

Previsão de doença de diabetes usando máquina Aprendizagem em Big Data de Saúde

Ayman Mir

Departamento de Engenharia Informática

Instituto de Tecnologia Sardar Patel

Mumbai, Índia

ayman.mir@spit.ac.in

Sudhir N. Dhage

Departamento de Engenharia Informática

Instituto de Tecnologia Sardar Patel

Mumbai, Índia

sudhir.dhage@spit.ac.in

Resumo — O domínio da saúde é uma pesquisa muito proeminente campo com rápido avanço tecnológico e dados crescentes dia a dia. Para lidar com grande volume de cuidados de saúde dados, precisamos de Big Data Analytics, uma abordagem emergente no domínio Saúde. Milhões de pacientes procuram tratamento em todo o mundo com vários procedimentos. Analisando as tendências no tratamento de pacientes para diagnóstico de uma doença específica ajudará na tomada de decisões informadas e eficientes para melhorar a qualidade geral dos cuidados de saúde. O aprendizado de máquina é muito abordagem promissora que ajuda no diagnóstico precoce de doenças e pode ajudar os médicos na tomada de decisão para o diagnóstico. Este artigo tem como objetivo construir um modelo de classificador utilizando a ferramenta WEKA. para prever doenças diabéticas, empregando Naive Bayes, Support Máquina de vetores, floresta aleatória e algoritmo CART simples.

A pesquisa espera recomendar o melhor algoritmo baseado em resultado de desempenho eficiente para a previsão de doenças do diabetes.

Os resultados experimentais de cada algoritmo usado no conjunto de dados foram avaliados. Observa-se que a máquina de vetores de suporte executou melhor na previsão da doença com precisão máxima.

Palavras-chave - Saúde, Big Data, Aprendizado de Máquina, Doenças Predição, Naive Bayes, Máquina de vetores de suporte, Floresta aleatória, CART Simples

I. INTRODUÇÃO

A saúde é sempre uma prioridade, mesmo antes de a tecnologia existir.

O domínio da saúde oferece um amplo escopo de pesquisa, pois tem tremendamente evoluído. É necessário atualizar o tecnologia de saúde existente, adotando a digitalização de informações médicas, tanto em termos de dados fornecidos pelo paciente quanto bem como resultados médicos gerados a partir de equipamentos avançados. Um resultado comum desta revolução da informação é que estamos confrontados com a difícil tarefa de interpretar e compreender os enormes dados coletados. Uma vez que há uma grande quantidade de dados portanto, Big Data Analytics vem para resgatar. Big Data Analytics é a abordagem emergente que é lidando com diversos setores aplicando-o à saúde domínio resultará em melhores serviços de saúde. Enquanto abordagem de big data em saúde ainda está em fase de desenvolvimento, é claro que a concepção de uma plataforma de saúde para um melhor amanhã será de grande ajuda para melhorar a qualidade prestação de cuidados de saúde. O aprendizado de máquina é outro aplicativo emergente e popular abordagem que trabalha de perto para resolver os problemas em tempo real.

Atualmente no domínio da saúde, vários dados de mineração métodos são usados para encontrar um padrão interessante de doença usando

dados médicos estatísticos com a ajuda do aprendizado de máquina algoritmos. A abordagem de aprendizado de máquina pode ser aplicada para previsão de doenças e fornecer diagnóstico automatizado sob a validação do profissional médico.

O artigo se concentra na previsão da doença crítica diabetes. Diabetes é classificada como a quinta doença mais mortal do mundo

ampla. A figura a seguir, Fig. 1, representa o relatório da OMS para Diabetes.

