach. [S.l.]: Morgan Kaufmann Pub, 1986.

Ministério da saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância de Doenças e Agravos não Transmissíveis e Promoção da Saúde. Vigitel Brasil 2014: vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico – Brasília: Ministério da Saúde, 2015.

National kidney foundation guidelines. *Am J Kidney Dis*. 43 (Suppl 1):S1-S290. 2004.

National kidney foundation. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney Inter Suppl, v. 3, n. 1, p. 1-150, jan. 2013. Disponível em: http://goo.gl/gZcgU5. Acesso em: 5 out. 2016.

Pal, D., Chakraborty, C., Mandana, K.M. Data mining approach for coronary artery disease screening. In Image Information Processing (ICIIP), 2011 International Conference on. pp. 1-6. IEEE. 2011.

Pardo, T.A.S., Nunes, M. das G.V. Aprendizado bayesiano aplicado ao processamento de línguas naturais. Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC-USP. 2002.

Polito, M.G. Complicações clínicas e condutas na doença renal crônica. Universidade Federal do Maranhão. UNASUS/UFMA (Org.). São Luís. 2014.

Praxedes J.N. Diretrizes sobre hipertensão arterial e uso de anti-hipertensivos na doença renal crônica. *J Bras Nefrol*. 26:44-6. 2004.

Quinlan, J. C. C4.5: Programs for machine learning. San Mateo: Morgan Kaufmann, 1993. 302p. 1993.

Rajan, J.R., Chelvan, C.C. A survey on mining techniques for early lung cancer diagnoses. In Green Computing, Communication and Conservation of Energy (ICGCE), 2013 International Conference on pp. 918-922. IEEE. 2013.

Rambhajani, M., Wyomesh, D., and Neelam P. A survey on implementation of machine learning techniques for dermatology diseases classification." International Journal of Advances in Engineering & Technology 8.2. 2015.

Romão Júnior, J.E. Doença renal crônica: definição, epidemiologia e classificação. *J Bras Nefrol*. 26(3). 2004. Disponível em: http://www.jbn.org.br/26-31/v26e3s1p001.pdf.

Rosa, J.L.A. Classificação de Dados Através da Otimização do Método KNN Fuzzy em Ambiente de Computação Paralela. XIII, 97 p. Tese (Doutorado em Ciências em Engenharia Civil / Sistemas Computacionais) - UFRJ/COPPE, Rio de Janeiro. 2003.

Sesso R., Lopes, A.A, Thomé A.S., Bevilacqua, J.L., Romão Junior J.E., Lugon Jr. Relatório do Censo Brasileiro de Diálise. *J Bras Nefrol*. 30:233-8. 2008.

Sesso, R.C., Lopes, A.A., Thomé, F.S., Lugon, J.R., Martins, C.T. Brazilian Chronic Dialysis Census. *J Bras Nefrol*, 38, pp. 54-61. 2016.

Silva, S.B., Caulliraux, H.M., Araújo, C.A., Rocha, E. Cost comparison of kidney transplant versus dialysis in Brazil. Cad. Saúde Pública, Rio de Janeiro, 32(6):e00013515, jun, 2016.

Smola, A. J., Barlett, P., Schölkopf, B., and Schuurmans, D. (1999). Advances in Large Margin Classifiers. The MIT Press (http://www.kernelmachines.org/nips98/lmc-book.pdf).

Sodré, A.N; Oliveira, M.I.A. Estimativa da Taxa de Filtração Glomerular Através de Fórmulas. News Lab - edição 122. 2014.

Sodré, AB et al. Evaluation of renal function and damage: a laboratorial challenge. J Bras Patol Med Lab. v. 43, n. 5, p. 329-33, 2007.

Srinivas, K., Rao, G.R., Govardhan, A. Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In Computer Science and Education (ICCSE), 2010 5th International Conference on pp. 1344- 1349. IEEE. 2010.

Steinman, T.I, Perrone, R.D, Hunsicker, L.G, MDRD Study Group. GFR determination in chronic renal failure by 3 radionuclide markers and inulin: coeficient of variation of the methods (abstract). Kidney Int 1989; 35:201.

Stevens, L.A, Coresh J., Greene T., Levey A.S. Assessing kidney function - Measured and estimated glomerular filtration rate. *New Engl J Med*. 354:2473-83. 2006.

Su, J.L., Wu, G.Z., Chao, I.P. The approach of data mining methods for medical database. In Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE Vol. 4, pp. 3824-3826. IEEE. 2001.

Vapnik, V.N. The nature of statistical learning theory. New York, NY, USA: Springer-Verlag New York, Inc., 1995. ISBN 0387945598. Disponível em: <https://dl.acm.org/citation.cfm?id=211359>. Acessado: 10/05/2018.

Vecina Neto, G; Malik, A.M. O futuro dos serviços de saúde no Brasil. In: Vecina Neto G, Malik AM, organizadores. Gestão em saúde. Rio de Janeiro: Editora Guanabara Koogan, p. 351-7, 2012.

Viana, T.A.M.N. Uma análise comparativa sobre ferramentas de mineração de dados adotadas na academia e na indústria. *Revista Tecnologia*, 33(1), article no. 9. 2012.

Witten, I.; Frank, E. Data Mining – Practical Machine Learning Tools. Morgan Kaufmann, 2000.

Witten, I.H, Frank, E., Trigg, L., Hall, M., Holmes, G. and Cunningham, S.J. Weka: Ferramentas e técnicas práticas de aprendizado de máquina com implementações Java. (Documento de trabalho 99/11). Hamilton, Nova Zelândia: University of Waikato, Departamento de Ciência da Computação. 1999.

World health organization. Noncommunicable diseases country profiles 2011. Geneva: WHO Obesity Technical Report Series, n. 284, 2011.

Xing, Y., Wang, J., Zhao, Z., Gao, Y. Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In Convergence Information Technology, 2007. International Conference on. pp. 868-872. IEEE, 2007.

Xun, L., Xiaoming, W., Ningshan, L., Tanqi, L. Application of radial basis function neural network to estimate glomerular filtration rate in Chinese patients with chronic kidney disease. In Computer Application and System Modeling (ICCASM), 2010 International Conference on Vol. 15, pp. V15-332. IEEE. 2010.

Capítulo I

ARTIFICIAL INTELLIGENCE IN PREDICTING CHRONIC KIDNEY DISEASE

Vanessa D. Martins [AB], Viviane S. Ferreira [A], Marta O. Barreiros [A], Ilka P. Belfort [A], Allan K. Barros [A]

[A] Department of Electrical Engineering, Laboratory for Biological Information Processing (PIB), Federal University of Maranhao (UFMA) Sao Luis-MA, Brazil.

[B] Corresponding author. Email: vanessa.duartema@gmail.com

ABSTRACT

The prediction of the future is becoming an increasingly easy and discussed task in the literature, especially in healthcare, with predictive analyzes of medical data using the machine learning, which evolved after the development of new informed technologies that originated multiple search fields. Much dedication is fulfilled periodically to deal with an explosion of medical data, to gain knowledge of it, to predict disease, and to anticipate healing. In order to extract useful knowledge and aid decision-making, researchers are increasingly applying technical innovations, including database analysis, predictive analysis, machine learning and learning algorithms. Thus, aimed to conduct a review of literature on the use of Artificial Intelligence in the prediction of Chronic Kidney Disease.

Key Words: Machine learning, Classifiers, Prediction, Chronic Kidney Disease.

INTRODUCTION

Artificial intelligence (IA) is the area of computer science that aims to simulate the processes of human thinking, having the ability to learn, store knowledge and solve problems (Krittanawong, 2017). The use of IA techniques has become widely accepted in medical applications, showing a growing number of medical devices available on the market, along with a fast pace of medical journal publishing, with more than 500 academic publications each year (Gant, 2001).

The medical field makes an extreme contribution to the magnitude of medical data because of some innovations in the field, such as cloud computing, laparoscopic surgery, and robotic surgery, which replaced classical surgery (Gabriel, 2010). There are also

intelligent applications or software that can analyze body signals using integrated sensors for monitoring purposes, as well as technologies that support new biological, behavioral and environmental data collection methods. These include sensors that monitor phenomena with high precision (Steve, 2014).

All of these innovations come from the grandiosity of medical data by multiplying electronic medical data sources and records containing diagnostic images, laboratory results, and biometric data (Steve, 2014; Weil, 2014 and Groves, 2013). Researchers have deduced that this explosion of medical data has the potential to improve point-of-care decisions.

The physician will be able to extract relevant knowledge for each patient, which provides better decisions and outcomes (Huang, 2015). There are many classification and prediction algorithms that can be applied to predict various diseases such as breast cancer, heart disease, motor neuron, diabetes, chronic kidney disease, among others. There is ongoing research work using Artificial Intelligence techniques in the field of medical diagnosis for these diseases (Boukenze, 2017).

Kidney Disease is currently considered a global health problem as it affects millions of people worldwide. This disease is considered dangerous if not treated immediately in time, and can be fatal. If doctors have a good tool that can identify patients who are likely to have kidney disease in advance, they can start treatment faster, thus avoiding complications of the disease (Levey, 2012 and Wang, 2016). Thus, the objective was to perform a literature review of research conducted with Artificial Intelligence in the prediction of Chronic Kidney Disease (CKD).

Research using Artificial Intelligence

Ho *et al.* (2012) presented a computer-assisted diagnostic implement based on ultrasound image analysis. The system was used to detect and classify different stages of CKD. They used the K-means machine learning algorithm to detect after the image preprocessing step. The study collected multiple ultrasound images of patients with kidney disease, and selected representative CKD images were applied for pre-analysis and comparison training. They concluded that transition sites calculated as reference indicators could provide physicians with an objective and auxiliary computational aid diagnostic tool for CKD identification and classification.

Valderrama *et al.* (2014) suggested the feasibility study of using a distributed approach for alarm management of patients with kidney disease. They handled alarms

related to monitoring CKD patients within the eNefro project. The results section shows, through the proof of concept studied, the feasibility of Data Distribution Service (DDS) for enabling emergency protocols in terms of prioritization and personalization alarm, as well as some observations on security, privacy and performance real-timecommunication.

Rosmani *et al.* (2015) developed self-care guidelines for CKD patients using Adobe Flash CS5.5. This CKD patient self-care site was developed using Adobe Dreamweaver, and has been helping to manage CKD patient self-care daily by creating a more effective channel of information designed for them.

Hsieh *et al.* (2014) suggested that a real-time system for analyzing chronic kidney disease could be developed using ultrasound images only. The learning set was also used to classify chronic kidney disease by constructing a classifier using Support Vector Machine (SVM) to predict and classify the stage of CKD with ultrasound images.

Singh *et al.* (2014), showed different methods to leverage the hierarchical structure in ICD-9 codes for CKD and heart failure assessment through high dimensionality data. This research proposed and evaluated a new feature of the engineering approach to leverage this hierarchy while improving the performance of predictive methods.

Classification Techniques

Classification and prediction are a data mining technique that first uses training data to develop a model and then the resulting model is applied to test data to obtain prediction results (Mandli, 2014). Several classification algorithms were applied to data sets for the diagnosis of chronic kidney disease and the results were considered very promising (Table 1).

Table 1. Classification Algorithms for CKD Prediction

Authors (Year)	Location	Database	Instance	Attributes	Methods	Accuracy (%)
Kusiak *et al* (2005)	USA	UIHC	188	50	DT RS	75 67
Abhishek *et al* (2012)	India	Hospitals	1199	7	BPA RBF SVM	81 62 60
Chiu *et al* (2012)	Taiwan	HEC	430	6	BPN MNN GFNN	94,75 93,23 86,63

Vijayarani and Dhayanand (2015)	India	KFT	583	6	SVM NB	76,32 70,96
Charleonnan *et al* (2016)	Thailand	UCI	400	24	SVM KNN LR DT	98,3 98,1 96,55 94,8
Boukenze *et al* (2016)	Marroco	UCI	400	24	C4.5 SVM NB	63 60,25 57,5
Anantha and Parthiban (2016)	India	CDC	600	12	DT NB	91 86
Tazin *et al* (2016)	Bangladesh	UCI	400	15	DT SVM KNN NB	99 98 97 96
Manish (2016)	India	UCI	400	25	RF SMO NB RBF MLPC SLG	100 97 95 98 98 98
Chimwayi *et al* (2017)	India	UCI	400	24	Neuro-Fuzzy	97

Kusiak *et al.* (2005) used preprocessing, transformations and data mining to gain insight into the interaction between many of the measured parameters and patient survival. Two different data mining algorithms were employed to extract knowledge in the form of decision rules. These rules were used by a decision-making algorithm, which predicts the survival of new hidden patients. They used Reed-Solomon (RS) and Decision Tree (DT) algorithms in 188 patients at the University of Iowa Hospitals and Clinics (UIHC). They totaled 50 important parameters identified by data mining which were interpreted by specific physicians. The decision tree algorithm (DT) produced 75% and RS with 67% correct predictions for the test data set. They introduced a new concept in their research, which was applied and tested using data collected at four sites with dialysis patients. The approach presented in his paper reduced the cost and effort of selecting patients for clinical studies. Patients can be chosen based on the prediction results and the most significant parameters discovered.

Abhishek *et al.* (2012) used three neural network techniques: The Back Propagation Algorithm (BPA), Radial Basis Function (RBF) and a nonlinear Support Vector Machine (SVM) classifier and compared them according to their efficiency and accuracy. They used the WEKA 3.6.5 implementation tool to find the best technique among the three algorithms for kidney stone diagnosis. The main objective of his thesis work was to propose the best diagnostic tool, such as kidney stones identification, to reduce diagnostic time, efficiency and accuracy. The data set for kidney disease was obtained from medical reports of patients from different hospitals and 1199 patients with 7 attributes each were used: age, sex, lymphoctins, monocytes, eosinophils, neutrophils, creatinine. From the experimental results they concluded that the Back Propagation Algorithm (BPA) significantly improved the conventional classification technique for use in the medical field with 81% accuracy over RBF and SVM with 62% and 60%, respectively.

