Universidade Federal do Maranhão
Coordenacão de Pós-Graduação em Engenharia de Eletricidade
Curso de Pós-graduação em Pós-Graduação em Engenharia de Eletricidade

Áurea Celeste da Costa Ribeiro

Rastreamento não invasivo para diabetes tipo 2

São Luís
2015

Áurea Celeste da Costa Ribeiro

Rastreamento não invasivo para diabetes tipo 2

Tese apresentada ao Curso de Pós-Graduação em Engenharia de Eletricidade da UFMA, como requisito para a obtenção do grau de doutora em Engenharia de Eletricidade

Orientador: Prof. Dr. Allan Kardec Duailibe Barros Filho

Universidade Federal do Maranhão

Co-orientador: Prof. Dr. Ewaldo Eder Santana

Universidade Estadual do Maranhão

São Luís

2015

Ribeiro, Áurea

Rastreamento não invasivo para diabetes tipo 2 / Áurea Ribeiro - 2015

xx.p

1.Automação e Controle 2. Processamento de sinais.. I.Título.

CDU 536.21

Áurea Celeste da Costa Ribeiro

Rastreamento não invasivo para diabetes tipo 2

Tese apresentada ao Curso de Pós-Graduação em Engenharia de Eletricidade da UFMA, como requisito para a obtenção do grau de doutora em engenharia de eletricidade.

Aprovado em 07 de Agosto de 2015

BANCA EXAMINADORA

Prof. Dr. Allan Kardec Duailibe Barros Filho

Universidade Federal do Maranhão

Prof. Dr. Dráulio Barros de Aráujo

Universidade Federal do Rio Grande do Norte

Prof. Dr. José Felipe Souza de Almeida

Universidade Federal Rural da Amazônia

Prof. Dr. Orlando José dos Santos

Universidade Federal do Maranhão

Prof. Dr. Francisco das Chagas de Souza

Universidade Federal do Maranhão

A Deus que me fortalece com esperança e saúde a cada novo dia.

A minha família pelo incentivo, amor e compreensão.

Resumo

O rastreamento do diabetes tipo 2 tornou-se um recurso importante devido ao grande aumento desta doença no mundo moderno, estima-se que haja 385 milhões de diabéticos no mundo e que 46% deste número desconhece sua condição. Isto dificulta seu tratamento e muitos pacientes no diagnóstico já apresentam alguma complicação devido a falta deste nos estágios iniciais da diabetes. Haviam discussões sobre a efetividade do rastreamento para diabetes tipo 2, no Brasil por exemplo, o último rastreamento teve um custo considerado desnecessário, de quase 40 milhões de reais. Métodos mais simples e eficazes de rastreio são estudados, como nos EUA e China que utilizam alguns métodos não invasivos para calcular o risco de diabetes. Este estudo propõe um método de rastreamento não invasivo baseado na técnica de codificação eficiente para extrair características de uma base de dados brasileira(HIPERDIA) para formar uma nova representação concisa destes, com a diminuição de redundância. A principal hipótese trabalhada nesta fase foi a busca das componentes independentes, que possivelmente estiveram presentes na formação da doença. Desta forma, os dados originais foram decompostos pelo método de análise de componentes independentes. Na fase de classificação para assegurar a discriminação entre as classes utilizou-se o método de maquinas de vetores de suporte para uma classe. Testes foram feitos para verificar o desempenho do classificador após à fase de extração de características, e mostraram que ela aumenta o desempenho da máquina de vetor de suporte para uma classe em fazer a discriminação entre diabéticos e não diabéticos. Alcançou-se resultados de (100%)com a combinação de certas características, e o método demonstra a promessa em obter-se um rastreamento de diabetes não invasivo confiável. Outros testes foram feitos para verificar a influência de cada marcador não invasivo no resultado final e a generalidade do método utilizando outras bases de dados, como a base de índios Pima e de americanos de origem africana. Diminuindo o número de características utilizadas para treinar o método e testando-se todas as possibilidades de combinações entre as restantes, retirando-se uma a uma, com um total de 12.910 possibilidades. Observou-se as características que mais afetavam no resultado final foram idade e as características relacionadas com a gordura corporal. Testando-se a generalidade do método em outras bases de dados verificou-se que o método trabalha melhor com bases balanceadas.

Palavras-chaves: Processamento de sinais, Rastreamento, Diabetes, Não Invasivo

---

Abstract

The type 2 diabetes screening has become an important resource due to the increase in this disease in the modern world, it is estimated that there are 385 millions of diabetics in worldwide and that 46% of this number are unaware of their condition. This complicates their treatment and many patients at diagnosis already present any complications due to lack this in the early stages of diabetes. Researchers have discussed the effectiveness of type 2 diabetes screening, for example: The Brazil made a screening in 2001 and it was considered an unnecessary cost of almost 40 million. The tracking of type 2 diabetes has become an important resource due to the large increase in this disease in the modern world, it is estimated that there are 385 million diabetics worldwide and that 46% of this number are unaware of their condition. This complicates their treatment and many patients the diagnosis already present any complications due to lack this in the early stages of diabetes. There were discussions about the effectiveness of screening for type 2 diabetes, in Brazil for example, the last scan was considered unnecessary cost of almost 40 million. Simplest and most effective methods of screening are studied, such as the US and China that use some non-invasive methods to calculate the risk of diabetes. This study proposes a non-invasive screening method based on efficient coding technique to extract features of a Brazilian database (HIPERDIA) to form a new concise representation thereof, with the decrease of redundancy. The main hypothesis worked at this stage was the pursuit of independent components, which possibly it were present at the formation of the disease. Thus, the original data were decomposed by the independent component analysis method. In the classification stage to ensure discrimination between classes was used the method of support vector machines for one class. Tests were done to check the performance of the classifier after the feature extraction phase, and showed that it increases the performance of support vector machine to one class in making the discrimination between diabetics and non-diabetics. Results were reached (100%) with the combination of certain characteristics, and the method shows promise in obtaining a non invasive type 2 diabetes screening. Other tests were done to determine the influence of each non invasive marker in the final result and the generality of the method using other databases, as t of the Pima Indians and African Americans data sets. Then, reducing the number of features used to train the method and testing whether all possible combinations among the remaining, removing one by one, a total of 12,910 possibilities. It was observed the characteristics or markers that most affected the final outcome were age and characteristics related to body fat. Testing the generality of the method in other databases found that the method works best with balanced data set.

Keywords: Signal processing. Screening. Diabetes. Non invasive

Agradecimentos

A Deus por ter permitido tudo.

À UFMA e ao PPGEE.

Aos professores Allan Kardec Duailibe Barros Filho, Ewaldo Eder Santana e José Carlos Príncipe.

Ao meu marido Neilson e meu filho João Gabriel que são meus motivos de motivação.

Aos meus pais João e Júlia e a minha avó Laíla, pelos anos de amor e dedicação, por terem ensinado-me a sempre seguir em frente e em especial pelas fartas orações.

A minha mãe e a minha avó que sempre deram exemplos de esperança e amor.

Aos meus irmãos Paulo, Jr. e a minha irmã Amélia por serem meus primeiros exemplos e testes de como agir em sociedade.

À família Mendes (Segunda Família), em especial ao companheiro com insônia da madrugada, Raimundo Mendes e em *memoriam* à Antonieta Mendes, Socorro Mendes, José Mendes e Ivone Mendes que sempre receberam meu filho com cuidado e carinho, facilitando meu doutorado.