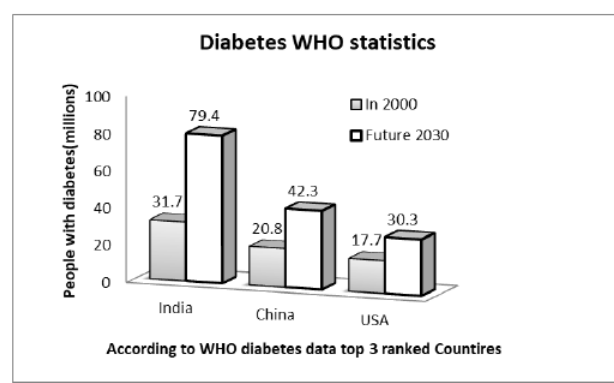


Fig. 1. Estatísticas de diabetes da OMS

De acordo com o relatório da OMS, a Índia é classificada como No. 1 com 31,7 milhões não. de paciente diabético em 2000 e é provável para aumentar até 79,4 milhões. Uma vez que existe um grande risco probabilidade de aumentar não. do paciente diabético, onde preciso o diagnóstico será a necessidade da hora. Daí esta estatística fornece a motivação para realizar a pesquisa para encontrar o algoritmo de melhor desempenho. Neste artigo, faremos uso de quatro algoritmos de aprendizagem supervisionada e realizar comparação em seu resultado para recomendar a melhor abordagem para prever o diabetes, que será útil para tomar decisões informadas com precisão.

O restante do artigo está organizado da seguinte forma : Seção II é Pesquisa de Literatura descrevendo o que já existe trabalho. Seção III é sobre Metodologia que destaca o conjunto de dados em trabalho e a metodologia proposta. A seção IV descreve os resultados experimentais que são obtidos depois de criar classificadores por meio do WEKA. Seção V é sobre Resultados e Discussões que discute sobre o avaliação de desempenho de todos os classificadores. Seção VI é sobre Conclusão que conclui os resultados gerais.

II. LITERATURE SURVEY

Esta seção analisa o trabalho de literatura recente existente e fornece insights na compreensão dos desafios e tenta encontrar as lacunas nas abordagens existentes.

Várias técnicas de computação são aplicadas na área da saúde domínio. O foco da pesquisa de literatura aqui está no uso de Big Data Analytics e Machine Learning na área de saúde domínio. A fim de fazer um sistema de saúde de aprendizagem inteligente existem desafios de dados analíticos não resolvidos [1]. Através dos Análise de big data, a relação entre os padrões de dados são compreendida e valor adicional do enorme sistema de saúde os dados são descobertos [2]. Existem várias tendências de pesquisa e desafios ao longo do ciclo de vida dos dados durante a implementação Big Data Analytics e está bem descrito em [3] [4] [5]. Saúde- os desafios do domínio do cuidado estão na melhoria das fases de pesquisa. a mineração da realidade é uma nova abordagem, ou seja, usando big data para estudar o comportamento do paciente através de sensores de telefones celulares que ajuda a melhorar a qualidade da saúde [5]. Extraíndo útil as informações do Electronic Health Record são pesquisadas em [6]. Um projeto inteligente usando Big Data é proposto em [7], o qual é um aplicativo baseado na web que fornece uma plataforma eficiente para simplificar a avaliação complexa da saúde, juntamente com procedimento de monitoramento.

Atualmente no domínio dos cuidados de saúde, implementando algoritmo de aprendizagem de máquina e fazendo uso de estatísticas médicas dados padrões interessantes em dados de doenças são descobertos [8] [9]. Para melhor diagnóstico e prognóstico, o aprendizado de máquina pode ser usado de forma eficaz é demonstrado com dois estudos de caso para diagnóstico da doença em [10] [11]. O suporte de decisão do especialista sistema usando técnicas de mineração de dados para classificação de doenças é proposto em [12], onde se concentra no diagnóstico de diabetes doença e usa a árvore de decisão e o algoritmo KNN. EM- sistema de psiquiatria é um sistema de ambiente inteligente em atendido para atendimento psiquiátrico de emergência e faz uso de técnica de aprendizado de máquina que é a entropia máxima de Markov modelo (MEMM) [13]. Outro sistema é o iHANDS que é uma saúde inteligente e agente de suporte à decisão construído com vários mecanismos de inteligência artificial que auxiliam indivíduos em seu processo de tomada de decisão em saúde [14]. UMA abordagem semelhante de aprendizado de máquina é aplicada a outro E-sistema de medicação para diagnóstico de doença falciforme onde tenta preencher a lacuna entre o médico e o paciente por meio aplicativo no smartphone [15]. Desde a máquina domínio de aprendizagem consiste em várias técnicas que as pesquisas fazer comparações de modo a que seja eficiente e mais rápido resulta em previsão. Uma comparação realizada usando SVM e implementou redes neurais perceptron multicamadas para previsão de doenças cardíacas em que o SVM oferece maior precisão [16]. O SVM tem preferência sobre outro aprendizado de máquina algoritmos devido à sua precisão é melhor em comparação [16]. A seguinte tabela TABELA I consiste em várias técnicas que diferentes pesquisadores usaram para fazer comparações e seu conjunto de dados de trabalho é mencionado junto com o resultado e as possíveis limitações são listadas