Chiu *et al.* (2012) presented an intelligent model for detecting kidney disease and assessing the severity of a patient. This intelligent model utilizes three types of artificial neural networks including back-propagation networks (BPN), generalized feeding neural networks (GRNN) and modular neural networks (MNN). The input data set for the development of neural networks was collected from the health examination cases provided by this study's collaborative hospital (HCE), which used 430 patients with 6 instances each: creatinine, glucose, systolic pressure, proteinuria, hematuria and urea. The best performing model was chosen for system development. The BPN algorithm obtained the highest accuracy of 94.75% in relation to MNN (93.23%) and GRNN (86.63%). The system developed in line with the best model was deployed on Google's cloud platform, leveraging the Google Application Engine.

Vijayarani and Dhayanand (2015) aimed to predict CKD using Support Vector Machine (SVM) and Naive Bayes (NB). The objective was to compare the performance of these two algorithms based on their accuracy and execution time. A synthetic kidney function test (KFT) dataset was created for the analysis of kidney disease. The data set with 584 instances and 6 attributes used in the comparative analysis were: Age, Sex, Urea, Creatinine and Glomerular Filtration Rate (GFR). This data set consists of affected kidney disease in formation. From the experimental results they observed that the SVM performance was better (76.32% accuracy) compared to the other algorithm (70.96% accuracy).

Charleonnan *et al.* (2016) used four machine learning methods including the K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR) and decision tree classifiers (DT) to predict kidney disease with the aid of the WEKA tool, with a database collected from the UCI Machine Learning Repository (University of California Irvine), consisting of 400 attributes and 24 instances (age, blood pressure, severity specific, albumin, sugar level, red blood cells, pus cells, agglomerates of pus cells, bacteria, blood glucose, blood urea, serum creatinine, sodium, potassium, hemoglobin, cell volume, white blood cell count, red blood cells, hypertension, diabetes, coronary artery disease, appetite, edema and anemia).From the experimental results, they concluded that the SVM classifier provided higher accuracy and, moreover, the SVM has higher sensitivity after training and testing by the proposed method. The SVM classifier showed the highest accuracy than others with 98.3%, while the KNN, Logistic Regression (RL) and Decision Tree (DT) can produce the average accuracy of 98.1%, 96.55% and 94.8%, respectively.

Boukenze *et al.* (2016) used machine learning algorithms such as Support Vector Machine (SVM), Decision Tree (C4.5), and Naïve Bayes (NB). These predictive models are constructed from the chronic kidney disease (UCI) dataset using Weka. Simulation results showed that the C4.5 classifier proved its predictive performance with better results in terms of accuracy and execution time obtained accuracy of 63% followed by SVM (60.25%) and NB (57.5%).

Anantha and Parthiban (2016), obtained high accuracy using Decision Tree for early detection of CKD. In their work, they aimed to predict early detection of chronic kidney disease for diabetic patients with the help of machine learning methods and finally suggested a decision tree to arrive at concrete results with desirable accuracy, measuring their performance to their specification and sensitivity. The Clinical Foundation Heart Disease's available data set of 600 clinical records was collected from a major Chennai-based diabetes research center with 12 instances each: gender, age, heredity, weight, smoking, blood pressure, fasting glucose, postprandial glucose, glycolyzed hemoglobin test, LDL, HDL, VLDL. They tested the data set for classification using Naïve Bayes (NB) and the Decision Tree (DT) method. By comparing the classification algorithms, they concluded that the accuracy is up to 91% for the Decision Tree classification compared to 86% for Naïve Bayes. In order to increase the accuracy of the prediction result, they also used neural network algorithms and clustering data that helped a lot in the mission and also provided room for future research.

Tazin *et al.* (2016) used classification algorithms Supporting Machine Vector (SVM), Decision Tree (DT), Naïve Bayes (NB) and K-Nearest Neighbor (KNN), in the analysis of Chronic Kidney Disease Data collected. In the UCI repository to predict the presence of kidney disease. In the study, the decision tree (DT) shows promising results (99% accuracy) when implemented using the WEKA data mining tool, followed by SVM, KNN and NB with 98, 97 and 96% accuracy values, respectively. The classification algorithm provides vital improvements in classifications with appropriate numeric attributes.

Manish (2016) in his work predicted the risk of chronic kidney disease by comparing numerous algorithms that were implemented using the WEKA tool. The researcher focused on the application of several classifier algorithms including Random Forest (RF), Minimal Sequential Optimization (SMO), Naive Bayes (NB), Radial Basis Function (RBF) and Multilayer Perceptron Classifier (MLPC) and Simple Logistic (SLG), and obtained high accuracy values of 100, 97, 95, 98, 98 and 98, respectively, comparing them with the numerous methodologies applied. The researcher also used validation to classify each classifier.

Chimwayi *et al.* (2017) applied the neuro-fuzzy algorithm to determine the risk of CKD in patients. Predictions made using neuro-fuzzy gave 97% accuracy from the chronic kidney disease (UCI) dataset. Using selected resources, prediction for chronic kidney disease is designed to identify the risk. Prediction results are grouped to identify the percentage of patients at high risk for kidney disease who are most likely to be diabetic. Using hierarchical grouping, three groups formed show that there is a strong relationship between chronic kidney and diabetes.

Therefore, classification methods are a good solution because they provide a more accurate prediction about an individual health because it is a process that separates data into groups whose members have one or more characteristics in common. In addition, artificial intelligence using machine learning is an excellent working tool for health professionals in decision-making (Lenart, 2016).

Final Considerations

There is an extreme need to develop a new classification technique that can accelerate and simplify the process of diagnosing chronic diseases. According to research,

it has been observed that CKD can be predicted using various classifiers in data mining as well as predicting disease stage using Artificial Intelligence. The different experiences observed have shown that most classifiers provide high accuracy values, above 90%, which can be implemented in an easy-to-run interface to assist physicians and healthcare professionals in decision making and the accuracy of patient outcomes of patients.

Many technology companies, such as IBM, Apple, and Google, are investing heavily in healthcare analytics to make disease management easier. It is important to note that IA will not replace physicians, but it is important for physicians to know how to use IA sufficiently to generate their hypotheses, perform "big data" analysis, optimize IA applications and software in clinical practice to bring the era of precision medicine.

REFERENCES

Abhishek, G.S.M.T and Gupta, D. 2012. Proposing Efficient Neural Network Training Model for Kidney Stone Diagnosis. *International Journal of Computer Science and Information Technologies,* vol. 3, 3900-3904.

Anantha Padmanaban, K.R. and Parthiban, G. 2016. Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease. *Indian Journal of Science and Technology,* 2016, vol. 9 (29).

Boukenze, B., Haqiq, A. and Mousannif, H. 2017. Predicting Chronic Kidney Failure Disease using Data Mining Techniques. *Advances in Ubiquitous Networking, Springer*, pp 701-712.

Boukenze, B., Mousannif, H. and Haqiq, A. 2016. Performance of Data Mining Techniques to Predict in Healthcare Case Study: Chronic Kidney Failure Disease. *International Journal of Database Management Systems* (IJDMS), vol.8, n.3.

Charleonnan, A., Fufaung, T., Niyomwong, F., Chokchueypattanakit, W., Suwannawach, S. and Ninchawee, N. 2016. Predictive analytics for chronic kidney disease using machine learning techniques. Management and Innovation Technology International Conference (MIT icon), IEEE.

Chimwayi, K.B., Haris, N., Caytiles, R.D. and Iyengar, N.C.S. 2017. Risk Level Prediction of Chronic Kidney Disease Using Neuro- Fuzzy and Hierarchical Clustering Algorithm (s). *International Journal of Multimedia and Ubiquitous Engineering,* vol.12, n.8.

Chiu. R.K, Chen, R.Y, Wang, S, Jian, S. 2012. Intelligent systems on the cloud for the early detection of chronic kidney disease. *Machine Learning and Cybernetics,* IEEE; p. 1737 – 1742.

Gabriel I. Barbas, Sherry A. Glied, 2010. New Technology and Healthcare Costs - The Robot Assisted Surgery Case "; *The new England Journal of Medicine,* n. 363, p. 707-704.

Gant, V., Rodway, S. and Wyatt, J. 2001. Artificial neural networks: Practical considerations for clinical application. In R. Dybowski & V. Gant, Clinical applications of artificial neural networks (pp. 329-356). Cambridge, MA: Cambridge University Press.

Groves, P and Kayyali, B. 2013. The bigdate revolution in health. McKinsy and Company Health System Reform Center EUA. Available in: http://digitalstrategy.nl/wp-content/uploads/E2-2013.04-The-big-data-revolution-in-US-health-care-Accelerating-value-and-innovation.pdf.

Ho, C., Pai, T., Peng, Y., Lee, C., Chen, Y. 2012. Ultrasonography Image Analysis for Detection and Classification of Chronic Kidney Disease," IEEE Complex, Intelligent and Software Intensive Systems, p. 624 – 629.

Hsieh, J.W., Hung, C., Lee, Y., Chih, C., Shan Lee, W., Fen Chiang, H. 2014. Stage Classification in Chronic Kidney Disease by Ultrasound Image. International Conference on Image and Vision Computing New Zealand, ACM, p. 271-276.

Huang, T. and Lan, L. 2015. Promises and Challenges of Big Data Computing in Health Sciences. Big Data Research, vol. 2, pp 2-11.

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., Kitai, T. 2017 Artificial intelligence in precision cardiovascular medicine. *J. Am. Coll. Cardiol.* Vol. 69, n. 2657–2664.

Kusiaka, A., Dixonb, B. and Shah, S. 2005. Predicting survival time for kidney dialysis patients: a data mining approach. Computers in Biology and Medicine, p. 311 – 327.

Lenart, M., Mascarenhas, N., Xiong, R., Flower, A. 2016 Identifying Risk of Progression for Patients with Chronic Kidney Disease Using Clustering Models. *IEEE Systems and Information Engineering Design Conference* (SIEDS '16) p. 221-226.

Levey, A.S. and Coresh, J. 2012. Chronic kidney disease. Lancet, 379, pp 165–180.

Mandli, I. and Panchal M. 2014 Selection of Most Relevant Features from High Dimensional Data using IG-GA Hybrid Approach. *International Journal of Computer Science and Mobile Computing,* vol.3 Issue. 2, p. 827-830.

Manish, K., 2016. Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm. *International Journal of Compute Science and Mobile Computing,* vol 5, Issue 2, p. 24-33.

R. Weil, 2014. Big Data in Health: A New Era for Research and Patient Care Alan R. Weil. Health Affair, vol. 33, n. 7, pp 1110.

Rosmani, A., Mazlan, U., Ibrahim, A., Zakaria, D. 2015. I-KS: Composition of Chronic Kidney Disease (CKD) Online Informational Self-Care Tool. Computer, Communication, and Control Technology, IEEE, p. 379 – 383.

Singh, A., Nadkarni, G., Guttag, J. and Bottinger, E. 2014. Leveraging hierarchy in medical codes for predictive modeling. *Bioinformatics, Computational Biology and Health Informatics,* ACM, p. 96-103.

Steve G. Peters, James D. Buntrock, 2014. Big Data and the Electronic Health Registry. *Ambulatory Care Manage,* vol. 37, n. 3, pp. 206–210.

Tazin. N., Sabab, S.A, Chowdhury, M.T. 2016. Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique. *International Conference on Medical Engineering, Health Informatics and Technology* (MediTec).

Valderrama, M.A.E, Barroso, T.A., Roa, L.M, Hernández, D.N., Tosina, J.R., Fosalba N.A., Martín, J.A.M. 2014. A Distributed Approach to Alarm Management in Chronic Kidney Disease," IEEE Transl. Biomedical and Health Informatics, vol. 18, p. 1796 – 1803.

Vijayarani, S., Dhayanand, S. 2015. Data mining classification algorithm for kidney disease prediction. *International journal on cybernetic and information,* Volume 4, Issue 4, p.14-24.

Wang, H., Naghavi, M., Allen, C. 2016. GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and causespecific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet, 388, pp 1459–544.

Capítulo II

DEVELOPMENT OF A COMPUTER SYSTEM TO SCREENING PATIENTS WITH CHRONIC KIDNEY DISEASE

[1]Vanessa D. Martins, [2]Antonino C. Santos, [2]Jonnison L. Ferreira, [1]Viviane S. Ferreira, [3]Ewaldo C. Santana, [4]Érika R. Carneiro, [1]André B. Cavalcante and [1]Allan K. Barros

[1]Department of Electrical Engineering, Laboratory for Biological Information Processing (PIB), Federal University of Maranhao (UFMA) Sao Luis 65085680, MA, Brazil.

[2]Applied Computing Core (NCA), Federal University of Maranhao (UFMA) Sao Luis-MA, Brazil.

[3]LAPS- Lab of Signals Acquisition and Processing, State University of Maranhão.

[4]Kidney Disease Prevention Center, University Hospital of Federal University of Maranhao, Sao Luis 65080805, MA, Brazil.

ABSTRACT

This work aims to construct a computer system to aid in the early diagnosis of Chronic Kidney Disease (CKD) using noninvasive clinical data, exploring machine learning techniques. Data collection was performed at a referral center for treatment of chronic kidney disease. The database consists of 443 participants (instances), of whom 178 have no renal disease (control) and 265 have chronic kidney disease. The clinical data collected were: Gender, Age, Stature, Weight, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP) and Diabetes. To classify chronic kidney disease, four classifier algorithms were tested: Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM) and K-nearest neighbors (KNN). The classifier that obtained the best result was applied a graphical interface. Among the classifiers, the SVM showed more accurate results than the other classifiers with 93.18% of accuracy, the sensibility and specificity parameters were also higher than the other methods, 0.96 and 0.88, respectively. Thus, SVM was the classifier used to obtain the computer system that is available *online* for health professionals and the general population, presenting a low cost and easy execution alternative for screening patients with CKD.