Aos amigos do departamento de Engenharia de Computação-UEMA.

Aos antigos e novos Pibianos.

Ao professor Edson Nascimento, pelo incentivo, motivação e direção na hora oportuna.

A amiga Rose pelo incentivo e apoio moral.

A equipe do HIPERDIA.

Aos amigos do CNEL e a família Simmitt pelo apoio e amizade na minha estadia na Universidade da Flórida.

Ao CNPQ pela bolsa concedida.

À CAPES pela bolsa de doutorado sanduíche concedida.

A todos meu eterno agradecimento, sem a participação destes, sem os comentários oportunos, discussões, seminários, aprendizados, sem a oportunidade de ver a ciência desenvolver-se dentro de um laboratório, e poder fazer um questionamento no meio deste conjunto universo de problemas em que vivemos, com a esperança no ser humano, não teria tido a oportunidade de chegar até aqui e é com grande estima que os cito e dedido a vocês os créditos deste trabalho!

"Mas os que esperam no Senhor renovam as suas forças, sobem com asas como águias, correm e não se cansam, caminham e não se fatigam."

Isaías 40.31

"Não existe nenhum caminho lógico para a descoberta das leis do Universo -o único caminho é o da intuição."

Albert Einstein

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 O diabetes e suas complicações

O diabetes mellitus(DM) é uma doença causada pela falência do pâncreas em produzir insulina, ou alternativamente, quando o corpo não pode usar este hormônio efetivamente. É uma doença incurável e seu tratamento é baseado em dietas, exercícios físicos e remédios. Os órgãos mundiais competentes publicam estatísticas alarmantes sobre o aumento do número de diabéticos no mundo, que cresce cada vez mais. Há 13 anos atrás a Organização mundial de saúde (OMS) estimava o número de diabéticos no mundo na casa dos 180 milhões de pessoas. Atualmente, a Federação Internacional de Diabetes (FID) estimou o número de diabéticos no mundo na casa dos 382 milhões de casos, sendo que deste número, 46% ainda desconhe sua condição. Recentemente, a OMS publicou um número mais próximo da FID: 347 milhões de pessoas no mundo com diabetes, 80% vivem em países de renda baixa ou média e o número de mortes devido a doença irá dobrar entre os anos de 2005 e 2030.

Existem vários tipos de manifestações do diabetes mellitus no ser humano. Por isso há diferentes classificações da doença, entre as mais comuns estão o diabetes tipo 1 (DM1), diabetes tipo 2 (DM2) e o diabetes gestacional(DG).

O DM1 surge quando o organismo deixa de produzir a insulina ou produz uma pequena quantidade deste hormônio, é mais comum em pessoas com menos de 35 anos, mas pode surgir em qualquer idade. O DM2 tem como peculiaridade a contínua produção de insulina pelo pâncreas, que por consequência, ocorre uma anomalia denominada “resistência insulínica". Esta impede que as células metabolizem glicose suficiente da corrente sanguínea, além disso o DM2 possui um fator hereditário maior do que no tipo 1 e uma grande relação com a obesidade e o sedentarismo, sua incidência é maior após os 40 anos. O diabetes gestacional é a

alteração das taxas de açúcar no sangue que aparece ou é detectada pela primeira vez na gravidez, pode persistir ou desaparecer após o parto.

O Diabetes Mellitus é considerado, hoje, como uma epidemia mundial, sendo um grande desafio para os sistemas de saúde de todo o mundo. Sua incidência e prevalência estão relacionadas ao envelhecimento populacional, a demora de avanços terapêuticos no tratamento da doença e ao estilo de vida atual, caracterizado pela falta de atividade física e aos hábitos alimentares que predispõem ao acúmulo de gordura corporal.

Estudos mostraram que metade dos indivíduos diagnosticados diabéticos desconhecia sua condição, e uma grande parcela já havia desenvolvido as complicações crônicas da doença que estão relacionadas ao tempo de exposição à hiperglicemia, como microangiopatia, retinopatia, nefropatia e neuropatias. Estas complicações são consideradas muito debilitantes ao paciente e onerosas ao sistema de saúde. Isto significa que os serviços de saúde tem diagnosticado casos de DM2 tardiamente, dificultando o sucesso do tratamento em termos de prevenção das complicações crônicas.[1]

Para que se entenda as complicações crônicas do DM2, explica-se estas em por menores: A doença cardiovascular é a primeira causa de mortalidade de indivíduos com DM2; a retinopatia representa à principal causa de cegueira adquirida e a nefropatia uma das maiores responsáveis pelo ingresso a programas de diálise e transplante; o pé diabético constitui-se em importante causa de amputações dos membros inferiores. Desta forma, procedimentos diagnósticos e terapêuticos como cateterismo, bypass, coronariano, fotocoagulação, retiniana, transplante renal, hospitalizações, absenteísmo, invalidez e morte prematura elevam substancialmente os custos diretos e indiretos da assistência à saúde da população diabética. Ainda, o DM2 é acompanhado de outras morbidades que podem tornar os custos totais exorbitantes[2]. No Brasil, como exemplo, só as doenças cardiovasculares são responsáveis por 1.150.000 das internações/ano, com um custo aproximado de 475 milhões de reais, sendo que nestes números não estão inclusos os gastos com procedimentos de alta complexidade.

Além do fator financeiro, o DM gera vários transtornos sociais considerados como custos indiretos para a sociedade e para o indivíduo doente, tais como: morte prematura, incapacidades, absenteísmo, a diminuição do retorno da educação oferecida ao indivíduo, diminuição da renda do chefe de uma família, aumento de aposentadorias precoces e desemprego. [3]

1.1.1 Os valores dos custos com o diabetes

A ADA (American Diabetes Association) publicou em 2007 e em 2012 os valores com os gastos do diabetes nos Estados Unidos. Em 2007, quando começou-se a orçar estes gastos, o custo foi estimado em U$ 174 bilhões de dólares, sendo U$ 116 bilhões em despesas médicas e U$ 58 bilhões com os custos indiretos (redução da força de trabalho nacional) [4], este número aumentou para U$ 245 bilhões em 2012, com U$ 176 bilhões com os custos médicos diretos e U$ 69 bilhões com os custos indiretos.[5]. Um crescimento de quase 40% nos custos totais com a doença em menos de 10 anos.

No Brasil, em 2007, foi feito um estudo sobre os custos do diabetes e hipertensão arterial em uma unidade de saúde pública de Recife/PE. Só os custos diretos com a assistência aos portadores das doenças, neste ano, totalizaram o valor de R$ 4.885.291,82 bilhões de reais, os valores repassados para reembolso pelo SUS(Sistema Único de Saúde) totalizaram um valor de R$2.118.893,56 bilhões de reais. Os medicamentos para as doenças cardiovasculares foram responsáveis por 24%, os grupos de maiores despesas foram com medicamentos: 36%, serviços de terceiros: 20,5%, e com pessoal especializado: 20,1%; Excluindo-se os valores dos medicamentos de R$ 3.092.867,40 bilhões, ainda resultou-se em um saldo negativo de R$ 973.973,84 reais para a unidade de referência.[6]

Os custos indiretos com a doença não foram orçados neste estudo. No Brasil, há uma carência de estudos proporcionais aos dos Estados Unidos para que se possa medir os custos reais do diabetes, bem como sua incidência de mortalidade e prevalência. Os recursos disponíveis no sistema de saúde são limitados, desta forma,

torna-se essencial a correta administração destes para um melhor aproveitamento e para que possam ser planejadas ações de combate a doença.[7]

1.2 Proposta de Tese

Nos parágrafos anteriores fez-se uma introdução da problemática do crescimento da diabetes, das suas complicações e desafios a serem vencidos. Como visto pelas estatísticas, o número de pacientes diabéticos e os custos com a doença crescem rapidamente, com a demora no diagnóstico desta, mais problemas aparecem, devido ao surgimento de suas complicações.