Table I: Comparison of various Literature Works

Ref	Methodology	Dataset	Outcomes	Limitations
Bhargava et al. 2017 [17]	Simple CART algorithm in WEKA to predict heart attack	Real world Male Heart disease dataset. Instances Used = 209	Accuracy of correctly classified instances is 79.9 %	Only one algorithm considered therefore couldn't state if it is best approach.
Dhomse, Ma-hale, 2016 [18]	SVM, Decision Tree and Naive Bayes applied with and without feature selection to predict heart disease	Heart disease dataset from Cleveland Clinic Foundation. Instances Used = 303	After reducing dataset SVM outperforms Naive Bayes	Accuracy results not mentioned direct graph plotted.
Dhomse, Ma-hale, 2016 [18]	SVM and Naive Bayes applied with and without feature selection to predict diabetes disease using WEKA tool	Diabetic patients dataset is collected from hospital repository. Instances Used = 1865	Naive Bayes have better accuracy results and takes less time for building the training model than SVM	Classification Accuracy of Naive Bayes is 34.89% which is quite risky for prediction
Ramzan, 2016 [19]	Naive Bayes, J48 Decision Tree, Random Forest are used to compare classifiers to predict critical disease using WEKA tool	Disease classification dataset collected from Global Health Data Exchange. Instances Used = 9242	Random Forest turns out with an Accuracy of 99.83% beating both Naive Bayes and J48	Random Forest requires more time for building the training model
Naik, Samant, 2016 [20]	Decision Tree, K-Nearest Neighbours, Naive Bayes are completely reviewed to predict liver disorder using WEKA, Orange, Tungara, KNIME and Rapid miner tool	Liver patient dataset collected from Indian Liver Patient Dataset. Instances Used = 583	As for tool's performance all three algorithms performed well using KNIME tool. As for algorithm then Decision Tree and KNN outperformed Naive Bayes using all the tools	Requires a powerful machine learning to analyze the outcome of model using all tools to improve classification accuracy.
Iyer et al., 2015 [21]	J48 Decision Tree and Naive Bayes approach for diagnosis of diabetes	Pima Indians Diabetes Database. Instances Used = 768	Naive Bayes gives least error rate and thus outperforms J48 decision Algorithm	Comparison of only 2 algorithm is not sufficient to build best diagnosis model.

III. METODOLOGIA

Esta seção inclui a metodologia que descreve a abordagem que é usado para realizar a pesquisa, a fim de realizar análise comparativa

A. Descrição da ferramenta WEKA

A ferramenta WEKA é brevemente descrita abaixo:

WEKA [Waikata Enviroment for Knowledge Analysis]

- É um aprendizado de máquina e mineração de dados muito popular

kit de ferramentas para a realização de pesquisas orientadas a dados.

- Desenvolvido na Nova Zelândia na Universidade de Waikato

- A coleção de aprendizado de máquina e mineração de dados

algoritmos presentes são escritos em Java

- A versão do WEKA usada para experimentação neste o papel é WEKA versão 3.82

A pesquisa fez uso da ferramenta WEKA, pois ajuda na realização

avaliação de desempenho e comparação de desempenho de vários

técnicas de aprendizado de máquina convenientemente em dados em tempo real.

B. Conjunto de dados de doença de diabetes

Aqui está a descrição do conjunto de dados que foi usado como

uma entrada para classificadores implementados usando vários algoritmos.

O nome do conjunto de dados que foi considerado é **Pima Banco de dados de diabetes dos índios** coletado no National Instituto de Diabetes e Doenças Digestivas e Renais. O número total de instâncias é 768 e o tamanho é 37 KB. O total não. de atributos são 9 incluindo a classe alvo atributo. O nome de duas classes de destino são testados positivamente

e testado negativo. O não. de instâncias para teste positivo são 268 e o não. de ocorrências para teste negativo são 500.