Key Words: Classifier algorithms, DRC, Prediction, Software.

INTRODUCTION

Chronic Kidney Disease (CKD) is defined as abnormalities in the structure or functions of the persistent kidneys for more than three months. This pathology affects more than 10% of the general population, decreasing the quality of life of millions of people, becoming one in the last years a public health problem at a global level (Eckardt *et al*., 2013).

DRC is associated with the presence of comorbidities as cardiovascular diseases and stroke (Baumgarten *et al*., 2011). The main underlying diseases, prevalent in patients with CKD, are hypertension and diabetes, as well as obesity, which has been a worrying factor in recent decades (Baumgarten *et al*., 2011). Other risk factors are also associated with CKD, such as family history of kidney disease, advanced age, chronic use of anti-inflammatories, chronic glomerulonephritis, chronic pyelonephritis, prolonged acute kidney injury, autoimmune diseases, lifestyle (smoking, low water consumption, sedentary lifestyle, among others) (Chimwayi, 2017 and Draibe *et al*., 2014).

DRC is classified in five stages based on the degree of reduction of the glomerular filtration rate, going from the normal/elevated condition to dialysis or transplantation (Draibe, 2014). Because it is asymptomatic in its early stages (Baumgarten *et al.,* 2011), the development of diagnostic and/or screening methods for the early detection of CKD is of great importance for public health. A computational tool that can identify in advance whether or not the patient has kidney disease, for example, may assist health professionals in the early diagnosis of this pathology, thus preventing the progression of the disease and preventing its complications.

Following this line of prevention and early diagnosis, several studies have proposed methods of evaluating CKD, through computational models. Ho *et al*. (2012), for example, presented a computer-aided diagnostic tool based on ultrasound imaging used to detect and classify different stages of CKD. Estudillo-Valderrama *et al*. (2014), have suggested the feasibility study of a distributed approach for the management of alarms related to the monitoring of patients with CKD within the eNefro project. Rosmani *et al*. (2015), developed self-care guidelines for patients with CKD, and implemented a communication channel that assists patients in their daily self-care. Jun-Wei *et al*. (2014), have developed

a system that evaluates in real time the patient's ultrasound images in order to verify the probability of having CKD.

Other studies have used machine learning (ML) techniques, Singh *et al* (2014), used hierarchical methods for assessing CKD and heart failure through high dimensional data. Chiu *et al* (2012), proposed an intelligent model using artificial neural networks that detects and evaluates the severity of renal disease. Anantha Padmanaban (2016), obtained high accuracy in the early detection of CKD using Decision Tree as the classification method. Therefore, it is verified that ML methods are a solution to classification problems such as screening of patients with CKD. For, they offer a more accurate prediction about the health of the individual (Lenart, 2016). In the field of health care, this work aims to construct a classification model to assist in the early diagnosis of CKD using non-invasive clinical data, low cost and easy application, exploring machine learning techniques.

METHODOLOGY

Database: Data collection was performed at a referral center for the treatment of renal disease from July 2017 to July 2018. The present study is approved by the Research Ethics Committee of the Federal University of Maranhão, according to CAAE opinion: 67030517.5.0000.5087. The database consists of 443 adult patients (instances), aged between 20 and 80 years, of whom 178 presented no underlying disease (healthy) and 265 presented CKD.

Input variables: The noninvasive data set presents seven attributes (characteristics): Gender, Age, Stature, Weight, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP) and Diabetes. The attributes used in the data set are presented in Table 1.

Table 1. Set of input attributes used in the experiment.

Variable	Type	Description
Gender	Binominal	M/F
Age	Interger	Age of the patient
Stature	Interger	Stature of the patient
Weight	Interger	Body Mass of the patient
SBP	Numeric	Systolic Blood Pressure

DBP	Numeric	Diastolic Blood Pressure
Diabetes	Nominal	Yes or No

Proposed Method: Four machine learning methods were used to predict the case of Chronic Kidney Disease with the aid of WEKA (Waikato Environment for Knowledge Analysis) software, written in Java, developed at Waikato University (Hall *et al*., 2009). The work methodology is shown in Figure 1. Figure 1 show which WEKA software was used to perform the classification experiments. The experiments consisted of two steps: training and testing of the classifiers using the respective 90% and 10% databases. After the classifier training phase, the 10-fold cross-validation method was used for testing. The best classifier result was an easy-to-use graphical interface to obtain DRC predictor software.

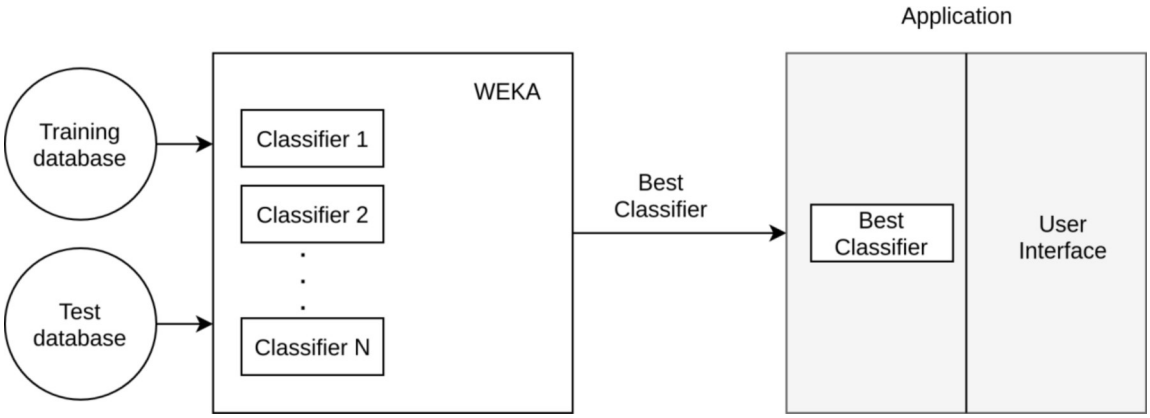


Figure 1. Work methodology

Machine Learning Methods

K-nearest neighbors: The KNN algorithm is the supervised machine learning method used to classify unknown elements, seeking the similarity of the data in a standard space (Jun-Wei, 2014). The KNN calculates the distance between two points to predict the class, being more used the Euclidean distance, represented as:

$$d(x, y) = \sqrt{\sum_{i=1}^{k} (xi - yi)^2}$$

The euclidean distance d (x, y) measures the square root of the absolute distance between two points to find the examples of nearest k in a d-dimensional space. The unknown element class is identified by the closest category of its neighbor in common (Galit, 2010).

Support Vector Machine: The support vector machines (SVM) constitute the popular technique of data mining that aims at classification problem, predicting data class (Brereton, 2010). The SVM seeks to determine the optimal hyperplane, that is, a linear separator between two classes in the training data so that the distance is maximized between classes (Estudillo-Valderrama, 2014). The hyperplane generated by the SVM is determined by a subset of the points of the two classes, called support vectors (Ho, 2012).

Naive Bayes: The Naive Bayes algorithm is a simpler probabilistic classifier of the Bayesian networks, it uses only a formula to combine the previous probability and the conditional probabilities, so that it can calculate the probability of all the possible classes. To do this, the choice of the highest ranking of a given set of mutually exhaustive and exclusive classifications with previous probabilities and n attributes followed by the instance values (Dilli, 2017). The subsequent class probability that occurs for the specified instance can be shown as proportional to previous probabilities along with their respective values. In the assumption that if the attributes are independent, the value of the expression can be calculated using the product by calculating this product for each value from 1 to k, the highest value classification can be chosen (Dilli, 2017).

Decision tree: The decision tree is a machine learning technique where a complex problem is decomposed into simpler subproblems, and presents as the main advantage the "decision making" taking into account the attributes that are considered more relevant, according to the metric chosen, besides of being understandable to people (Gama, 2000). The most important attribute is presented in the tree as the first node, and the least important attributes, according to the used criteria, are shown in the subsequent nodes. By choosing and presenting attributes in order of importance, Decision Trees allow users to know which factors most influence their work.

According to Garcia (2000), Decision Trees consist of:

- nodes that represent the attributes,

- arcs (branches) from the nodes and which receive the possible values for these attributes (each descending branch corresponds to a possible value of this attribute) and
- leaf nodes (tree leaves), which represent the different classes of a training set, that is, each leaf is associated with a class.
- Each path in the tree (from root to leaf) corresponds to a classification rule.

Statistical analysis: In the evaluation and comparison of the algorithms regarding the rate of correctness in the classification, the area values under the ROC curve, Kappa Satisitc, accuracy, sensitivity and specificity with the support of WEKA 3.8 software (Massad, 2004).

Implementation of the computational system: All the tests were implemented in the Python programming language with the help of the machine learning libraries: Scikit-Learn (Pedregosa, 2011) and Auto-Sklearn (Pedregosa, 2011) and the Django web development framework (http://www. djangoproject.com/).

RESULTS AND DISCUSSION

The characteristics of the sample composed of 443 patients are shown in Table 2, which consists of 178 negative for CKD and 265 positive for CKD, with a total of 296 women and 147 males, with age, height and mean weight (48.69 - 1.58 - 74.27) for the negative group and (61.2 - 1.55 - 63.94) for the positive group. The mean systolic (SBP) and diastolic (DBP) pressure of the negative group were 122.71 and 77.61, respectively, and for the positive group 141.72 (SBP) and 82.75 (DBP). The presence of diabetes observed in the CKD positive group showed a total of 96 cases.

Table 2. Sample characteristics of the negative and positive group for Chronic Kidney Disease (CKD) database

Variables	Negative CKD (n=178)	Positive CKD (n=265)
Gender		
Female (total)	n= 123	n= 173

Male (total)	n= 55	n= 92
Age (years)	48,69 ± 11,66	61,2 ± 11,54
Height (meter)	1,58 ± 0,09	1,55 ± 0,08
Weight (Kg)	74,27 ± 13,71	63,94 ± 11,72
SBP (mmHg)	122,71 ± 11,49	141,72 ± 25,11
DBP (mmHg)	77,61 ± 7,76	82,75 ± 14,90
Diabetes (total)	n= 0	n= 96

Data presented as mean ± standard deviation and total value (n).

Abbreviations: meter; kg-kilogram; SBP - Systolic Blood Pressure; DBP- Diastolic Blood Pressure; mmHg-millimeter of mercury.

The performance of the four classifier algorithms are compared using area under ROC curve, Kappa statistics and Precision, which can be observed in Table 3.

Table 3. Classifier performance for the selected dataset

Classifier Algorithms	Área ROC	Kappa statistic (K)	Precicion (%)
SVM	0,93	0,85	93
KNN	0,87	0,71	87
Random Forest	0,90	0,65	83,9
Naive Bayes	0,85	0,56	81,8

Abbreviations: SVM: Support Vector Machines, KNN: K-nearest neighbors, ROC- Receiver operating characteristics curve.

The Receiver Operational Characteristic Curve (ROC) graphically represents the exchange between false positive and false negative. The area under the curve evaluates the performance of the classifiers based on: excellent (0.90-1), good (0.80-0.90), regular (0.70-0.80), poor (0.60- 0.70) and failure (0.50-0.60). Table 3 shows that the SVM and Random Forest classifiers have higher ROC values of 0.93 and 0.90, respectively, considered excellent in the classification scale. The KNN and Naive Bayes also provide satisfactory measurements, so it can be said that for the selected dataset all the classifiers showed the

characteristics of a good classifier. From Table 3, it is observed that all classifiers presented K value greater than zero which represents a chance agreement. Among all classifiers the SVM has a higher value of K = 0.85, which means the perfect agreement between the classifier and the fundamental truth for the given data. Regarding precision, the SVM showed again superior to the others with 93%.

The accuracy of the machine learning techniques trained and tested by the proposed method are compared among the classifiers shown in Figure 2.

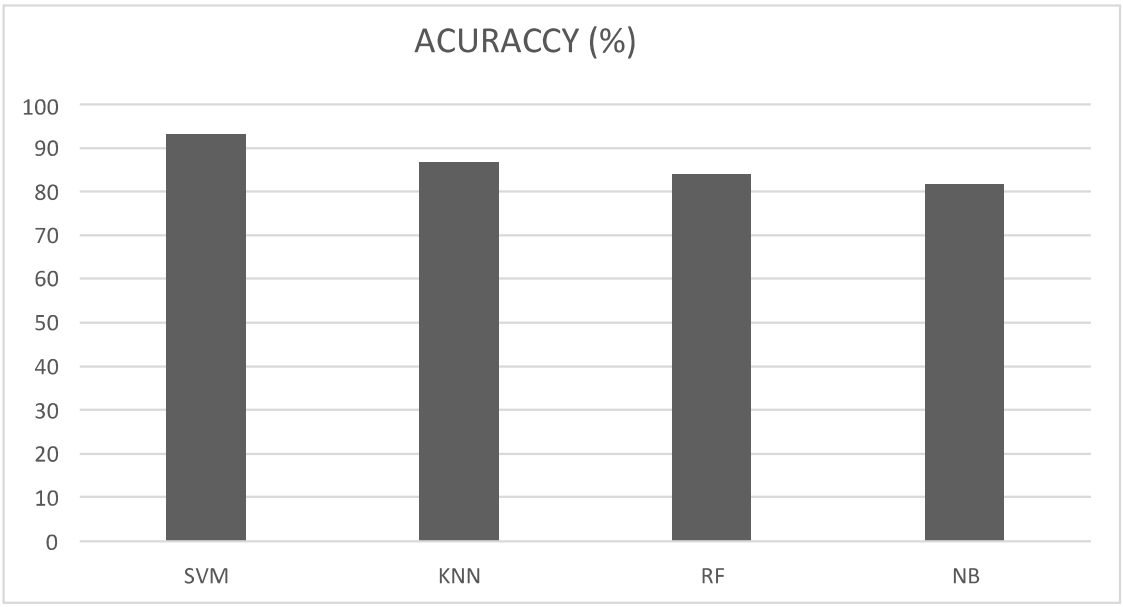


Figure 2. Acuraccy of classifier algorithms

It is observed in Figure 2 the results associated to the four classifiers, where it shows that the SVM classifier presented greater accuracy in relation to the others with 93.18%, while, while KNN, Random Forest (RF) and Naive Bayes (NB) can provide accuracy of 86.36%, 84.09% and 81.81%, respectively.