Na mesma proporção que o número de casos da doença aumenta, os dados sobre ela aumentam em bases de dados no mundo inteiro. Esses dados são utilizados para levantar, estimar ou propor metodologias de combate a diabetes. Pode-se citar estes dados como: Pressão, IMC, peso, glicemia em jejum, altura e etc.

Nos pacientes diabéticos estes dados são considerados marcadores da doença, marcam ou descrevem o padrão da diabetes em determinado individuo. São como sinais que o corpo emite e ao conhecer estes padrões, pode-se identificar a que rótulo ou classe pertencem. Classificam-se estes marcadores como invasivos ou não invasivos. Os invasivos são aqueles que podem ser estimados ou encontrados por meio de um exame de sangue(glicemia em jejum, H1ac, insulina). Já os não invasivos, podem ser detectados por meio não invasivo (pressão, peso, altura, presença de certas características e etc).

O objetivo deste trabalho é propor um método não invasivo de rastreamento para a diabetes tipo 2, por meio de marcadores não invasivos da doença com confiabilidade e baixo custo. A identificação precoce dos casos de diabetes e o estabelecimento entre os portadores da doença e as unidades de saúde são elementos imprescindíveis para o sucesso do controle dos agravos da doença. O acompanhamento e o controle do diabetes mellitus no âmbito da atenção básica pode evitar o surgimento e a progressão de complicações, reduzindo o número de internações hospitalares, bem como a mortalidade devido à esses agravos.

Na metodologia proposta foram utilizados dois métodos computacionais: Um para extração das características, no qual foi utilizado o método de codificação eficiente através da análise das componentes independentes para obter-se uma representação mínima e sem redundâncias dos dados, e outro para a classificação dos pacientes em diabéticos e não-diabéticos, que foi o classificador de máquinas de vetor de suporte para uma classe.

1.3 Organização do texto

Este trabalho está organizado da seguinte forma: No capítulo 2 descreve-se a base de dados utilizada e apresenta-se a revisão teórica dos métodos implementados. Primeiro foi feita uma discussão sobre esta base de dados e o estado da arte em rastreamento de diabetes, com as calculadoras de diabetes já validadas por organizações responsáveis como a Sociedade de diabetes americana (ADA) do inglês *American diabetes association*, bem como suas vantagens nos dias atuais, frente ao crescimento alarmante da diabetes tipo 2. Segundo, fez-se a descrição do referencial teórico sobre o método proposto, da extração de características através da análise das componentes independentes até ao método de classificação.

O passo de extração foi crucial para a fase de classificação, pois buscou-se uma representação concisa e fiel dos dados originais pelo conceito de indepêndencia estatística que minimizou o trabalho de classificação, por último descreve-se o classificador utilizado que foi o One-class SVM (Máquinas de vetor de suporte para uma classe).

No capítulo 3, a metodologia utilizada e os resultados obtidos são discutidos, ilustrados por meio de gráficos de disperção em 3D e por uma medida que mostra claramente a separação das classes. Por fim no capítulo 4, descreve-se os resultados apresentados, a possível relação do método com a fisiologia e trabalhos futuros são discutidos.

2 Revisão teórica

2.1 Descrição das bases de dados

Em muitos trabalhos da literatura utiliza-se a base de dados dos índios Pima para testes de sistemas CAD (*Computer Aided Design*) [8], isto porque é uma base de livre acesso e uma das poucas disponíveis em diabetes. Esta base tem alguns pontos de deficiência como ser desbalanceada e homogênea.

O desbalanceamento em domínio da classificação de padrões, significa que existem muito menos casos de algumas classes do que outras e isso pode causar uma valorização das classes predominantes e a ignorar classes de menor representação[9]. Quanto a homogeneidade da base, significa que ela pode ser representativa apenas para populações homogêneas (Índios Pima). Para populações heterogêneas como a brasileira, que é altamente miscigenada, os testes com a base Pima em sistemas CAD para auxílio a classificação da diabetes não são representativos.

Nossa motivação é implementar um método que possa fazer o auxílio e a detecção de pessoas com diabetes aproveitando informações simples, disponíveis em bases de dados. Promovendo uma forma simples e barata de auxilio ao rastreamento de diabetes tipo 2 na população brasileira que é heterogênea.

Desta forma, procurou-se uma base de dados que pudesse representar a população brasileira e em cooperação, o DATASUS disponibilizou a base de dados do sistema HIPERDIA (Hipertensão e diabetes)[10], que é um programa do governo federal brasileiro responsável pelo cadastramento de pacientes diabéticos e hipertensos atendidos pelo SUS (Sistema único de saúde). Esse programa possibilita a gestão dos cuidados necessários aos pacientes cadastrados no momento da vinculação destes com as unidades de atenção básica de saúde.

A base disponibilizada pelo HIPERDIA tem 17 características clínicas

de 1000 pacientes aproximadamente, por estado brasileiro. Totalizando um número aproximado de 27.000 pacientes, as características clínicas são:

1- Pressão arterial sistólica

2- Pressão arterial diastólica

3- cintura (cm)

4- Peso (kg)

5- Altura (cm)

6- Glicemia em jejum

7- Antecedentes familiares

8- Idade

9- Tabagismo

10- Sedentarismo

11- Sobrepeso

12- Infarto

13- AVC

14- Amputação

15- Pé diabético

16- Outras coronariopatias

17- Doença renal.

Primeiramente, utilizou-se apenas a base de dados do estado do Maranhão, com 501 pacientes diabéticos e 501 não-diabéticos.

Segundo o Hiperdia o cadastro destes pacientes portadores de diabetes e hipertensão, possibilita também o monitoramento de forma contínua da qualidade clínica (outcome), controle dos agravos e dos fatores de risco destas doenças na população assistida.

O programa tem a missão de fornecer informações gerenciais que permitam subsidiar os gestores públicos para tomada de decisões na adoção de estratégias gerais ou pontuais em relação a doença, como estimar o acesso aos serviços de saúde, planejar demanda para referenciamentos, estimar o uso de materiais, recursos humanos e capacitações, fornecer informações que subsidiem a gerência e gestão da assistência farmacêutica, instrumentalizar a vigilância à saúde fornecendo informações que permitam conhecer o perfil epidemiológico do diabetes mellitus, seus fatores de risco e suas complicações na população. Desta forma, possibilitar o controle social por meio de informações que permitam analisar acesso, cobertura e qualidade de atenção na população assistida pelo programa. [11]

Todas estas medidas são aprazíveis ao combate, tratamento, e controle da doença em seu portador no âmbito das políticas públicas, mas há outro objetivo desejável neste combate que é o diagnóstico da doença em sua fase assintomática, antes que esta se desdobre nas complicações do diabetes, como: doenças cardiovasculares, cegueira, amputações e outros. Por fim, aumentando os custos diretos e indiretos com a doença.