O pré-processamento de dados é realizado automaticamente pelo WEKA ferramenta.

A seguinte tabela TABELA II descreve os 9 atributos de o conjunto de dados de diabetes brevemente

Tabela II: Conjunto de dados de doença de diabetes

Sr No.	Attribute Used	Attribute Type	Attribute Description
1	preg	Numeric	No. of times pregnant
2	plas	Numeric	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	pres	Numeric	Diastolic blood pressure (mm Hg)
4	skin	Numeric	Triceps skin fold thickness (mm)
5	insu	Numeric	2-Hour serum insulin (mu U/ml)
6	mass	Numeric	Body mass index (weight in kg / (height in square m)
7	pedi	Numeric	Diabetes pedigree function
8	age	Numeric	Age (years)
9	Class	Nominal	Class variable (tested_positive or tested_negative)

C. Fluxograma da metodologia proposta

Aqui está uma breve descrição sobre o fluxo de propostas metodologia.

O modelo do Classificador proposto especificamente considerado diabetes doença e leva entrada do conjunto de dados para diabetes. O conjunto de dados de entrada é processado usando quatro aprendizado de máquina algoritmos que são Naive Bayes, SVM, Random Forest, Simple CART e para cada algoritmo respectivo modelo de classificador é treinado e testado e os resultados são coletados. Baseado em os resultados experimentais, o algoritmo de melhor desempenho pode ser determinado, o que ajudará na previsão precisa do doença.

A figura a seguir, Fig 2. mostra a abordagem que tem foi aplicado para realizar a análise comparativa a fim de para recomendar o melhor algoritmo para a construção de classificação de modelo para prever a doença diabética.

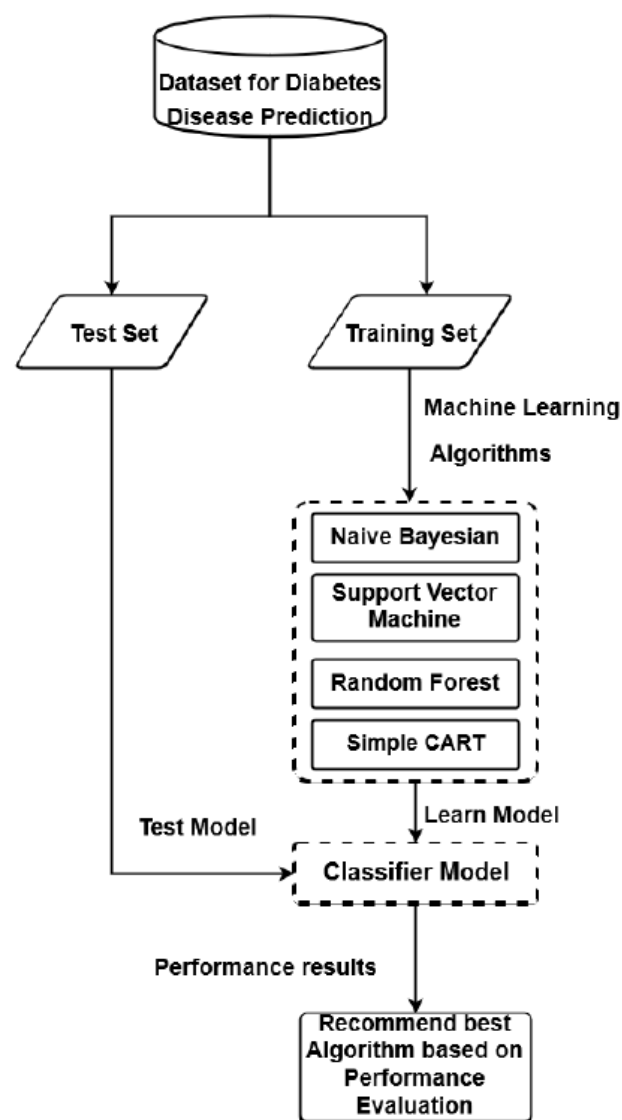


Fig. 2. Fluxograma da metodologia proposta

O seguinte descreve as etapas envolvidas no procedimento da Figura 2. Metodologia do Classificador Proposta