Calculating the accuracy is relevant, because although evaluating the general effectiveness of the algorithm, if the classifier demonstrates an incorrect prediction, it can bring harm to the patient. Therefore, the sensitivity and specificity value is used in the experiments to evaluate the performance of the proposed methods.

Figure 3 illustrates the sensitivity and specificity parameters of the classifiers used in the experiment. The SVM classifier indicates slightly higher values of sensitivity of 0.96 compared to KNN with 0.92. The sensitivity values of the Random Forest (RF) and Naive Bayes (NB) classifiers were lower with 0.86 and 0.85, respectively. Regarding the specificity value, the SVM classifier also presented slightly higher than the other methods in 0.88, Random Forest (RF), KNN and Naive Bayes (NB) showed specificity of 0.80, 0.77 and 0.75, respectively.

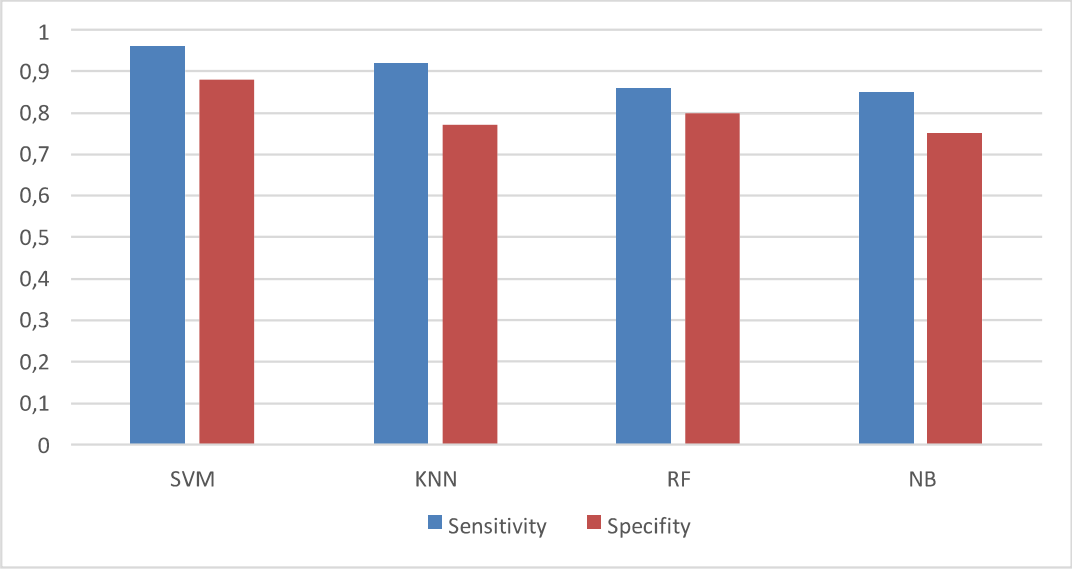


Figure 3. Comparison between sensitivity and specificity of the classifiers

Several investigations using other methods of classification were applied in a set of data for diagnosis of chronic kidney disease and obtained good results of accuracy, it is worth noting that they used invasive clinical data as input to the classifier. Our study, compared to previous studies, can be seen in Table 4.

Table 4. Comparison of results with previous surveys

Authors	Database	Instances	Attributes	Methods	Acuraccy (%)
Kusiak, *et al.* (2005)	UIHC	188	50	RS DT	75 57

Abhishek, et *al.* (2012)	Hospitals	1199	7	BPA RBF SVM	81 62 60
Chiu, *et al.* (2012)	HEC	430	6	BPN GFNN MNN	94,75 86,63 93,23
Vijayarani and Dhayanand (2015)	KFT	583	6	ANN SVM	87 76
Anusorn, *et al.* (2016)	UCI	400	24	SVM KNN LR DT	98,3 98,1 96 94
Anantha and Parthiban (2016)	CDC	600	13	DT NB	91 86
Tazin, *et al.* (2016)	UCI	400	15	DT SVM KNN NB	99 98 97 96
Kerina, *et al.* (2017)	UCI	400	25	Neuro-Fuzzy	97
Our study	HUUFMA	442	7	SVM KNN RF NB	93 86 84 81

Abbreviations: University of Iowa Hospitals and Clinics (UIHC), Clinic Foundation Heart Disease (CHD), Synthetic renal function (KFT), Rear Propagation Algorithm (BPA), Radial Base Function (RBF), Support Vector Machine (SVM), Reed-Solomon (RS), Decision tree (DT), Back-propagation network (BPN), generalized feeding neural networks (GRNN), modular neural networks (MNN), Naive Bayes (NB), University of California Irvine (UCI), Decision Tree (C4.5), K-nearest neighbors (KNN), Logistic Regression (LR), Minimal Sequential Optimization (SMO), Radial Basis Function (RBF), Multilayer Perceptron Classifier (MLPC) Simple Logistic (SLG).

In Table 4 it can be observed that in our study, compared to the results of Anusorn, *et al*. (2016) and Tazin, *et al*. (2016) on the prediction of CKD using the SVM classifier algorithm, were slightly similar with accuracy greater than 90%. Choosing the appropriate input variables is the most important feature of the model that can improve prediction accuracy, and our study, using noninvasive data, was possible to obtain high accuracy in relation to the other studies with invasive variables.

This was the first study performed with data from a population from the Brazilian Northeast, more precisely from the University Hospital of Maranhão, using a machine learning technique using less invasive data related to Chronic Renal Disease (CKD) with a

high sensitivity value, which may be added to the health professionals in the aid to the early diagnosis and in the treatment of the patients. With the classifier implementation, seen in Figure 4 and 5, our model can be used as a central computational component in a medical decision support system, and assist physicians in making appropriate decisions.

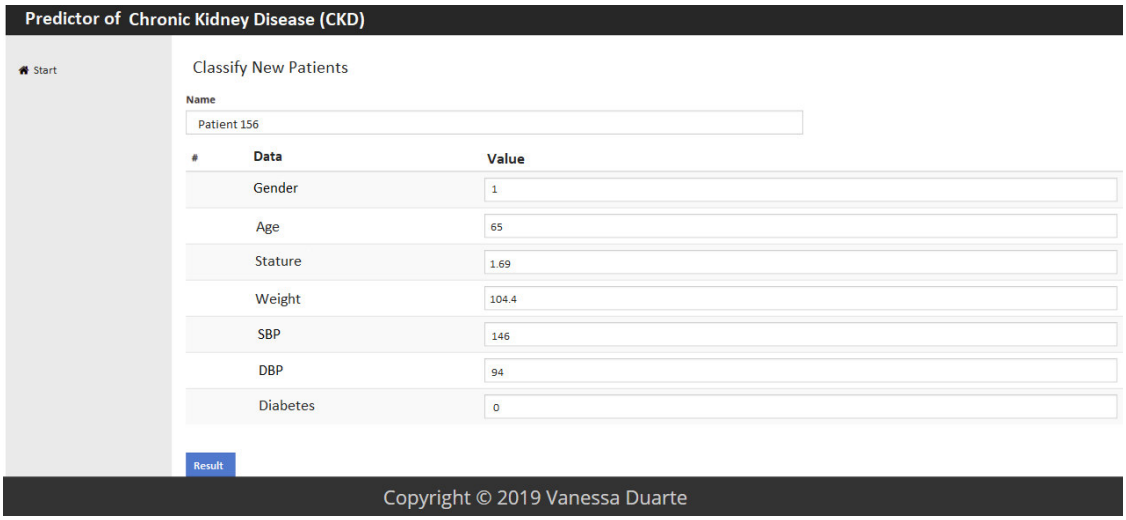


Figure 4. Predicting software for CKD running

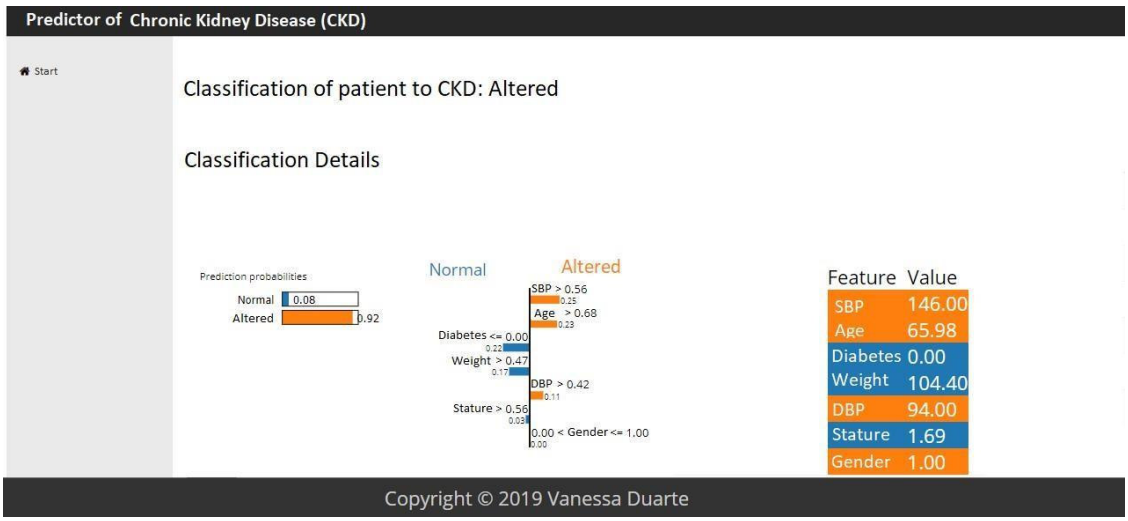


Figure 5. Result of Predicting software for CKD

In Figure 4 shows the running Chronic Kidney Disease Predictor (CKD) software, where it is possible to enter the patient's name with their respective data for further evaluation. Figure 5 presents the software results regarding CKD patient classification, as well as the details of the classification, that is, the variables that most influenced the

classifier in obtaining the result. The Software is available online at: <https://rins.picos.ufpi.br/> and is registered with the National Institute of Industrial Property (INPI) under process No.: BR512019001220-8.

There is an extreme need to develop classification techniques that can accelerate and simplify the process of diagnosing chronic diseases. In this study, the objective was reached, in which the prediction of CKD can be made with noninvasive data related to the disease, in order to simplify the classification process and facilitate health professionals to manage their patients. In the future, other noninvasive parameters such as nutrition, physical activity, water consumption, smoking and alcoholism can be considered for the detection of CKD, as well as the performance of other classifiers such as Artificial Neural Networks, Fuzzy Logic and Logistical Regression can be compared using the WEKA tool for situations and dataset.

Acknowledgements

CAPES and Departamento de Saúde Pública da Universidade Federal do Maranhão, forassigningpatient data toobtaining the study.

Conflicts of Interest: The authors declare no conflict of interest.

CONCLUSION

According to the experimental results, the SVM classifier algorithm showed superior to the others in all evaluated parameters. Thus, SVM was the machine learning technique used to obtain software for predicting chronic kidney disease using noninvasive data.

The software generated from the chosen classifier is available *on line*, presenting a low-cost alternative and easy execution. From it, the medical team can effectively use with ability to assist in the early diagnosis of CKD without invasive exams and accurately treat patients. In addition, public users can take advantage of this model to conduct self-detection so that the necessary precaution can be taken in advance to avoid any risk of causing the disease or preventing it from worsening to later stages.

REFERENCES

Abhishek, Gour Sundar Mitra Thaku e Dolly Gupta (2012). Proposing Efficient Neural Network Training Model for Kidney Stone Diagnosis. International Journal of Computer Science and Information Technologies, Vol. 3, 3900-3904.

Anantha Padmanaban, KR., Parthiban, G. 2016. Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease. *Indian Journal of Science and Technology,* v. 9(29), p. 1-7.

Anusorn, C., Thipwan, F., Tippawan N., Wandee, C., Sathit, S., Nitat, N. (2016) Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques, the Management and Innovation Technology International Conference.

Baumgarten, M., Gehr, T. 2011. Chronic Kidney Disease: Detection and EvaluationAmerican Family Physician, Number 10.

Brereton, RG., Lloyd, GR. 2010 "Support Vector Machines for classification and regression," Analyst, vol. 135, no. 2, p. 230-267.

Chimwayi, K.B., Haris, N., Caytiles, R.D.,Iyengar, S.N. 2017. Risk Level Prediction of Chronic Kidney Disease Using Neuro- Fuzzy and Hierarchical Clustering Algorithm (s). *International Journal of Multimedia and Ubiquitous Engineering,* vol.12, n.8.

Chiu, R.K, Chen, R.Y, Wang, S.A, Jian, S.J. 2012. Intelligent systems on the cloud for the early detection of chronic kidney disease, Machine Learning and Cybernetics, IEEE, p. 1737 – 1742.

Draibe, S.A. 2014. Overview of Chronic Renal Disease in Brazil and the World. UNASUS/UFMA - Sao Luis, 2014.

Dilli, SA., Thirumalaiselvi, R. (2017) Review of Chronic Kidney Disease based on Data Mining Techniques. International Journal of Applied Engineering Research. Vol. 12, n. 23, p. 13498-13505.

Eckardt, K.U., Coresh, J., Devuyst, O. 2013. Evolving importance of kidney disease: from subspecialty to global health burfen. Lancet, p.158–169.

Estudillo-Valderrama, M.A, Talaminos-Barroso, A., Roa, L.M., Naranjo-Hernández, D., Javier, R.T., Nuria, A.F, José, M.M. 2014. A Distributed Approach to Alarm Management in Chronic Kidney Disease. *IEEE Jounal of Biomedical and Health Informatics*, vol. 18, p. 1796 – 1803.