2.2 Rastreamento de diabetes

Há uma grande diferença entre o diagnóstico e o rastreamento de uma doença. No caso do diagnóstico, dado um indivíduo que apresenta sintomas ou sinais de uma possível doença, os testes de corte são realizados, estes testes não representam um rastreamento, mas um diagnóstico isolado.

Já no rastreamento, o objetivo é identificar vários indivíduos assintomáticos que estão propensos a ter diabetes antes do diagnóstico final isolado. Dessa forma, diminuindo as chances do surgimento de complicações desta. Isto

é viável para a diabetes do tipo 2, pois a doença tem uma fase assintomática, que compreende um período de 7 a 10 anos. Assim, prevenindo as possíveis complicações da doença em questão. [12]

No mundo todo existam divergências quanto a aplicação e efetividade do rastreamento para o diabetes tipo 2, segundo[13] existem sete condições que reunidas indicam a possibilidade de um rastreamento em determinada doença, que são:

1. A doença representa um importante problema de saúde que impõe uma carga significativa sobre a população?
2. A história natural da doença é compreendida?
3. Há uma fase assintomática em que a doença pode ser diagnosticada?
4. Existem testes disponíveis que possam detectar a doença no seu estágio pré-clínico? Estes testes são aceitáveis, confiáveis?
5. O tratamento após a detecção precoce produz benefícios superiores aos obtidos quando o tratamento é tardio?
6. Os custos para cada caso detectado e tratamento são razoáveis e equilibrados em relação aos gastos com a saúde como um todo? Existem hospitais e recursos disponíveis para novos casos diagnosticados?
7. O rastreamento será um processo contínuo, sistemático e não apenas um esforço em um tempo isolado?

Até 2008 estudos diziam que o diabetes não cumpria todas estas condições para a realização do rastreamento [14]. Faltavam estudos clínicos aleatorizados comprovando os riscos e benefícios do rastreio para a diabetes tipo 2, como redução na incidência da doença e diminuição de custos com as complicações da doença.

Nos últimos anos, estudos clínicos aleatorizados foram feitos e garantiram os benefícios do rastreamento da diabetes tipo 2 [15],[16]. Os benefícios em linhas gerais estão na redução dos custos em pacientes com fatores de alto risco que

passaram pelo rastreamento em relação aos custos dos pacientes do mesmo grupo que não foram rastreados.

O estudo comparou a economia feita com a realização do rastreamento com cinco tipos diferentes de exames invasivos: [**1**]Glicemia plasmática de 1h depois do consumo de 50 g de glicose oral(GCTpl), [**2**]Glicemia capilar de 1 h depois da ingestão de 50 g de glicose oral(GCTcap), [**3**]Medição de glicemia plasmática aleatória (RPG), [**4**]Medição de glicemia capilar aleatória (RPG,) e [**5**]Hemoglobina Glicada (A1c). Os resultados mostram vantagens econômicas e sociais no diagnóstico antecipado do diabetes tipo 2 através do rastreamento do que no diagnóstico isolado caso a caso (sem rastreamento).

A média de economia financeira obtida com o rastreamento para a diabetes foi de -25,72%, sendo que o exame mais econômico utilizado para o rastreamento foi o de GCTpl, com -30.12 % e o menos econômico o de A1c, com uma diminuição de -14,88%. Todas estas diminuições foram em comparação ao valor gasto com o não rastreamento. Quanto aos custos sociais, a economia média do rastreamento em relação ao não rastreamento foi de -21,08%. Neste caso o mais econômico foi o de GCTpl, com -24,91 % e o menos econômico o de A1c, com -13.09 %.

Também foram orçadas as porcentagens econômicas em relação ao diagnóstico antecipado de pré-diabéticos oriundos do rastreamento. Neste caso, obtiveram-se números significantes de economia financeira e social com o rastreamento. Nos custos financeiros, a economia média em relação ao não rastreamento foi de -9,31%. O exame de rastreamento mais econômico foi o de GCTpl, com uma econômia de -11,01% e o menos econômico o de A1c, com uma econômia de -5.13%. Com relação a econômia gerada nos custos sociais para os pré-diabéticos, com o rastreamento, o percentual médio de diminuição dos custos foi de - 1,36%, com a maior econômia de 2,82% no exame de GCTpl e a menor econômia de 0.53% no exame de A1c.

No mesmo estudo foi analisada a economia em 3 anos com o tratamento, acompanhamento e impacto dos casos diagnosticados como diabéticos, pré diabéticos

e não diabéticos oriundos do rastreamento para os principais fatores observados pelo sistema de saúde nos EUA. Os fatores observados foram:

1a - Índice de Massa Corporal menor que $25kg/m^2$ ($IMC < 25kg/m^2$)

1b - Índice de Massa Corporal entre $25kg/m^2$-$35kg/m^2$ ($IMC25 - 35kg/m^2$)

1c - Índice de Massa Corporal maior que $35kg/m^2$ ($IMC > 35kg/m^2$)

2a - Idade menor que 40 anos (Idade $<$ 40 anos)

2b - Idade entre 40 e 55 anos (Idade 40-55 anos)

2c - Idade maior que 55 anos (Idade $>$ 55 anos)

3a - Pressão menor que 130 mmKg ($Pressão < 130mmKg$)

3b - Pressão maior ou igual a 130 mmKg ($Pressão \geq 130mmKg$)

4a - Triglicéridos baixo risco (TG baixo risco)

4b - Triglicéridos alto risco (TG alto risco)

5a - HDL baixo risco (HDL baixo risco)

5b - HDL alto risco (HDL alto risco)

6a - Cintura baixo risco (Cintura baixo risco)

6b - Cintura alto risco (Cintura alto risco)

7a - Histórico familiar negativo (Histórico familiar negativo)

7b - Histórico familiar positivo (Histórico familiar positivo)

Para todos os fatores observados, o rastreamento foi mais econômico que o não rastreamento tanto no campo financeiro como no social, mas as econômias mais significativas foram observadas nos fatores de alto risco como o 1b, 2b,3b, 4b, 5b, 6b e 7b com o tratamento, acompanhamento e impacto para os diabéticos, pré diabéticos e dos não diabéticos oriundos do rastreamento.

O estudo mostra a economia do rastreamento nos custos sociais(Somam-se diminuição de horas trabalhadas, custos da família com medicamentos e etc)e nos custos financeiros (Somam-se gastos hospitalares com exames, medicamentos, campanhas e acompanhamento dos diabéticos e pré-diabéticos)tanto no diagnóstico antecipado oriundo do rastreamento, como no acompanhamento clínico dos pacientes oriundos do rastreamento.

Pode-se ver estes resultados de custos, em dólar americano, para os diabéticos e pré diabéticos no **Anexo 1**: Nas **tabelas 6.1 e 6.2** para os custos financeiros e sociais dos diabéticos, respectivamente. E nas **tabelas 6.3 e 6.4** para os custos financeiros e sociais com os pré diabeticos, respectivamente.

No Brasil, o último rastreamento realizado foi em 2001. Utilizou-se o teste de glicemia capilar para o rastreamento com um custo total para o governo federal de R$38.620,775,00. Neste custo não estão orçados os custos locais dos municípios, nem os custos com os tratamentos e acompanhamento de cada novo caso diagnosticado.[17] Foi considerado desnecessário na época, pelos altos custos envolvidos e por não atender os pré-requisitos citados acima, que envolvem o rastreamento de uma doença.