Procedimento Stepwise da Metodologia Proposta

- Etapa 1: - Pré-processar o conjunto de dados de entrada para diabetes
- doença na ferramenta WEKA

- **Etapa 2:** - Realize a divisão percentual de 70% para dividir conjunto de dados como conjunto de treinamento e conjunto de teste
- **Etapa 3:** - Selecione o algoritmo de aprendizado de máquina, ou seja Naive Bayes, máquina de vetores de suporte, floresta aleatória e algoritmo CART simples.
- **Etapa 4:** - Construir o modelo do classificador para o mencionado algoritmo de aprendizado de máquina com base no conjunto de treinamento.
- **Etapa 5:** - Teste o modelo do classificador para o algoritmo de aprendizado de máquina baseado em conjunto de teste
- **Etapa 6:** - Realizar avaliação de comparação da experiência resultados de desempenho mental obtidos para cada classificador.
- **Etapa 7:** - Depois de analisar com base em várias medidas conclua o algoritmo de melhor desempenho. O modelo do classificador proposto foi construído usando WEKA ferramenta e com base na execução bem-sucedida de cada etapa, podemos avaliar os resultados experimentais.

4. RESULTADOS EXPERIMENTAIS

Esta seção descreve os resultados experimentais que são obtido após o treinamento Naive Bayes, Support Vector Machine, Classificadores Random Forest e Simple CART no diabetes conjunto de dados do paciente. O propósito desses resultados experimentais são para avaliação de desempenho de todos os quatro classificadores e para recomendar o melhor algoritmo adequado para previsão.

A. Matriz de confusão

No aprendizado de máquina, uma matriz de confusão é usada para analisar o desempenho do algoritmo de classificação. A confusão matrix é uma estrutura tabular onde as linhas representam o real classe e colunas representam a classe prevista.

CONFUSION MATRIX STRUCTURE			
Total no. of instances		Predicted Class	
		No a=tested_negative	Yes b=tested_positive
Actual Class	No a=tested_negative	TrueNegative	FalsePositive
	Yes b=tested_positive	FalseNegative	TruePositive

Fig. 3. Estrutura geral da matriz de confusão

Determinada terminologia que aparece na Confusão geral A estrutura da matriz é descrita a seguir. Essa terminologia irá ser usado posteriormente para avaliação de desempenho de cada classificador.

- **Classe real:** rótulo de classe que representa a classe real antes de construir o classificador
- **Classe prevista:** rótulo de classe que representa o previsto Aula depois de construir o classificador
- **TruePositive:** Número de casos previstos positivos e são realmente positivo
- **FalsePositive:** Número de instâncias previstas negativas e são realmente negativos
- **TrueNegative:** Número de casos previstos positivos, mas são realmente negativos

- **FalseNegative:** número de instâncias previstas negativas, mas são realmente positivos

- **Nº total de instâncias:** a soma de todas as instâncias que foram classificados pelo classificador.

A matriz de confusão assim obtida após a construção da classe usando Naive Bayes, Support Vector Machine, Random Floresta, algoritmo de aprendizado de máquina simples CART no WEKA ferramenta são mostrados abaixo:

```

Naive Bayes
=== Confusion Matrix ===

 a  b  <-- classified as
133 25 | a = tested_negative
 28 44 | b = tested_positive

```

Fig. 4. Matriz de confusão Naive Bayes

Conforme a Fig 4. De acordo com a matriz de confusão Naive Bayes os valores de TrueNegative = 133, FalseNegative = 28, FalsePositive = 25, TruePositive = 44

```

Support Vector Machine
=== Confusion Matrix ===

 a  b  <-- classified as
143 15 | a = tested_negative
 33 39 | b = tested_positive

```

Fig. 5. Matriz de confusão de vetores de suporte

Conforme a Fig. 5. De acordo com o Support Vector Machine Confusão Matrix os valores de TrueNegative = 143, FalseNegativo = 33, FalsePositive = 15, TruePositive = 39.

```

Random Forest
=== Confusion Matrix ===

 a  b  <-- classified as
138 20 | a = tested_negative
 34 38 | b = tested_positive

```

Fig. 6. Matriz de confusão florestal aleatória

Conforme a Figura 6. De acordo com a confusão florestal aleatória Matriz os valores de TrueNegative = 138, FalseNegative = 34, FalsePositive = 20, TruePositive = 38.

```

Simple CART
=== Confusion Matrix ===

 a  b  <-- classified as
133 25 | a = tested_negative
 29 43 | b = tested_positive

```

Fig. 7. Matriz de confusão CART simples

Conforme a Fig 7. De acordo com Simple CART Confusion Matrix os valores de TrueNegative = 133, FalseNegative = 29, FalsePositive = 25, TruePositive = 43.