Galit, S., Nitin, RP., Peter, CB. 2010. Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner: Wiley Publishing.

Gama, J. (200) Decision Trees. Available in: http://www.liacc.up.pt/jgama/Masters/ECD1/Trees.html.[accessed august 14, 2018].

Garcia, SC. (2000) The Use of Decision Trees in Health Knowledge Discovery. Academic Week. Federal University of Rio Grande do Sul.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. 2009. The weka data mining software: an update. ACM SIGKDD Explorations Newsletter, ACM, v. 11, n. 1, p. 10–18.

Ho, C., Pai, T., Peng, Y., Lee, C., Chen, Y. 2012. Ultrasonography Image Analysis for Detection and Classification of Chronic Kidney Disease, IEEE Complex, Intelligent and Software Intensive Systems, p. 624 – 629.

Jun-Wei, H., Hung, C., Lee, Y., Chih, C., Shan, L.W. Fen Chiang H. 2014. Stage Classification in Chronic Kidney Disease by Ultrasound Image, International Conference on Image and Vision Computing New Zealand, ACM, p. 271-276.

Kusiaka, A., Bradley, D., Shital, S. (2005) Predicting survival time for kidney dialysis patients: a data mining approach. Computers in Biology and Medicine, p. 311 – 327.

Lenart, M., Mascarenhas, N., Xiong, R., Flower, A. 2016. Identifying Risk of Progression for Patients with Chronic Kidney Disease Using Clustering Models. IEEE Systems and Information Engineering Design Conference (SIEDS '16), p. 221-226.

Massad, E., Menezes, RX., Silveira, PSP., Ortega, NRS. (2004). Métodos quantitativos em medicina. São Paulo: Manole.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, p. 2825–2830.

Rosmani, A., Mazlan, U., Ibrahim, A., Zakaria, D. 2015. I-KS: Composition of Chronic Kidney Disease (CKD) Online Informational Self-Care Tool, Computer, Communication, and Control Technology, IEEE p. 379 – 383.

Singh, A., Nadkarni, G., Guttag, J., Bottinger E. 2014. Leveraging hierarchy in medical codes for predictive modeling, Bioinformatics, Computational Biology, and Health Informatics, ACM, p. 96-103.

Tazin, N., Sabab, S.A, Chowdhury, M.T. (2016) Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique. International Conference on Medical Engineering, Health Informatics and Technology (MediTec).

Vijayarani, S., Dhayanand. (2015) Data mining classification algorithm for kidney prediction", International journal on cybernetic and information, Vol. 4, Issue 4, p.14-24.

Capítulo III

Support System for Chronic Kidney Disease Prediction using Machine Learning

Vanessa Edilene Duarte Martins[1]*, Antonino Calisto dos Santos Neto[3], Jonnison Lima Ferreira[3], Marta de Oliveira Barreiros[2], Viviane Sousa Ferreira[1], Ilka Kassandra Pereira Belfort[1], Erika Ribeiro Carneiro[4], and Allan Kardec Barros[1],[2]

[1] Department of Biotechnology, Laboratory for Biological Information Processing (PIB), Federal University of Maranhao (UFMA) Sao Luis 65085680, MA, Brazil.

[2] Department of Electrical Engineering, Laboratory for Biological Information Processing (PIB), Federal University of Maranhao (UFMA) Sao Luis 65085680, MA, Brazil.

[3] Applied Computing Core (NCA), Federal University of Maranhao (UFMA) Sao Luis-MA, Brazil.

[4] Kidney Disease Prevention Center, University Hospital of Federal University of Maranhao, Sao Luis 65080805,MA, Brazil.

* Corresponding author. Email: vanessa.duartema@gmail.com

Abstract

Background. Chronic kidney disease (CKD) has no signs and/or symptoms in its early stages, and the study and development of alternative methods of diagnosis and / or screening that are highly sensitive is extremely important. Thus, the objective was to build and validate a software for predicting chronic kidney disease based on a classifier algorithm for screening patients. **Method**. First, a classification experiment was carried out, using 4 classifying algorithms: Random Forest (RF), Naive Bayes (NV), Support vector machine (SVM) and K- closest neighbors (KNN). The classifier that obtained the best response was the SVM in all evaluated parameters. From the SVM, a graphical interface was implemented to obtain the software. Afterwards, it was validated using two databases, the study (HUUFMA) and data from the University of California (UCI). Two sets of inputs from each bank were evaluated, with all data and non-invasive data. **Results**. It is observed that with all the data the parameters of accuracy, sensitivity, specificity and precision are superior (0.95 - 1.00 - 0.91 - 0.90) for both the HUUFMA and UCI databases (0.98 -1.00 - 1.00- 0.99). Despite the lower accuracy (0.83) of the HUUFMA's non-invasive data compared to the ICU data (0.94), the sensitivity was higher (0.94). **Conclusions**. The SVM was a classifier used to obtain the CKD predictor software, demonstrated good

performance in the validation which can be used in clinical practice as a way of screening patients with the disease and for the population in general, presenting a low cost and easy alternative execution.

Keywords: CKD, artificial intelligence, prediction, implementation, software.

1. Introduction

Chronic kidney disease (CKD) consists of kidney damage with progressive loss and irreversible effect of kidney function. It can be classified into five stages, according to the degree of reduction in glomerular filtration, ranging from normal/elevated condition to dialysis or transplantation. Complications of chronic kidney disease can affect the entire organism and present at any stage of the disease's evolution, often leading to death, without progression to stage 5 renal failure (Draibe, 2014).

CKD shows no signs and symptoms in patients who are in the stages initial demands, requiring doctors to maintain an adequate level of suspicion, especially in those patients with a higher risk factor for the development of CKD, including hypertension, diabetes, advanced age and family history of CKD.

The glomerular filtration rate (GFR) is the standard measure of renal function, used to diagnose and classify CKD. Defined as the ability of the kidneys to eliminate a blood substance and expressed as the blood volume is completely cleared in a unit of time. GFR is determined, in routine clinical conditions, by the measurement of serum creatinine and/or by its clearance by the kidney. Clearance creatinine can be performed on urine collected within 24 hours, however the collection inadequate urination, either due to lack of understanding of the procedure or type of activity of the patient, is a limiter of the method (Bastos & Kirsztajn, 2011).

Since it is an asymptomatic disease in its early stages, it is extremely important the study and development of alternative diagnostic and/or screening methods that have high sensitivity. Thus, classification methods present a good solution for offering a more accurate forecast about the health of the individual, as it deals with a process which separates data into groups, whose members have one or more characteristics in common (Lenart et al., 2016).

In the machine learning (ML) area, several classifying algorithms are used in studies to predict various diseases, such as heart disease (Xing et al., 2007; Lee, Noh & Ryu, 2008; Srinivas, Raghavendra Rao & Govardhan, 2010; Pal, Chakraborty & Mandana, 2011), cancer, epilepsy, diabetes, Parkinson disease (Jenn-Lung Su, Guo-Zhen Wu & I-Pin Chao; Bonato et al., 2004; Ghannad-Rezaie & Soltanian-Zadeh, 2008; Sison & Gerlai, 2011; Rajan & Chilambu Chelvan, 2013; Ilayaraja & Meyyappan, 2013) including recent studies for CKD detection (Xun et al., 2010; Chiu et al., 2012; Lakshmi, Nagesh & Veerakrishna, 2014; Kunwar et al., 2016). Therefore, the ML techniques using classifying algorithms become an attractive computational tool for the solving complex problems such as the early detection of individuals with CKD. Thus, the objective was to build and validate a software predictor of chronic kidney disease based on a classifier algorithm for patient screening.

2. Methodology

2.1 Database

At the Kidney Disease Prevention Center of the University Hospital da Federal University of Maranhão (HUUFMA) data collection was carried out in the period from July 2017 to July 2018. All study procedures were approved by the University ethics committee on the CAAE number: 67030517.5.0000.5087.

A total of 443 patients were collected and their variables such as: Gender, Age, Height, Weight, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (PAD), Diabetes, Glucose, Creatinine, Cholesterol, Lipoprotein High Intensity (HDL), Low Intensity Lipoprotein, Triglycerides (LDL) and estimate of the Glomerular Filtration Rate (e-TFG). The input data used in the classification experiment include anthropometric, hemodynamic and biochemical indices, totaling 13 attributes, are described in Table 1.

Table 1. Set of input attributes used in the development of the computer system for screening patients with CKD.

Variable	Type	Description
Gender	Binominal	M/F

Age	Interger	Age of the patient
Stature	Interger	Stature of the patient
Weigtht	Interger	Weight of the patient
SBP	Numeric	Systolic Blood Pressure
DBP	Numeric	Diastolic Blood Pressure
Diabetes	Nominal	Yes or No
Glucose	Numeric	Blood glucose
Creatinine	Numeric	Creatinine clearance test
Cholesterol	Numeric	Total cholesterol
HDL	Numeric	High Density Lipoprotein
LDL	Integer	Low Density Lipoprotein
Triglycerides	Numeric	Type of fat in the blood

2.2 Proposed method

Four machine learning methods were used to predict the case of Chronic Kidney Disease with the help of the WEKA tool (Waikato Environment for Knowledge Analysis), written in Java, developed at the University of Waikato (Hall et al., 2009). The work methodology is shown in Figure 1.

Figure 1. Classification experiment methodology using the Weka tool to choose the best classifier to be implemented.

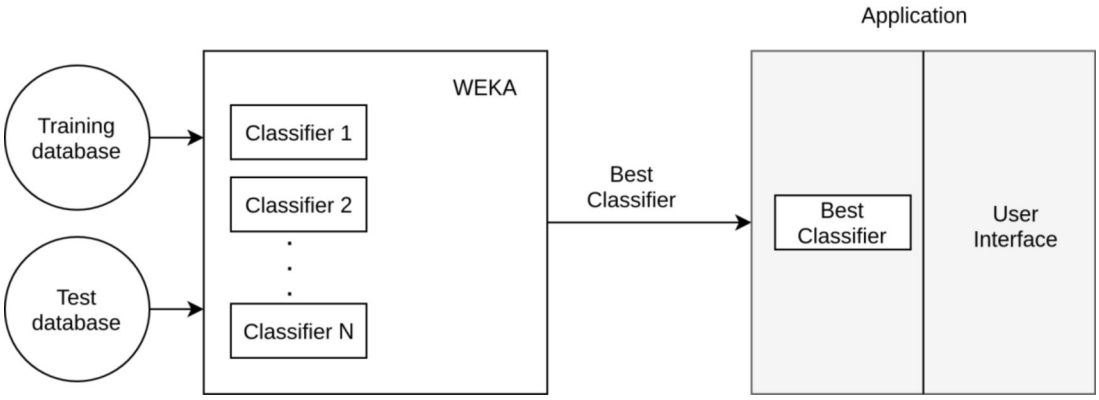


Figure 1 shows which WEKA software was used to carry out the classification experiments. The experiments consisted of two stages: training and testing the classifiers using the databases, with a ratio of 90% and 10% respectively. After the classifier training phase, the 10-fold cross-validation cross-validation method was used for testing. The result of the best classifier was applied to an easy-to-use graphical interface to obtain the software.

2.3 Statistical analysis

In the evaluation and comparison of the algorithms regarding the correctness rate in the classification, the area values under the ROC curve, Kappa Satistic, Accuracy, Sensitivity and Specificity were used as a metric for the selection of the algorithm with the aid of the WEKA 3.8 software (Eduardo Massad, Renée X. de Menezes, Paulo S. P. Silveira, 2004).

2.4 Machine Learning Algorithms

The machine learning algorithms used in this work were: K-nearest neighbors (KNN), Support vector machine (SVM), Naive Bayes and Decision tree (TREE).

The KNN algorithm is the supervised machine learning method used to classify unknown elements, looking for similarity of data in a standard space. The distance between two points is calculated to predict the class, the Euclidean distance (x, y) being more used, which measures the square root of the absolute distance between two points to find the examples of k nearest in a d-dimensional space. The class of the unknown element is identified by the category closest to its common neighbor (Kumar, 2012).

The SVM seeks to determine the optimal hyperplane, that is, a linear separator between two classes in the training data so that the distance is maximized between the classes (Estudillo-Valderrama et al., 2014). The hyperplane is determined by a subset of the points of the two classes, called support vectors (Ho et al., 2012).

The Naive Bayes algorithm is a simpler probabilistic classifier of the Bayesian networks, it uses only one formula to combine the previous probability and the conditional probabilities, so that it can calculate the probability of all possible classes. To do this, choosing the classification with the highest value for a given set of k mutually exhaustive and exclusive classifications with previous probabilities and n attributes followed by the

instance values. The subsequent class probability that occurs for the specified instance can be shown as proportional to previous probabilities together with the respective values. In the assumption that, if the attributes are independent, the value of the expression can be calculated using the product by calculating this product for each value from 1 to k, the classification with the highest value can be chosen (Dilli Arasu & Thirumalaiselvi, 2017).

Random Forest is a supervised learning algorithm based on several decision trees, it creates a forest in a random way, each one with its particularities and combined the result of the classification of all of them. Thus, with this combination of models, it makes it an algorithm much more powerful than the Decision Tree (Breiman, 2001).

2.5 Software validation

The software was developed using the Python programming language with the aid of machine learning libraries: Scikit-Learn and Auto-Sklearn. The classifier used in the implementation of the software was the SVM, because among the classifiers with recent testing, the best performance is presented.

For validation of software, the study database and the University of California, Irvine data repository (UCI), called Chronic_Kidney_Disease_DataSet (Rubini, 2015), were used. The data set was provided by the Apollo Hospital, India, and contains 400 instances and 24 attributes with two classes, 250 (62.5%) patients with CKD and 150 (37.5%) records of people without CKD. Table 2 describes the attributes present.

Table 2: Attributes of the UCI database.