2.3 Calculadoras de risco de diabetes

Em 2008, a ADA adicionou como medida preventiva e de cálculo do risco do diabetes a calculadora de risco. O usuário pode prever o risco da doença por meio de um sistema que tem como entrada dados clínicos do usuário como altura, peso, circunferência da cintura, idade, sexo, etnia, utilização de medicação para pressão alta, utilização de medicação para colesterol alto, diabetes gestacional, pressão alta, hitórico de diabetes na família e exercícios físicos.

O sistema utiliza estas características clínicas como entrada para fazer o cálculo de risco para pré-diabetes e diabetes não diagnosticada. Utiliza o método de árvore de regressão, com uma sensibilidade de 75.36% e com uma especificidade de 64.69% para a detecção de diabetes. [18]

Outra calculadora de diabetes empregada na Ásia é validada para o povo chinês e utiliza o método de regressão logística com os seguinte resultados para sensibilidade 83.3% e para especificidade de 66.5%. [19]

Todos estes métodos não utilizam uma fase de minimização de redundâncias ou de mínima representação, o que chamamos em aprendizado de máquina de extração de características. Acredita-se que é um passo de extrema importância na classificação, pois minimiza o trabalho do classificador.

2.4 Extração de características

O problema de dados correlacionados em aprendizado de máquina é que decrescem a performance do classificador, isto por causa da informação redundante presente nos dados. Os dados reais são geralmente muito correlacionados, e métodos para executar a extração de características antes da etapa de classificação tornam-se indispensáveis. Isto pode aumentar o poder do classificador, porque provê uma representação concisa, invariante e fiel dos padrões de entrada.

O trabalho da etapa de extração de características, é mapear o conjunto de dados do vetor observação para o espaço de características [20]. Aqui, para alcançar este objetivo, utilizou-se a codificação eficiente modelada por ICA do inglês *Independent Component Analysis*.

2.5 Extração de características através da Análise dos Componentes independentes

A extração de características que usa estatística, tem sido influenciada pelo modelo de processamento da informação neural[21]. Estudos de neurociência sugerem que o processo de processamento neural estimula a informação de acordo com o conceito de “Codificação Eficiente"[22] [23][24] [25]. Debaixo deste conceito, as respostas neurais são mutuamente estatisticamente independentes, o que significa

que não há “informação redundante"processada.

O objetivo computacional da codificação eficiente, é de estimar a partir das estatísticas do conjunto de padrões, um código compacto que reduza a redundância nos dados com uma perda mínima da informação. Os dados são transformados por um conjunto de filtros lineares $\mathbf{W}$ em uma saída $\mathbf{s}$, na forma matricial:

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{2.1}$$

equivalentemente, em termos da matriz das bases $\mathbf{x} = \mathbf{W}^{-1}\mathbf{s} = \mathbf{A}\mathbf{s}$ onde $\mathbf{s}$ é uma estimação das componentes independentes, métodos para derivar o código eficiente no modelo da equação 2.1 podem ser obtidos pela análise de componentes independentes que é um modelo de distribuição estatística de um conjunto de padrões, ela é utilizada para encontrar as características (ou funções bases).

2.5.1 Análise de componentes independentes

A análise de componentes independentes, do inglês independent component analysis (ICA), é uma técnica baseada no modelo de independência, que é um requisito do código eficiente [26]. Esta técnica foi desenvolvida para atender ao requisito de trabalhar com componentes independentes e não-gaussianas.

Foi difundido e popularizado pela resolução do problema de separação de fontes cegas (BSS, blind source separation), onde o problema está na estimação da saída de uma fonte conhecida, sendo que esta fonte recebe vários sinais misturados e desconhecidos.

Esta técnica tem sido aplicada em áreas como áudio, radar, comunicação móvel, engenharia biomédica e outras. Como a técnica é baseada no modelo de independência, as fontes devem ser mutuamente estatisticamente independentes, definiremos independência estatística para entender o modelo de ICA.

2.5.2 O que é independência estatística?

Duas variáveis aleatórias s_1 e s_2 são estatisticamente independentes se a partir de s_1 não for possível estimar ou inferir algum valor ou informação de s_2. Exemplos simples e rotineiramente utilizados de variáveis aleatórias independentes, são os sinais de eletrocardiograma de um feto e ruídos em sistemas de comunicação[27].

Matematicamente, a independência estatística de s_1 e s_2 ocorre satisfazendo-se a seguinte condição :

$$P_{s_1,s_2}(s_1, s_2) = P_{s_1}(s_1)P_{s_2}(s_2) \tag{2.2}$$

Assim, a probabilidade conjunta de duas variáveis estatísticamente independentes pode ser calculada apenas como o produto das marginais.

2.5.3 Modelo do método de ICA

O modelo de ICA é similar ao modelo de codificação eficiente de acordo com a equação 2.1. De uma forma geral, cosideremos que $\mathbf{x} = [x_1; x_2; \ldots; x_n]^T$ e $\mathbf{s} = [s_1; s_2; \ldots; s_n]^T$ são vetores aleatórios, sendo que cada elemento x_i é uma mistura dos elementos de $\mathbf{s}$.

As componentes de $\mathbf{s}$ não são observadas diretamente, pois $\mathbf{s}$ além de independentes, são latentes .

Desta forma, para simplificar, utiliza-se um modelo matemático da forma:

$$\mathbf{x} = \mathbf{As} \tag{2.3}$$

em que $\mathbf{A}$ é uma matrix de mistura e corresponde a equação 2.1, quando $\mathbf{x} = \mathbf{W}^{-1}\mathbf{s} = \mathbf{As}$.

O objetivo da técnica, basicamente, é recuperar $\mathbf{s}$, por meio de $\mathbf{x}$, sem

nenhuma informação sobre as características de **A**, a equação 2.3 é um modelo estatístico chamado de análise das componentes independentes, que descreve os dados observados pelo processo de mistura das componentes independentes **s**, sendo que estas não podem ser observadas diretamente. Assim é preciso estimar tanto **s** quanto a matriz de mistura **A**, porque somente **x** é observável.

2.5.4 Particularidades em ICA

A técnica de ICA tem suas particularidades, dentre as quais pode-se citar 3 delas. Como dito acima, ICA foi criada para atender as variáveis aleatórias não-gaussianas, pois para variáveis aleatórias gaussianas as componentes são sempre descorrelacionadas e, consequentemente, independentes. Além da distribuição conjunta de misturas destas variáveis serem também gaussianas. Assim, haveria informação perdida na aplicação da técnica pois esta distribuição é rotacionalmente simétrica e a informação da rotação da mistura é perdida.

A segunda particularidade é que a informação de variância das componentes independentes é perdida no processo de estimação; e a terceira é que não se pode estabelecer uma ordem para as componentes independentes.

2.5.5 CI por maximização da não gaussianidade

Um dos problemas chaves de ICA é estimar as componentes independentes **s** a partir de **x** e a não-gaussianidade é um elemento chave para esta estimação, pois a matriz **A** segundo a equação 2.3 não é identificável quando mais de uma das componentes independentes tem distribuições gaussianas.

O teorema do limite central, implica que a distribuição de soma das componentes independentes é sempre mais próxima de uma gaussiana do que qualquer uma das distribuições das componentes isoladas. Dessa forma, para estimar uma componente independente s_i, considera-se a seguinte soma ou combinação linear dos vetores **x**

$$y = \mathbf{b}^T \mathbf{x} \tag{2.4}$$

em que $\mathbf{b}$ é um vetor a ser determinado. Se $\mathbf{b}$ for uma das linhas da inversa de $\mathbf{A}$, y será igual a uma das componentes independentes s_i.