B. Precisão de classificação

A precisão da classificação é uma das avaliações de desempenho medida de uação. A precisão representa o quão bem o classificador realiza a previsão das instâncias com base nos dados de treinamento.

• **Precisão:** é a proporção do não. de verdadeiro predito exemplo positivo e negativo para o total não. de instâncias.

$$\text{Precisão} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{Nº total de instâncias}}$$

A seguinte tabela Tabela III representa o experimental resultados de precisão de classificação de Naive Bayes, Vetor de Suporte Máquina, floresta aleatória e algoritmo CART simples. O tabela exibe o tempo de treinamento, tempo de teste e precisão valor de cada algoritmo.

Tabela III: Resultados de precisão de classificação experimental

Algorithm	Training Time	Testing Time	Accuracy Value
Naive Bayes	0.03 sec	0.02 sec	0.77
Support Vector Machine	0.14 sec	0.03 sec	0.7913
Random Forest	0.67 sec	0.06 sec	0.765
Simple CART	1.38 sec	0.02 sec	0.765

A figura a seguir, Fig. 8, mostra o Classificação de Precisão gráfico de valor plotado de todos os quatro classificadores

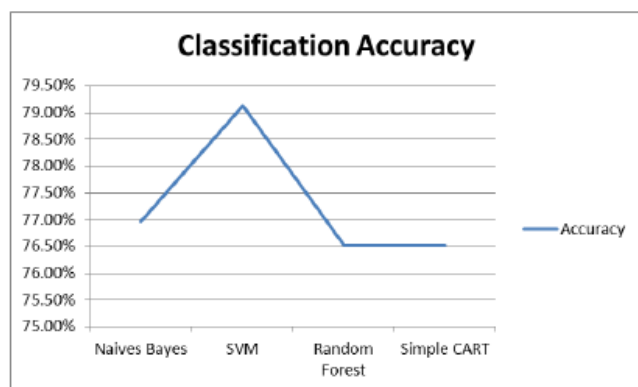


Fig. 8. Valores de precisão de classificação

C. Valores de medida de precisão

A seguir estão os valores de medida de precisão do classificador
Descrição:

• **TP - Taxa:** É a relação do nº. de predito positivo instâncias ao total real não. de exemplos positivos

$$\text{TP - Rate} = \frac{\text{True Positive}}{\text{TruePositive} + \text{FalseNegative}}$$

• **FP - Taxa:** É a relação do nº. de negativo previsto instâncias ao total real não. de instâncias negativas

$$\text{Taxa de FP} = \frac{\text{Falso positivo}}{\text{FalsePositive} + \text{TrueNegative}}$$

• **Precisão:** é a proporção de não. de positivos previstos instâncias ao total de todas as instâncias positivas previstas.

$$\text{Precisão} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}$$

• **Recall:** É a proporção do não. de predito positivo instâncias ao total real não. de exemplos positivos

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}$$

• **F- Measure:** usado para representar o desempenho geral. Isto é a média harmônica ponderada da precisão e recall

$$\text{F-Medida} = \frac{2 * \text{Precisão} * \text{Rechamada}}{\text{Precisão} + \text{recall}}$$

A seguinte tabela Tabela IV representa a Medida de Precisão Valor de todos os quatro classificadores obtidos após a construção de todos quatro classificadores no conjunto de dados de diabetes no WEKA

Tabela IV: Valores da Medida de Precisão Principal da Classificação

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure
Naive Bayes	0.770	0.317	0.767	0.770	0.768
Support Vector Machine	0.791	0.345	0.784	0.791	0.782
Random Forest	0.765	0.326	0.756	0.765	0.758
Simple CART	0.765	0.364	0.762	0.763	0.446

A figura a seguir, Fig. 9., mostra a Medida de Precisão Valor de todos os quatro classificadores com base nos valores obtidos após experimentação que é representada na Tabela IV