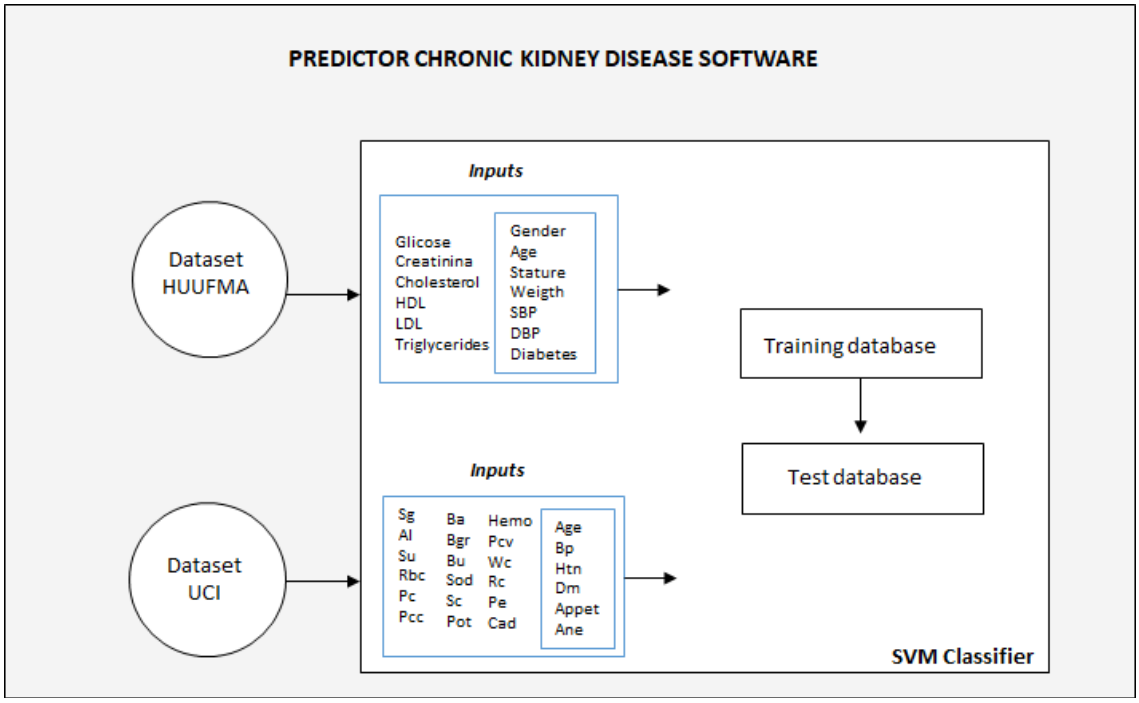
Attribute	Representation	Information attribute	Description
Age	Age	Numerical	Years
Blood pressure	Bp	Numerical	Mm/Hg
Specific gravity	Sg	Nominal	1.005,1.010,1.015,1.020,1.025
Albumin	Al	Nominal	0.1.2.3.4.5
Sugar	Su	Nominal	0.1.2.3.4.5
Red blood cells	Rbc	Nominal	Normal, abnormal

Pus cell	Pc	Nominal	Normal, abnormal
Pus cell clumps	Pcc	Nominal	Present, notpresent
Bacteria	Ba	Nominal	Present, notpresent
Blood glucose random	Bgr	Numerical	Mgs/dl
Blood urea	Bu	Numerical	Mgs/dl
Serum creatinine	Sc	Numerical	Mgs/dl
Sodium	Sod	Numerical	mEq/L
Potassium	Pot	Numerical	mEq/L
Haemoglobin	Hemo	Numerical	Gms
Packed cell volume	Pcv	Numerical	
White blood cell count	Wc	Numerical	Cells/cumm
Red blood cell count	Rc	Numerical	Millions/cmm
Hypertension	Htn	Nominal	Yes, no
Diabetes mellitus	Dm	Nominal	Yes, no
Coronary artery disease	Cad	Nominal	Yes, no
Appetite	Appet	Nominal	Good, poor
Pedal edema	Pe	Nominal	Yes, no
Anemia	Ane	Nominal	Yes, no
Class	Classe	Nominal	Ckd notckd

The databases were divided into 90% and 10% for training and testing, respectively. All data were initially normalized. For testing the software, the evaluation metrics used were the values of accuracy, sensitivity and specificity. Accuracy assesses the overall performance of the classifier, where VP is the true value of the positive rate and VN true value of the negative rate, FP is the false value of the positive rate and FN false value of the negative rate. Sensitivity is the model's ability to correctly classify the class defined as positive. Specificity is the model's ability to correctly classify the class defined as negative.

The summary of the software validation method is illustrated in Figure 2.

Figure 2: Validation method of CKD predictive software from the SVM classifier using two databases with all variables and non-invasive input variables.



Abbreviations: HDL- High Density Lipoprotein; LDL- Low Density Lipoprotein; SBP- Systolic Blood Pressure; DBP- Diastolic Blood Pressure; Bp- Blood pressure; Htn- Hypertension; DM- Diabetes mellitus; Appet- Appetite; Ane- Anemia; Sg- Specific gravity; Al- Albumin; Su- Sugar; Rbc- Red blood cells; Pc- Pus cell; Pcc- Pus cell clumps; Ba- Bacteria; Bgr- Blood glucose random; Bu- Blood urea; Sod- Sodium; Sc- Serum creatinina; Pot- Potassium; Hemo- Haemoglobin; Pcv- Packed cell volume; Wc- White blood cell count; Rc- Red blood cell count; Pe- Pedal edema; Cad- Coronary artery disease.

Figure 2 represents the software validation method, in which two databases were used, HUUFMA's with two sets of input, with 13 total attributes (gender, age, height, weight, SBP, DBP, diabetes, glucose, creatinine, cholesterol, HDL, LDL and triglycerides) and 7 non-invasive attributes (gender, age, height, weight, SBP, DBP and diabetes) and the UCI database, with 24 total attributes (Age, Bp, Htn, Dm, Appet , Ane, Sg, Al, Su, Rbc, Pc, Pcc, Ba, Bgr, Bu, Sod, Sc, Pot, Hemo, Pcv, Wc, Rc, Pe and Cad) and 6 non-invasive attributes (Age, Bp, Htn, Dm, Appet and Ane). SVM was the classifier used to obtain the software, which went through two stages: training and testing.

3. Results

The characteristics of the sample consisting of 443 patients are shown in Table 3, which is 178 negative for CKD and 265 positive for CKD, with a total of 296 women and 147 men, with age, height and average weight (48.69 - 1.58 - 74.27) for the negative group and (61.2 - 1.55 - 63.94) for the positive group. The mean systolic (SBP) and diastolic (DBP) blood pressure in the negative group was 122.71 mmHg and 77.61 mmHg, respectively, and for the positive group 141.72 mmHg (SBP) and 82.75 mmHg (DBP). The presence of diabetes observed in the positive CKD group showed a total of 96 cases.

Regarding biochemical tests, the negative group had mean values of glucose, creatinine, total cholesterol, HDL, LDL and triglycerides (93 - 0.7 - 194.40 - 48.75 - 116.29) and the positive group (120, 29 - 2.06 -174.01 - 45.32 - 110.74 - 139.29) respectively. Regarding the Glomerular Filtration Rate (e-GFR) estimate, the mean values were 126.94 mL / min / 1.73m² for the negative group and 53.89 mL / min / 1.73m² for the positive group.

Table 3: Sample characteristics of the negative and positive group database for Chronic Kidney Disease (CKD).

Variables	**Negative CKD (n=178)**	**Positive CKD (n=265)**
Genre		
Female (total)	123	173
Male (total)	55	92
Age (years)	48,69 ± 11,66	61,2 ± 11,54
Height (m)	1,58 ± 0,09	1,55 ± 0,08
Weight (Kg)	74,27 ± 13,71	63,94 ± 11,72
SBP (mmHg)	122,71 ± 11,49	141,72 ± 25,11
DPB (mmHg)	77,61 ± 7,76	82,75 ± 14,90
Diabetes (total)	0	96
Glucose	93 ± 17,34	120,29 ± 15,85
Creatinine	0,7 ± 7,19	2,06 ± 10,48
Total cholesterol	194,40 ± 23,56	174,01 ± 20,72
HDL	48,75 ± 24,81	45,32 ± 12,85
LDL	116,38 ± 26,69	110,74 ± 20,53

Triglycerides	146,29 ± 30,56	139,29 ± 40,53
eTFG (mL/min/1,73m²)	126,94 ± 17,45	53,89 ± 19,01

Data presented as mean ± standard deviation and total value (n).

Abbreviations: meter; kg- kilogram; SBP- Systolic Blood Pressure; DBP- Diastolic Blood Pressure; mmHg- millimeter of mercury; HDL: High density lipoproteins; LDL: low density lipoproteins; e-TFG: estimate of glomerular filtration rate.

Table 4 shows the performance of the classifiers, in which the SVM provided greater accuracy with 0.95 compared to Naive Bayes (0.88), Random Forest (0.80) and KNN (0.75). Regarding sensitivity, SVM is slightly higher than Naive Bayes (NB) of 1.00 and 0.96, respectively. Random Forest and KNN presented sensitivity with values of 0.84 and 0.66, respectively. The specificity of SVM was higher (0.88) in relation to Naive Bayes (NB), Random Forest and KNN with values of 0.78; 0.74; and 0.54. The SVM also showed higher values of ROC area (0.98), Kappa statistic (0.90) and precision (96%) in relation to the other classifiers.

Table 4: Classifier performance in relation to evaluation metrics for the HUUFMA data set using the Weka tool

Classifier	ACU	SEN	SPE	ROC Area	Kappa statistic (K)	Precision (%)
SVM	0,95	1,00	0,88	0,98	0,90	96,00
Naive Bayes	0,88	0,96	0,78	0,94	0,76	89,80
Random Forest	0,80	0,84	0,74	0,79	0,58	82,04
KNN	0,75	0,66	0,54	0,74	0,46	75,40

Abbreviations: ACU: Accuracy, SEN: Sensitivity, SPE: specificity, ROC: Receiver Operational Characteristic Curve, SVM: Support vector machines, KNN: K-nearest neighbors.

Table 5 describes the performance of the software validation, with the two databases and the divisions of the attributes, with all the data and with the non-invasive data. It is observed that with all data the parameters of accuracy, sensitivity, specificity and precision are superior (0.95 - 1.00 - 0.91 - 0.90) for both the HUUFMA database and the database. UCI data (0.98 - 1.00 - 1.00 - 0.99). Despite the lower accuracy (0.83) of the non-invasive data from the HUUFMA database in relation to the UCI database (0.94), the sensitivity was higher (0.94). However, the specificity and precision of non-invasive data in the UCI database showed maximum values (1.00).

Table 5: Performance of validation of the DRC software predictor using he classifier SVM in relation to evaluation metrics

Software	ACU	SEN	SPE	Precision	F-Score
Dataset HUUFMA					
13 attributes	0,95	1,00	0,91	0,90	0,95
7 no-invasive attributes	0,83	0,94	0,75	0,75	0,83
Dataset UCI					
24 atributes	0,98	1,00	1,00	0,99	0,98
6 no-invasive attributes	0,94	0,91	1,00	1,00	0,95

Abbreviations: ACU: accuracy, SEN: sensitivity, ESP: specificity, HUUFMA: University Hospital of the Federal University of Maranhão, UCI: University of California de Irvine.

Figures 3 and 4 represent the development of the Predictor Software for Chronic Kidney Disease (CKD) using 13 attributes from the HUUFMA data with and Figures 5 and 6 using 6 non-invasive attributes from the UCI data. In Figures 3 and 5 the software is running, in which the name of the patient can be entered, as well as its respective data for later evaluation. Figures 4 and 6 show the result of the patient's classification as altered (data from HUUFMA) and normal (data from UCI) for CKD, and the details, that is, which attributes influenced the results according to the learning of the SVM classifier during the

training. The Software is available online at the site: <https://rins.picos.ufpi.br/> and is registered at the National Institute of Industrial Property (INPI) under process No.: BR512019001220-8.

Figure 3: CKD Predictor software running with HUUFMA attributes.

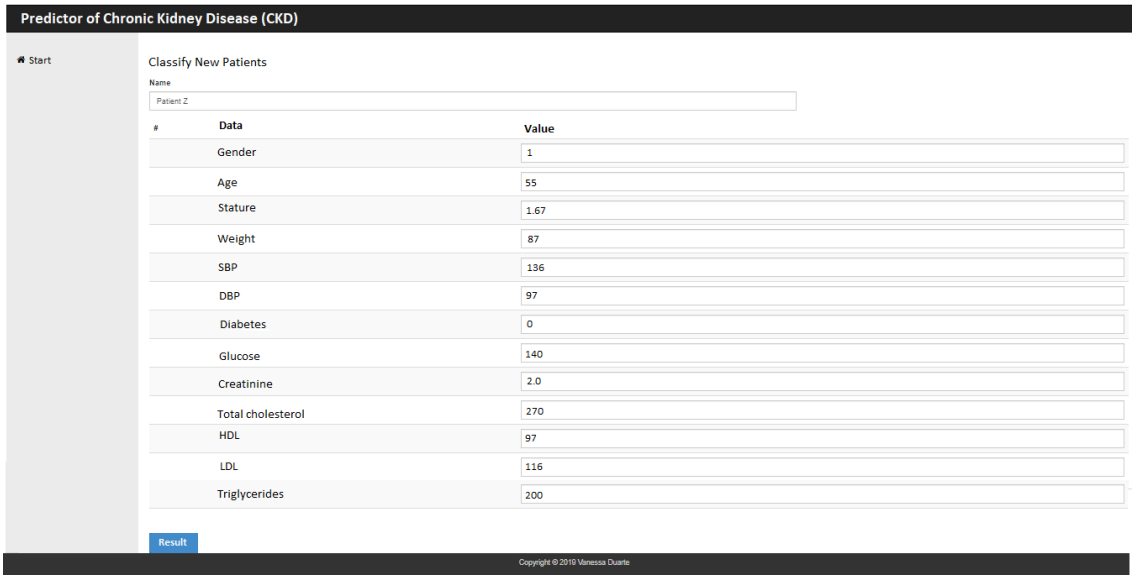


Figure 4: Result of patient classification for CKD with HUUFMA attributes.