Do modelo de ICA, equação 2.3, tem-se que

$$y = \mathbf{b}^T \mathbf{x} \tag{2.5}$$

$$y = \sum_i b_i x_i \tag{2.6}$$

$$y = \mathbf{b}^T \mathbf{A} \mathbf{s} \tag{2.7}$$

Denota-se o produto $\mathbf{b}^T \mathbf{A}$ como $\mathbf{q}$. Assim tem-se que:

$$y = \mathbf{b}^T \mathbf{x} = \mathbf{q}^T \mathbf{s} = \sum_i q_i s_i \tag{2.8}$$

Assim, da equação 2.7, observa-se que y também é uma soma das componentes independentes s_i; Desta forma, pelo teorema do limite central é possível concluir que a distribuição de y tende a ser gaussiana, diferente das outras distribuições individuais das outras componentes independentes s_i. Assim, um dos elementos q_i é diferente de zero (0).

Como na prática os valores q_i são desconhecidos, pode-se através das equações 2.4 e 2.8 dizer que

$$\mathbf{b}^T \mathbf{x} = \mathbf{q}^T \mathbf{s} \tag{2.9}$$

Assim, pode-se variar $\mathbf{b}$ e observar a distribuição de $\mathbf{b}^T \mathbf{x}$. Desta forma, pode-se tomar $\mathbf{b}$, como um vetor que maximiza a não-gaussianidade de $\mathbf{b}^T \mathbf{x}$, sendo que este vetor corresponde a $\mathbf{q} = \mathbf{A}^T \mathbf{s}$, sendo que este vetor possui apenas uma de suas componentes diferente de zero. Daí, pode-se concluir que y da equação

2.4 é igual a uma das componentes independentes. Logo, a maximização da não-gaussianidade de $\mathbf{b}^T\mathbf{x}$ permite encontrar uma das componentes.

2.5.6 Medindo a não gaussianidade

A entropia de uma variável aleatória está relacionada com a quantidade de informação que essa variável possui. Sendo y um vetor aleatório com função densidade de probabilidade $f(y)$, a sua entropia diferencial é dada por:

$$H(y) = -\int f(y) \log f(y) dy \tag{2.10}$$

Sabendo-se que uma variável gaussiana tem a maior entropia dentre todas as variáveis aleatórias de igual variância [28][27], tem-se que uma versão modificada da entropia diferencial pode ser usada como medida de não-gaussianidade. Tal medida é denominada negentropia, definida por:

$$HJ(y) = H(y_{gauss}) - H(y) \tag{2.11}$$

sendo y_{gauss} uma variável aleatória de mesma matriz de covariância que y. A negentropia é sempre não-negativa, e pode assumir zero se, e somente se, y tem distribuição gaussiana e é invariante para transformações lineares inversíveis.

Apesar de permitir que se possa medir não-gaussianidade, a negentropia é de difícil estimação, sendo necessária sua estimação por aproximações através de momentos de alta ordem. Assim,

$$J(y) \approx \frac{1}{12} E\{y^3\}^2 + \frac{1}{48} kurt(y)^2 \tag{2.12}$$

Sendo $kurt(y)$ a *kurtosis* de y, definida como o momento de quarta ordem da variável aleatória y, definida por :

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2 \tag{2.13}$$

2.6 Máquinas de vetores de suporte

2.6.1 Introdução

Os fundamentos teóricos sobre a Máquina de vetores de suporte, do inglês Support vector machine (SVM), foram introduzidos por V. Vapnick em 1995 e consistem em um método de classificação para duas classes [29]. A idéia básica deste método é construir um hiperplano com uma superfície de decisão em que a margem de separação entre as duas classes é maximizada.

O termo Máquina de vetores de suporte surgiu porque que os pontos do conjunto de treinamento que estão mais próximos da superfície de decisão são chamados de vetores de suporte. O SVM realiza essa tarefa baseado no princípio de Minimização do risco estrutural, que é baseado no fato da taxa de erro da máquina de aprendizado no conjunto de teste ser limitada pelo somatório taxa de erro de treinamento e por um termo que depende da dimensão de Vapnik-Chervonenkis (VC).[30]

A classificação de dados estatísticos pode ser atribuída a um problema de uma, duas ou várias classes. Na classificação binária, os dados de duas classes estão disponíveis, supondo que as amostras da base de dados contenham classes igualmente balanceadas. Uma base de dados desbalanceada, poderia levar a resultados insatisfatórios [31]. Um problema comum com esta abordagem é a decisão da criação do limite entre as duas classes, na pior das situações, poderia ter uma grande taxa de erro se as classes não forem bem separadas.

As Máquinas de Vetor de Suporte para uma classe constroem um classificador somente para o conjunto de exemplos positivos, chamados de amostras de treinamento positivas [32]. A classificação é feita basicamente pela geração de uma hiper-esfera para a decisão, limitando apenas uma classe de outras que possam

existir. A estratégia é mapear os dados em função do espaço de características e em seguida tentar usar a hiper-esfera para descrever os dados e para a inserção de dados. Por isso, a metodologia necessita aprender apenas sobre uma classe e assim as bases de dados desbalançeadas podem ser utilizadas nesta abordagem, sem problemas com o desempenho do classificador. Esta é uma vantagem deste em relação ao SVM para duas classes.

2.6.2 Máquinas de vetor de suporte para uma classe

Supondo que uma base de dados tenha uma distribuição probabilidade P no espaço de características. Encontrar um subconjunto S deste espaço de características, tal que a probabilidade que um ponto de P esteja fora de S é determinada por uma condição a priori especificada

$$\upsilon \in (0, 1) \tag{2.14}$$

A solução para este problema é obtida pela estimação da função f, que é positiva em S e negativa no complemento $\overline{S}$. Em outras palavras, Scholkopf desenvolveu um algoritmo que retorna uma função f [31]. Esta função toma valores $+1$ em uma pequena região, a hiper-esfera, capturando o maior número de dados e toma valores -1 em outro local.

$$f(x) = \begin{cases} +1 \ \ se \ \ x \in S \\ -1 \ \ se \ \ x \in \overline{S} \end{cases} \tag{2.15}$$

O algoritmo pode ser resumido como um mapeamento dos dados em um espaço de características H, usando uma função kernel apropriada, e então tentar separar os dados mapeados da origem com uma margem máxima.

No nosso contexto, tem-se amostras de treino $x_1, x_2, ..., x_l$ pertencentes a uma classe X, onde X é um pequeno subconjunto de R^N. Tem-se $\phi : X \rightarrow H$ sendo

o kernel que transforma as amostras de treinamento para outro espaço. Então, para separar o conjunto de dados da origem tem-se um seguinte função objetivo na forma primária

$$\min \mathbf{r}^2 + \frac{1}{vl} \sum\nolimits_i \zeta_i$$

sujeito a

$$\|\Phi(X_i) - c\|^2 \leq \mathbf{r}^2 + \zeta_i\,, \quad \zeta_i \geq 0 \; para \; i \in [l]$$

Sendo que $v \in [0, 1]$ representa a quantidade total das amostras de treinamento,$\mathbf{r}$ é um vetor ortogonal que separa as amostras de treinamento da origem até um limiar ρ, l representa a parte dos dados de treinamento rejeitados pela hiper-esfera, ζ é usado para rejeitar as amostras de treinamento da hiper-esfera.