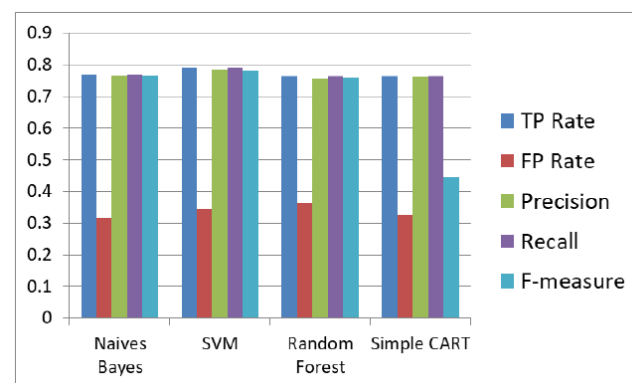


Fig. 9. Resultados da medição de precisão do classificador

V. RESULTADOS E DISCUSSÕES

Esta seção discute os resultados experimentais gerais, portanto obtido através da ferramenta WEKA.

De acordo com a figura de precisão de classificação Fig 8. A precisão do SVM é a mais alta, que é 0,7913. A precisão de Naive Bayes é 0,77 melhor do que Random Floresta e CART simples. A precisão de Random Forest e Simple CART é quase igual ao valor 0,765. O o tempo de treinamento de Naive Bayes é menor que o de SVM. O treinamento tempo de Simple CART é o mais alto. Geral de acordo com classificação Precisão O SVM superou todos os outros classificadores. De acordo com a Medida de Precisão de Classificação descrita na figura Fig 9. a máquina de vetores de suporte tem o maior F-valor de medida de 0,782 e o mínimo é de CART Simples que é 0,446. O valor de precisão da Máquina de vetores de suporte é mais alto com um valor de 0,784 e o valor de precisão de Random Forest é o mínimo com um valor de 0,756

VI. CONCLUSÃO

Neste trabalho de pesquisa, quatro classificadores baseados em máquina algoritmo de aprendizagem que são Naive Bayes, Vetor de SuporteMachine, Random Forest e Simple CART foram usado para experimentação na ferramenta WEKA para prever Diabetes doença. Os quatro classificadores assim construídos foram comparados com base no tempo de treinamento, tempo de teste e valor de precisão. Outro método de avaliação de desempenho foi classificador medida de precisão que incluiu taxa de TP, taxa de FP, precisão, recorde, F-Measure. O desempenho geral do Suporte Máquina de vetores para prever a doença do diabetes é melhor do que Naive Bayes, Random Forest e Simple Cart. Conseqüentemente a eficácia do modelo proposto é claramente representada ao longo dos resultados experimentais mencionados.

REFERÊNCIAS

[1] Rahul C. Basole, Mark L. Braunstein e Jimeng Sun, "Data e Desafios analíticos para um sistema de saúde de aprendizagem", ACM Journal of Data and Information Quality, vol. 6, No. 2-3, Artigo 10, Publicação data: julho de 2015

[2] J.Archenaa and EA Mary Anita, "A Survey Of Big Data Analytics em Saúde e Governo", 2º Simpósio Internacional do Big Dados e computação em nuvem (ISBCC'15), 2015

[3] Agusti Solanas, Fran Casino, Edgar Batista e Robert Rallo, "Trends e desafios na pesquisa de saúde inteligente: uma jornada dos dados para Sabedoria", IEEE 2017

[4] Fuad Rahman, "Application of Big-Data in Healthcare Analytics -Perspectivas e desafios", IEEE 2017

[5] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime e Thomas Noel, "Big Data em saúde: desafios e oportunidades", IEEE 2015

[6] Pranjul Yadav, Michael Steinbach, Vipin Kumar e Gyorgy Simon, "Mining Electronic Health Records (EHRs): A Survey", ACM Com-colocando Surveys, vol. 50, nº 6, artigo 85. Data de publicação: janeiro2018.