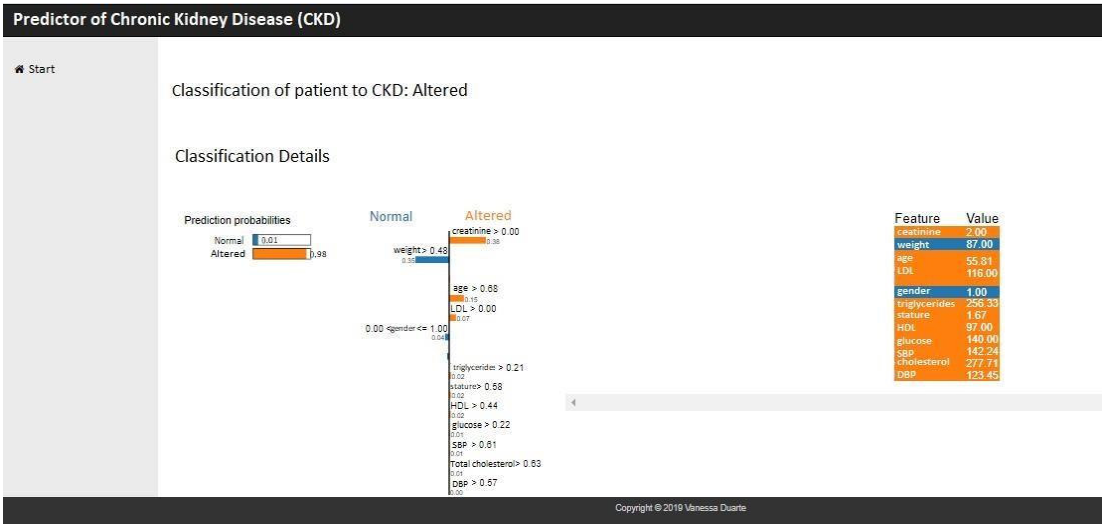


Figure 5: CKD Predictor software running with UCI attributes.

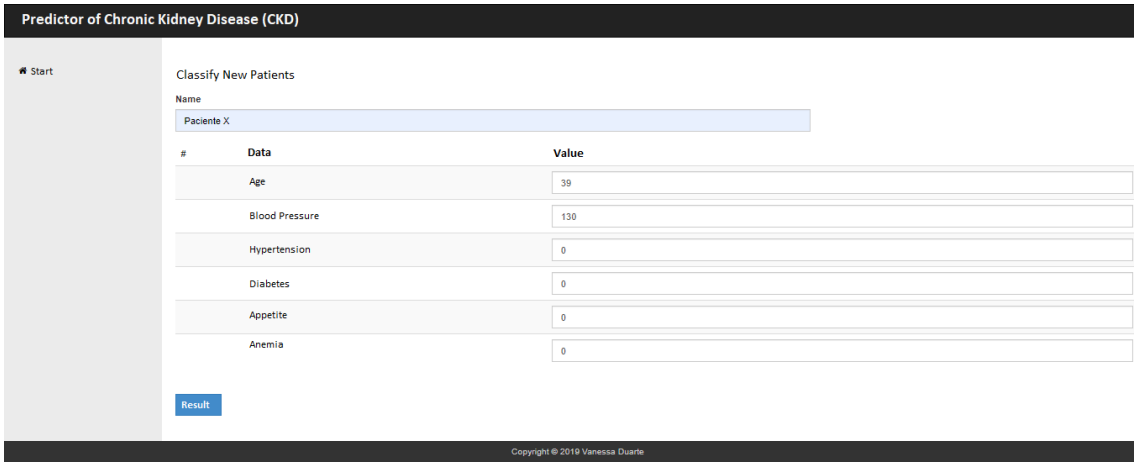
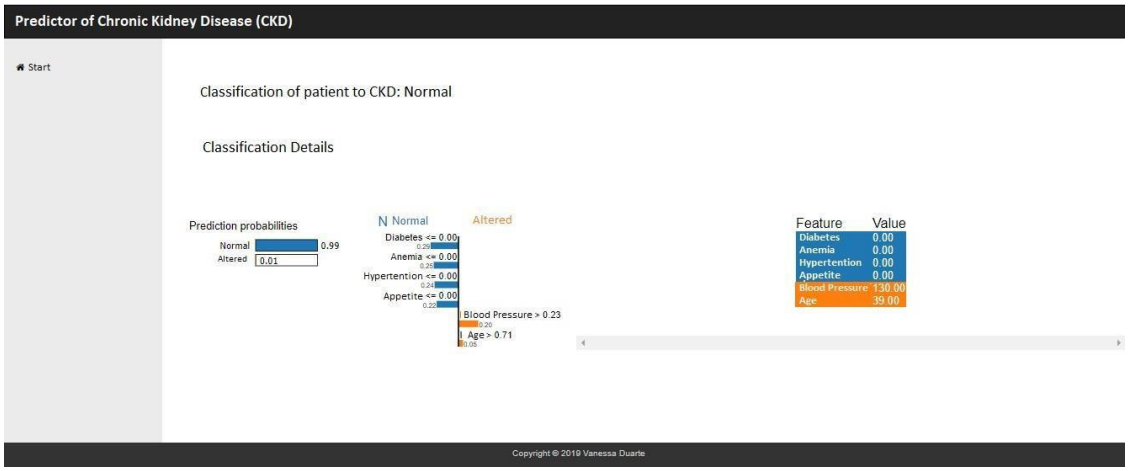


Figure 6: Result of patient classification for CKD with UCI attributes.



4. Discussion

There is a very significant fraction of patients with occult CKD, and one of the main concerns of doctors is the early diagnosis and regular monitoring of the slowdown in the progression of the disease and the prevention of its inevitable complications. In this research, we developed a new forecasting software for CKD based on some clinical parameters and non-invasive data. We evaluated four classifiers: Random Forest, Naive Bayes, KNN and SVM. Of these, the latter showed greater accuracy and the graphical interface was implemented from it.

Several predictors have been evaluated in the population with CKD, such as diseases (Sengur, 2008; Chang-Shing Lee & Mei-Hui Wang, 2011; Zhao et al., 2017). Our

results compared to the research by Yadollahpour et al. (Yadollahpour et al., 2018)on the implementation of a system to predict the progression of kidney disease, as well as other studies using methods with machine learning, are of great value, presenting very promising results.

The UCI database has been used in several studies to predict CKD with good accuracy results (Boukenze, Mousannif & Haqiq, 2016; Kumar, 2016; Charleonnan et al., 2017; Chimwayi et al., 2017; Tazin, Sabab & Chowdhury, 2017). In relation to the SVM algorithm, our study showed results similar to the research by Anusorn et al., (Charleonnan et al., 2017) and Tazin (Tazin, Sabab & Chowdhury, 2017) with 98% accuracy using invasive data as input. It is observed that there is a lack in these researches in implementing the evaluated algorithms in a graphical interface for users in general.

This work was carried out with a small population of northeastern Brazil, in the city of São Luis do Maranhão, in order to generate a support system to assist in the early dignity of CKD, mainly using non-invasive data related to the disease. The software showed good results of accuracy, after validation, in the two databases used. It is an easy-to-use device, it can be used on computers, tablets and cell phones with internet access as it is available online on a free platform. It is possible to insert other databases for evaluation, with other input variables, including more parameters influencing renal function. Thus, this system can help to reduce the costs of management CKD, in addition to reducing the mortality rates of the disease.

5. Conclusion

The SVM was a classifier used to obtain the CKD predictor software, demonstrated good performance in the validation which can be used in clinical practice as a way of screening patients with the disease and for the population in general, presenting a low cost and easy alternative execution.

6. References

Bastos MG, Kirsztajn GM. 2011. Chronic kidney disease: importance of early diagnosis, immediate referral and structured interdisciplinary approach to improve outcomes in patients not yet on dialysis. *Jornal brasileiro de nefrologia: orgão oficial de Sociedades Brasileira e Latino-Americana de Nefrologia*. DOI: 10.1590/S0101-

28002011000100013.

Bonato P, Sherrill DM, Standaert DG, Salles SS, Akay M. 2004. Data mining techniques to detect motor fluctuations in Parkinson's disease. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*. DOI: 10.1109/iembs.2004.1404319.

Boukenze B, Mousannif H, Haqiq A. 2016. Performance of Data Mining Techniques to Predict in Healthcare Case Study: Chronic Kidney Failure Disease. *International Journal of Database Management Systems*. DOI: 10.5121/ijdms.2016.8301.

Breiman L. 2001. Random forests. *Machine Learning*. DOI: 10.1023/A:1010933404324.

Chang-Shing Lee, Mei-Hui Wang. 2011. A Fuzzy Expert System for Diabetes Decision Support Application. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41:139–153. DOI: 10.1109/TSMCB.2010.2048899.

Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattanakit W, Suwannawach S, Ninchawee N. 2017. Predictive analytics for chronic kidney disease using machine learning techniques. In: *2016 Management and Innovation Technology International Conference, MITiCON 2016*. DOI: 10.1109/MITICON.2016.8025242.

Chimwayi KB, Haris N, Caytiles RD, Iyengar NCSN. 2017. Risk Level Prediction of Chronic Kidney Disease Using Neuro- Fuzzy and Hierarchical Clustering Algorithm (s). *International Journal of Multimedia and Ubiquitous Engineering*. DOI: 10.14257/ijmue.2017.12.8.03.

Chiu RK, Chen RY, Wang SA, Jian SJ. 2012. Intelligent systems on the cloud for the early detection of chronic kidney disease. In: *Proceedings - International Conference on Machine Learning and Cybernetics*. DOI: 10.1109/ICMLC.2012.6359637.

Dilli Arasu S, Thirumalaiselvi R. 2017. Review of chronic kidney disease based on data mining techniques. *International Journal of Applied Engineering Research*.

Eduardo Massad, Renée X. de Menezes, Paulo S. P. Silveira NRSO. 2004. *Métodos quantitativos em medicina*.

Estudillo-Valderrama MA, Talaminos-Barroso A, Roa LM, Naranjo-Hernández D, Reina-Tosina J, Aresté-Fosalba N, Milán-Martín JA. 2014. A distributed approach to alarm management in chronic kidney disease. *IEEE Journal of Biomedical and Health Informatics*. DOI: 10.1109/JBHI.2014.2333880.

Ghannad-Rezaie M, Soltanian-Zadeh H. 2008. Interactive Knowledge Discovery for Temporal Lobe Epilepsy. In: *Data Mining in Medical and Biological Research*.

InTech,. DOI: 10.5772/6411.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*.

Ho CY, Pai TW, Peng YC, Lee CH, Chen YC, Chen YT, Chen KS. 2012. Ultrasonography image analysis for detection and classification of chronic kidney disease. In: *Proceedings - 2012 6th International Conference on Complex, Intelligent, and Software Intensive Systems, CISIS 2012*. DOI: 10.1109/CISIS.2012.180.

Ilayaraja M, Meyyappan T. 2013. Mining medical data to identify frequent diseases using Apriori algorithm. In: *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*. IEEE, 194–199. DOI: 10.1109/ICPRIME.2013.6496471.

Jenn-Lung Su, Guo-Zhen Wu, I-Pin Chao. The approach of data mining methods for medical database. In: *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 3824–3826. DOI: 10.1109/IEMBS.2001.1019673.

Kumar N. 2012. Data Mining for Business Intelligence–Concepts, Techniques, and Applications in Microsoft Office Excel® with XLMiner®. *Journal of Quality Technology*. DOI: 10.1080/00224065.2012.11917885.

Kumar M. 2016. Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm Running Title: Prediction of Chronic Kidney Disease. *International Journal of Computer Science and Mobile Computing*.

Kunwar V, Chandel K, Sabitha AS, Bansal A. 2016. Chronic Kidney Disease analysis using data mining classification techniques. In: *Proceedings of the 2016 6th International Conference - Cloud System and Big Data Engineering, Confluence 2016*. DOI: 10.1109/CONFLUENCE.2016.7508132.

Lakshmi KR, Nagesh Y, Veerakrishna M. 2014. PERFORMANCE COMPARISON OF THREE DATA MINING TECHNIQUES FOR PREDICTING KIDNEY DIALYSIS SURVIVABILITY. *International Journal of Advances in Engineering & Technology*.

Lee HG, Noh KY, Ryu KH. 2008. A data mining approach for coronary heart disease prediction using HRV features and carotid arterial wall thickness. In: *BioMedical Engineering and Informatics: New Development and the Future - Proceedings of the 1st International Conference on BioMedical Engineering and Informatics, BMEI*

2008. DOI: 10.1109/BMEI.2008.189.

Lenart M, Mascarenhas N, Xiong R, Flower A. 2016. Identifying risk of progression for patients with Chronic Kidney Disease using clustering models. In: *2016 IEEE Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 221–226. DOI: 10.1109/SIEDS.2016.7489303.

Pal D, Chakraborty C, Mandana K. 2011. Data mining approach for coronary artery disease screening. In: *2011 International Conference on Image Information Processing*. IEEE, 1–6. DOI: 10.1109/ICIIP.2011.6108972.

Rajan JR, Chilambu Chelvan C. 2013. A survey on mining techniques for early lung cancer diagnoses. In: *Proceedings of the 2013 International Conference on Green Computing, Communication and Conservation of Energy, ICGCE 2013*. DOI: 10.1109/ICGCE.2013.6823566.

Sengur A. 2008. An expert system based on principal component analysis, artificial immune system and fuzzy -NN for diagnosis of valvular heart diseases. *Computers in Biology and Medicine* 38:329–338. DOI: 10.1016/j.compbiomed.2007.11.004.