Este problema de otimização é resolvido com os multiplicadores de Lagrange:

$$\begin{aligned} L(\mathbf{r},\, \zeta,\, c,\, \alpha,\, \beta) &= \mathbf{r}^2 + \sum_{i=1}^{l} \alpha_i \left[\|\Phi(X_i) - c\|^2 - \mathbf{r}^2 - \zeta_i\right] \\ &+ \frac{1}{vl} \sum_{i=1}^{l} \zeta_i - \sum_{i=1}^{l} \beta_i \zeta_i \end{aligned}$$

$$\frac{\partial L}{\partial \mathbf{r}} = 2\mathbf{r}\left(1 - \sum \alpha_i\right) = 0 \Rightarrow \sum \alpha_i = 1 \tag{2.16}$$

$$\frac{\partial L}{\partial \zeta_i} = \frac{1}{vl} - \alpha_i - \beta_i = 0 \Rightarrow 0 \leq \alpha_i \leq \frac{1}{vl} \tag{2.17}$$

$$\begin{aligned} \frac{\partial L}{\partial c} &= -\sum 2\alpha_i \left(\Phi(X_i) - c\right) = 0 \\ &\Rightarrow c = \sum \alpha_i \Phi(X_i) \end{aligned} \tag{2.18}$$

A equação 2.16 e 2.17 coloca para fora da hiper-esfera as amostras de treinamento rejeitadas, enquanto que a equação 2.18 informa o c (centro da hiper-esfera) que pode ser expresso como a combinação linear $\Phi(X)$, o que é possível resolver pela forma dual com a função kernel

$$\min \sum\nolimits_{i,j} \alpha_i \alpha_j k\left(X_i,\, X_j\right) - \sum\nolimits_i \alpha_i k\left(X_i,\, X_i\right)$$

sujeito a

$$0 \leq \alpha_i \leq \frac{1}{vl}, \quad \sum\nolimits_i \alpha_i = 1$$

Uma importante família de funções núcleo (kernel) é a função de base radial (RBF, *Radial Basis Function*), muito comumente utilizada em problemas de reconhecimento de padrões e também utilizada neste trabalho, que é definida por

$$K(\mathbf{x}, \mathbf{y}) = e^{-y\|x-y\|^2} \tag{2.19}$$

em que $\gamma > 1$ é um parâmetro que é definido pelo usuário.

3 Materiais e resultados

3.1 Metodologia

O método proposto tem o objetivo de fazer o rastreamento de pacientes diabéticos em sua fase assintomática de uma forma simples e eficaz. Assim, fazer a discriminação em duas classes ou categorias:

- Diabéticos
- Não diabéticos

Para isto, utilizou-se a metodologia que foi proposta ao longo do trabalho, resumida na Figura 3.1. Nos testes, foi utilizada a base de dados brasileira do sistema HIPERDIA. Primeiramente, foram feitos testes com a base de dados no estado do Maranhão que foi descrita no capítulo 2. Seguindo a linha de classificação com testes não invasivos, no teste principal as variáveis que representam o exame invasivo foram retiradas, estas são representadas pelas características ou marcadores: glicemia em jejum e doença renal que correspondem as numerações 6 e 17. A característica pé diabético que correspondente a numeração 15 também foi retirada dos testes por já ser uma complicação da doença.

Os pacientes que tinham dados sem preenchimento ou fora do padrão foram retirados na fase de pré-tratamento da base. A tabela 3.1 mostra as características utilizadas e as não utilizadas no teste não invasivo.

3.2 Aprendendo um subespaço através de ICA

Tomamos um vetor $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_n\}$ como sendo um conjunto de observações tomadas de um mesma base de dados ou em forma matricial $\mathbf{X}$.

Tabela 3.1: Características clínicas utilizadas no teste não invasivo

Características	Teste não invasivo
1. **Idade**	x
2. **Pressão Arterial Sistólica**	x
3. **Pressão Arterial Diastólica**	x
4. **Cintura(cm)**	x
5. **Peso(kg)**	x
6. **Glicemia**	
7. **Altura(cm)**	x
8. **Antecedentes familiares**	x
9. **Tabagismo**	x
10. **Sedentarismo**	x
11. **Sobrepeso**	x
12. **Infarto**	x
13. **Coronariopatias**	x
14. **AVC**	x
15. **Pé diabético**	
16. **Amputação**	x
17. **Doença renal**	

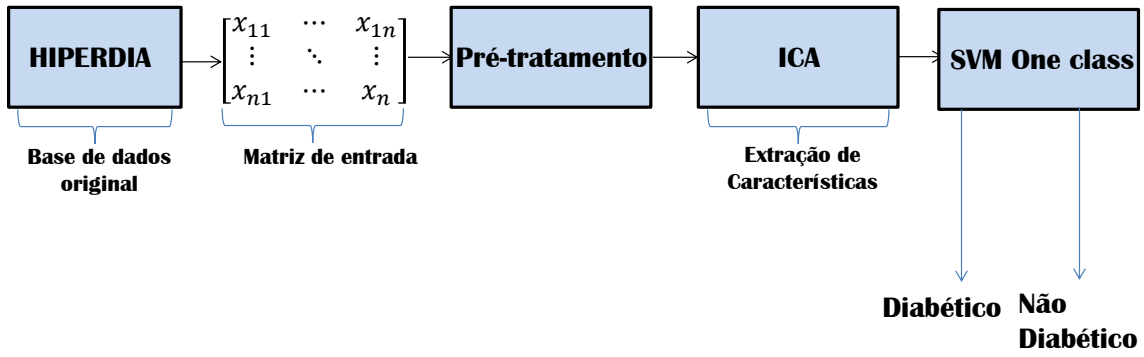

Figura 3.1: Fluxograma da metodologia proposta. A base HIPERDIA é tomada em uma matriz **X** que é tratada de forma a retirar itens faltantes ou fora do padrão, após isso é feita a extração de características através de ICA, acham-se as componentes independentes que são a entrada para o classificador One Class SVM, que faz a discriminação entre diabéticos e não diabéticos

Usando **x** como o conjunto de treinamento, ICA aprende as funções bases em colunas da matriz **A** para a classe de dados, de modo que o conjunto de variáveis que compõem o vetor **s** são mutuamente estatisticamente independentes, como na aquação 2.3.[33][26]

Para atingir a independência estatística, o algoritmo de ICA trabalha com estatística de alta ordem, que aponta a direção onde o dado é maximamente independente. Aqui, utilizou-se o algoritmo FastICA [27].

Por uma “unidade", referimo-nos a uma unidade computacional, eventualmente um neurônio artificial, tendo um vetor peso **w** que a referida unidade é apta para atualizar por uma regra de aprendizado. A regra de aprendizado do

FastICA procura uma direção, por exemplo, um vetor unitário $\mathbf{w}$, tal que a projeção $\mathbf{w^T} \cdot \mathbf{x}$ maximize a não-gaussianidade. A Não-Gaussianidade é aqui medida por aproximação de negentropia $J(\mathbf{w^T} \cdot \mathbf{x})$."[33]

$$J(y) \propto [E\{G(y)\} - E\{G(v)\}]^2 \tag{3.1}$$

Para muitos neurônios, necessita-se rodar o algoritmo FastICA de uma unidade repetidamente para estimar várias componentes independentes.