[7] Weider D. Yu, Jaspal Singh Gill, Maulin Dalal, Piyush Jha e SajjanShah, "Big Data Approach in Healthcare used for Intelligent Design", 2016 IEEE International Conference on Big Data (Big Data)

[8] Rohan Bhardwaj, Ankita R. Nambiar e Debojyoti Dutta, "A Study of Machine Learning in Healthcare", 2017 IEEE 41st Annual Computer Conferência de Software Aplicativos

[9] Athmaja S., Hanumanthappa M. e Vasantha Kavitha, "A Survey of Machine Learning Algorithms for Big Data Analytics", 2017 em Conferência internacional sobre inovações em informação, incorporada e Sistemas de Comunicação (ICIIECS)

[10] Oleg Roderick, Nicholas Marko, David Sanchez e Arun Aryasomajula, "Data Analysis And Machine Learning Effort In Healthcare: Organização, Limitações e Desenvolvimento de uma Abordagem", Internet of Things and Data Analytics Handbook, First Edition. 2017

[11] Niharika G. Maity, Dr. Sreerupa Das, "Machine Learning for Improved Diagnóstico e Prognóstico em Saúde", IEEE 2017

[12] Emrana Kabir Hashi, Md. Shahid Uz Zaman, Md. Rokibul Hasan, "Na Sistema de Apoio à Decisão Clínica Especializada para Prever Doenças Usando Clas-sification Techniques", Conferência Internacional de Eletricidade, Computação e Engenharia de Comunicação (ECCE), 16-18 de fevereiro de 2017, IEEE

[13] Md. Golam Rabiul Alam, Rim Haw, Sung Soo Kim, Md. Abul Kalam Azad, Sarder Fakhrul Abedin, Choong Seon Hong, "EM-Psychiatry: Um Sistema Inteligente Ambiental para Emergências Psiquiátricas", IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 12, NO. 6, DEZEMBRO 2016

[14] Brett Hannan, Xiaoqin Zhang, Kristen Sethares, "iHANDs: Intelligent Agente de Aconselhamento de Saúde e Apoio à Decisão", 2014 IEEE / WIC / ACM Conferências Conjuntas Internacionais sobre Web Intelligence (WI) e IntelligentTecnologias de Agente (IAT)

[15] Dhafar Hamed Abd, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B Naeem, "The Utilization of Machine Learning Approaches for Medical Classificação de dados e gerenciamento do sistema de cuidados pessoais para foice Cell Disease", Conferência Anual sobre Novas Tendências em Informação e Aplicações de tecnologia de comunicação (NTICT'2017) 7 9 de março 2017, IEEE 2017

[16] Parisa Naraei, Abdolreza Abhari, Alireza Sadeghian, "Pedido de Redes Neurais Perceptron Multicamadas e Máquinas Vetoriais de Suporte in Classification of Healthcare Data", FTC 2016 - Future Technologies Conferência 2016, 6 a 7 de dezembro de 2016 - São Francisco, Estados Unidos, IEEE 2016

[17] Dr. Neeraj Bhargava, Sonia Dayma, Abishek Kumar, Pramod Singh, "An Abordagem para classificação usando algoritmo CART simples em Weka", 2017 11ª Conferência Internacional de Sistemas Inteligentes e Controle (ISCO), IEEE 2017

[18] Dhomse Kanchan B., Sr. Mahale Kishor M, "Study of Machine Algoritmos de aprendizagem para previsão de doenças especiais usando o Principal Component Analysis", 2016 Conferência Internacional sobre Tendências Globais em Processamento de Sinais, Computação da Informação e Comunicação, IEEE 2016

[19] Munaza Ramzan, "Comparing and Evaluating the Performance of WEKA Classifiers on Critical Diseases", 2016 1st India International Conferência sobre Processamento de Informação (IICIP), IEEE 2016

[20] Amrita Naik, Lilavati Samant, "Revisão de correlação da classificação algoritmo usando ferramenta de mineração de dados: WEKA, Rapidminer, Tanagra, Orange e Knime", Conferência Internacional sobre Modelagem Computacional e Segurança (CMS 2016), Procedia Computer Science 85 (2016) 662 -668

[21] Aiswarya Iyer, S. Jeyalatha e Ronak Sumbaly, "Diagnosis Of Diabetes Using Classification Mining Techniques", International Journal of Processo de Mineração de Dados e Gerenciamento de Conhecimento (IJDGP) Vol.5, No.1, Janeiro de 2015