Sérgio Antonio Draibe. 2014. *Panorama da Doença Renal Crônica no Brasil e no mundo*. São Luis-MA.

Sison M, Gerlai R. 2011. Associative learning performance is impaired in zebrafish (Danio rerio) by the NMDA-R antagonist MK-801. *Neurobiology of Learning and Memory*. DOI: 10.1016/j.nlm.2011.04.016.

Srinivas K, Raghavendra Rao G, Govardhan A. 2010. Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In: *ICCSE 2010 - 5th International Conference on Computer Science and Education, Final Program and Book of Abstracts*. DOI: 10.1109/ICCSE.2010.5593711.

Tazin N, Sabab SA, Chowdhury MT. 2017. Diagnosis of chronic kidney Disease using effective classification and feature selection technique. In: *1st International Conference on Medical Engineering, Health Informatics and Technology, MediTec 2016*. DOI: 10.1109/MEDITEC.2016.7835365.

Xing Y, Wang J, Zhao Z, Gao A. 2007. Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease. In: *2007 International Conference on Convergence Information Technology (ICCIT 2007)*. IEEE, 868–872. DOI: 10.1109/ICCIT.2007.204.

Xun L, Wu Xiaoming, Li Ningshan, Lou Tanqi. 2010. Application of radial basis

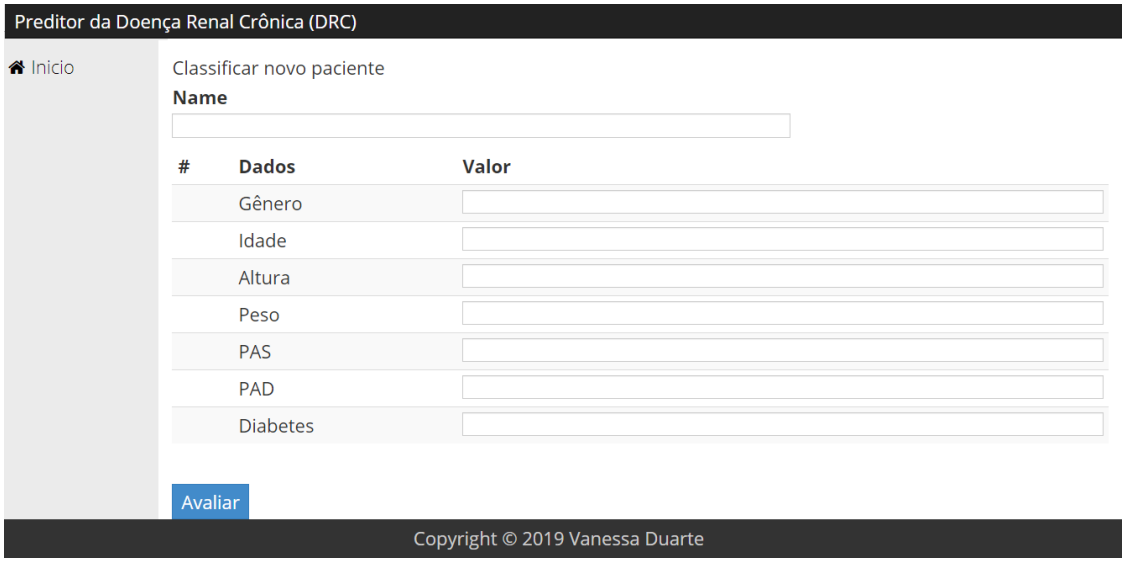
function neural network to estimate glomerular filtration rate in Chinese patients with chronic kidney disease. In: *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*. IEEE, V15-332-V15-335. DOI: 10.1109/ICCASM.2010.5622616.

Yadollahpour A, Nourozi J, Mirbagheri SA, Simancas-Acevedo E, Trejo-Macotela FR. 2018. Designing and Implementing an ANFIS Based Medical Decision Support System to Predict Chronic Kidney Disease Progression. *Frontiers in Physiology* 9. DOI: 10.3389/fphys.2018.01753.

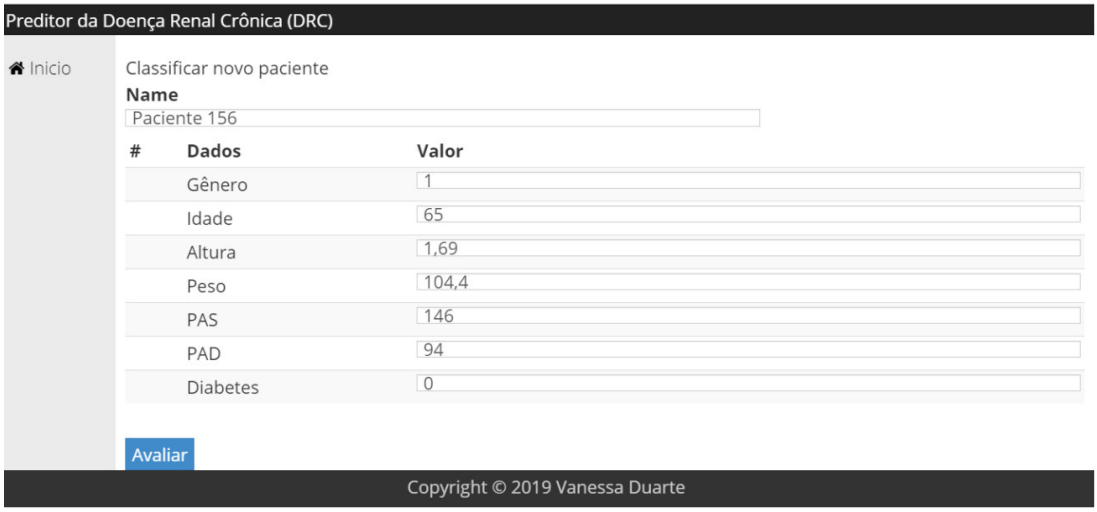
Zhao Y, Healy BC, Rotstein D, Guttmann CRG, Bakshi R, Weiner HL, Brodley CE, Chitnis T. 2017. Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLOS ONE* 12:e0174866. DOI: 10.1371/journal.pone.0174866.

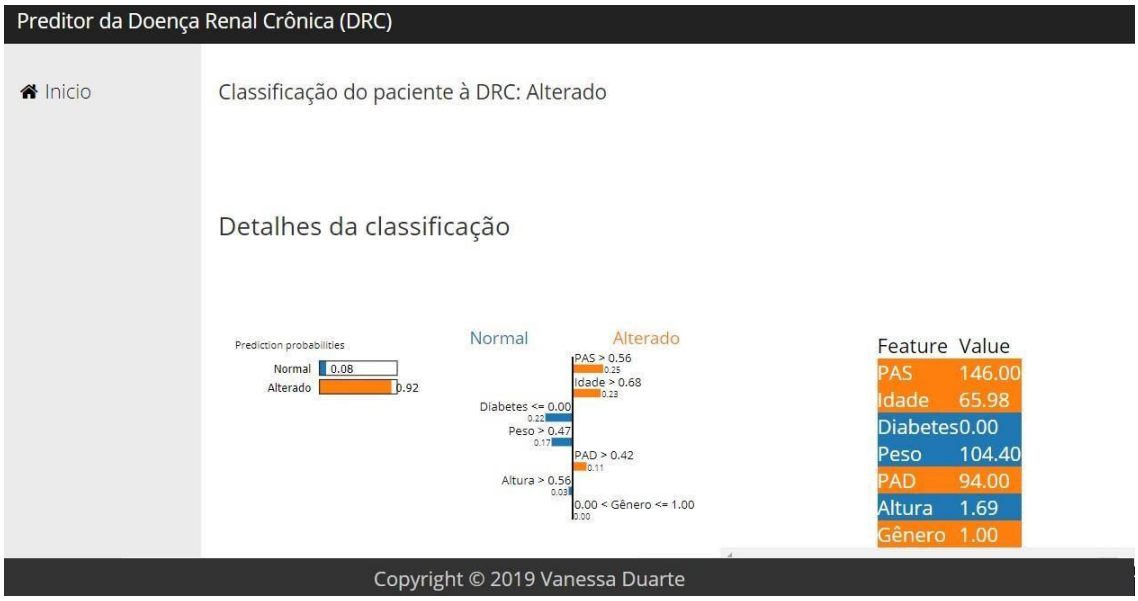
ANEXOS

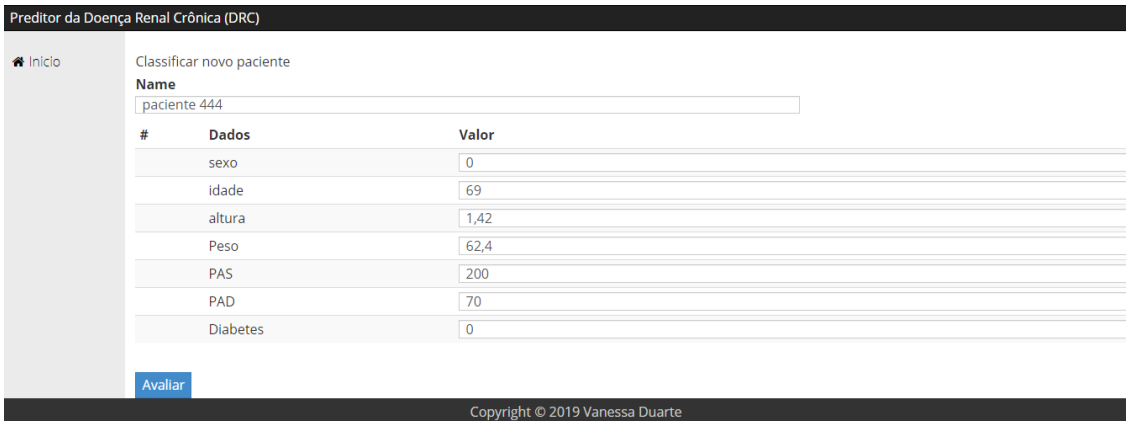
SISTEMA COMPUTACIONAL



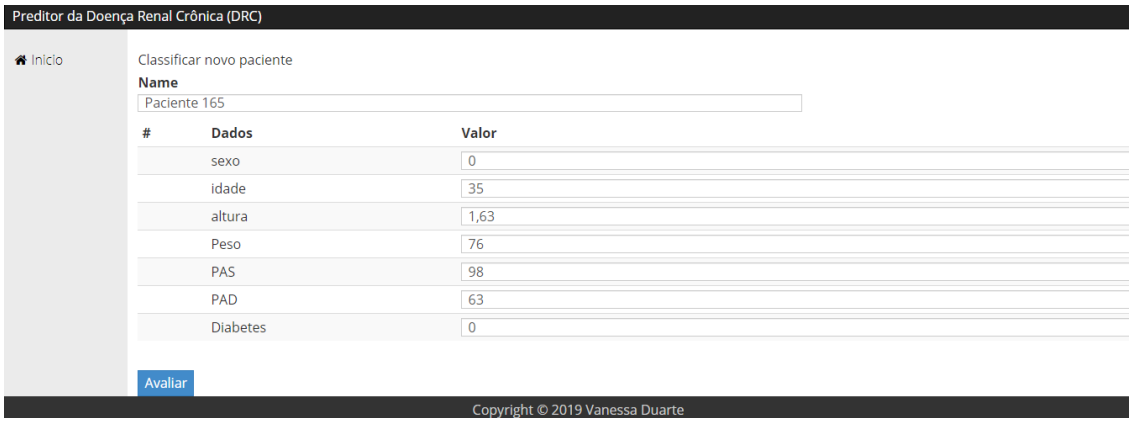
VALIDAÇÃO DO SISTEMA COMPUTACIONAL




Preditor da Doença Renal Crônica (DRC)
Inicio
Classificação do paciente à DRC: Alterado
Detalhes da classificação
Prediction probabilities
Normal 0.08
Alterado 0.92
Normal
Alterado
PAS > 0.56
Idade > 0.68
Diabetes <= 0.00
Peso > 0.47
PAD > 0.42
Altura > 0.56
0.00 < Gênero <= 1.00
Feature Value
PAS 146.00
Idade 65.98
Diabetes 0.00
Peso 104.40
PAD 94.00
Altura 1.69
Gênero 1.00
Copyright © 2019 Vanessa Duarte


Preditor da Doença Renal Crônica (DRC)
Inicio
Classificar novo paciente
Name
paciente 444
Dados Valor
sexo 0
idade 69
altura 1,42
Peso 62,4
PAS 200
PAD 70
Diabetes 0
Avaliar
Copyright © 2019 Vanessa Duarte


Preditor da Doença Renal Crônica (DRC)
Inicio
Classificação do paciente à DRC: Alterado
Detalhes da classificação
Prediction probabilities
Normal 0.00
Alterado 1.00
Normal
Alterado
Peso > 0.48
Idade > 0.66
altura > 0.58
PAS > 0.41
PAD > 0.45
Diabetes <= 0.00
sexo <= 0.00
Feature Value
Peso 62.40
Idade 69.00
altura 1.42
PAS 200.00
PAD 70.00
Diabetes 0.00
sexo 0.00


Preditor da Doença Renal Crônica (DRC)
Inicio
Classificar novo paciente
Name
Paciente 165
Dados Valor
sexo 0
idade 35
altura 1,63
Peso 76
PAS 98
PAD 63
Diabetes 0
Avaliar
Copyright © 2019 Vanessa Duarte


Preditor da Doença Renal Crônica (DRC)
Inicio
Classificação do paciente à DRC: Normal
Detalhes da classificação
Prediction probabilities
Normal 0.50
Alterado 0.50
Normal
Alterado
Peso > 0.49
idade > 0.66
altura > 0.58
PAS > 0.41
PAD > 0.45
sexo <= 0.00
Diabetes <= 0.00
Feature Value
Peso 76.00
Idade 35.00
altura 1.63
PAS 98.00
PAD 63.00
sexo 0.00
Diabetes0.00
Copyright © 2019 Vanessa Duarte