3.2.1 Análise de componentes independentes para a extração de características

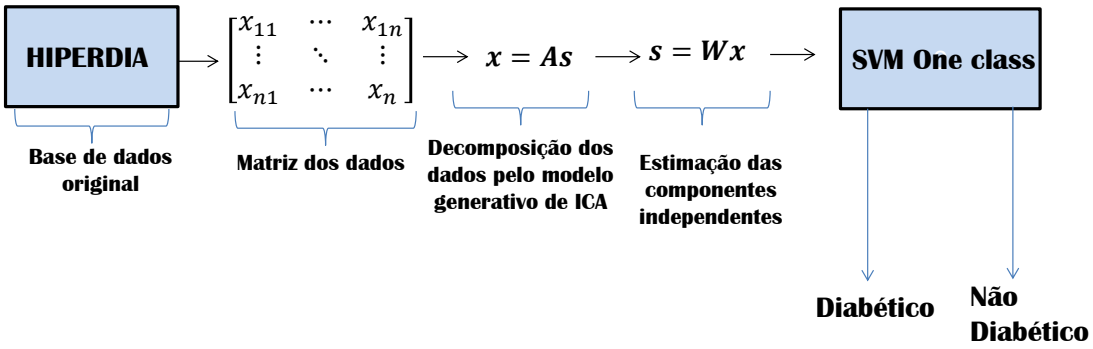

Figura 3.2: Fluxo do processo de decomposição dos dados por ICA e a estimação das componentes independentes para a entrada do classificador

Uma amostra que é representada pelo vetor observação de diabéticos ou não diabéticos, $\mathbf{x}_{treino}$, é transformada usando as funções bases ou as características de ICA , $\mathbf{W}$, que foram tomadas a partir do método de análise de componentes independentes descrito na sessão anterior. O novo subespaço,$\hat{\mathbf{x}}_{\mathbf{treino}}$ é usado como entrada do classificador. Isto é obtido através de uma simples projeção que consiste do produto interno entre o vetor observação e as funções bases de ICA. Podemos ver esta relação na seguinte equação, semelhante a equação 2.1

$$\hat{\mathbf{x}}_{\mathbf{treino}} = \mathbf{W}\mathbf{x}_{\mathbf{treino}} \tag{3.2}$$

onde $\mathbf{W} = \mathbf{A}^{-1}$ são as funções bases ou as características, veja a **Figura 3.2.1**

Como foi discutido nas sessões anteriores, ICA está relacionado com o conceito de "codificação eficiente ´´ que reduz a redundância nos padrões com o mínimo de perda da informação, isto garante que o objetivo no estágio de extração de características seja alcançado com eficácia, proporcionando ao classificador uma boa representação dos dados originais. Dessa forma, facilitando a tarefa deste. Veja a comparação entre as Figuras 3.3 e 3.5

Visualizando a Figura 3.3, pode-se observar o espaço multidimensional obtido a partir da base de dados original, os dados são misturados, com informação redundante. Cada ponto representa determinado valor para as características peso, pressão e cintura de um indivíduo ou observação, os de bolinha azul são não diabéticos e os de cruz vermelha, diabéticos. A figura 3.5 mostra os mesmos dados da figura anterior após a fase de extração de características, pode-se ver que ICA minimizou a dependência estatística nos dados, criando uma boa representação e entrada para a fase de classificação.

Outros autores na literatura utilizaram nas suas aplicações o estágio de extração de características para melhorar a performance do classificador. Polat [42] utilizou a Análise de componentes principais que trabalha com estatística de segunda ordem, diferente da técnica de ICA que trabalha que trabalha com estatística de alta

ordem e por isso pode tomar mais informação sobre os dados. Veja as Figuras 3.4 e 3.5 para ver a diferença de resultados entre as duas técnicas.

Para mostrar a qualidade da representação dos dados obtidos através da transformação com ICA, utilizou-se uma medida de distância entre os grupos diabéticos e não diabéticos para comparar os dados após a transformação com ICA e após a transformação com PCA.

A Tabela 3.2 mostra os resultados destas medidas entre as amostras nos dados originais, nos dados após o método de ICA e após o método de PCA. Esta relação é medida através da distância dos centros dos grupos (critério de separação) e a raiz quadrada da variância (critério de compactação) para cada classe.

Esta medida é baseada em dois critérios: Compactação e Separação. A primeira mede o quão perto estão os objetos um do outro dentro de cada grupo. A segunda mede o quão distintos e bem separados um grupo está do outro [35].Desta forma, mesmo se dois grupos estão distantes, deve-se ter certeza que eles tem uma baixa variância, desta forma os objetos em cada grupo serão bem compactados. ICA mostra que reduz a dependência estatística. Desta forma, os clusters serão devidamente compactados (de acordo com o critério de compactação), como pode ser visto na Figura 3.5 e Tabela 3.2.

Mesmo se as características dos dados reais (base de dados do HIPERDIA) são dependentes uma das outras como mostra a Figura 3.3, se diminuirmos esta dependência entre as variáveis, a classificação irá ter melhores resultados, como mostra-se a seguir:

Tabela 3.2: Medida de distância entre os grupos de diabéticos e não-diabéticos para os dados originais, para os dados antes de ICA e para os dados depois de PCA

Base de dados original	**Após PCA**	**Após ICA**
0.0128	0.5390	2.5381

Figura 3.3: Gráfico de dispersão da base de dados brasileira antes da fase de extração de características de 3 características pressão arterial sistólica, cintura e peso. A cruz vermelha mostra o grupo dos diabéticos e as bolas azuis mostra o grupo dos não-diabéticos. Este gráfico mostra que todos estão misturados.

Figura 3.4: Este é o gráfico de dispersão da amostra de treino utilizando PCA. Podemos ver que PCA pode fazer a descorrelação no conjunto de treino, mas por tomar somente informação dos momentos de segunda ordem, este método não tem a mesma boa performance de ICA, que trabalha com os momentos de alta ordem

Figura 3.5: Gráfico de dispersão da amostra de treino depois de ICA. Este gráfico mostra que ICA proporcionou uma melhor separação entre os grupos de diabéticos e não-diabéticos

3.3 Validação do método de classificação

De modo a avaliar o desempenho do classificador, é necessário que se quantifique a sua sensibilidade, especificidade e acurácia. No problema de classificação de diabéticos a sensibilidade é o parâmetro que informa o quanto o classificador é capaz de identificar as pessoas com diabetes; a especificidade informa o quanto o classificador é capaz de identificar as pessoas sem a doença ou saudáveis. Estes critérios tem as seguintes variáveis:

- Verdadeiro positivo (VP): Diagnóstico de pacientes corretamente classificados como diabéticos.
- Falso positivo (FP): Diagnóstico de pacientes não diabéticos classificados como diabéticos.
- Verdadeiro negativo (VN):Diagnóstico de pacientes corretamente classificados como não diabéticos.
- Falso negativo (FN):Diagnóstico de diabéticos classificados como não diabéticos.

e eles definem completamente a acurácia de um classificador, os quais são definidos como a seguir:

Sensibilidade	$VP/(VP+FN)$
Especificidade	$VN/(VN+FP)$
Acurácia	$(VP+FN)/(VP+VN+FP+FN)$

3.4 Resultados

Aqui descreve-se os resultados obtidos usando a metodologia proposta.

3.4.1 Teste Não invasivo

Dos 498 casos de diabéticos e dos 497 casos de não diabéticos restantes após o pré-processamento da base, separaram-se aleatoriamente as amostras usando o método de 20-validação cruzada para cada classe. Foi aplicada a decomposição dos dados através de ICA nos grupos de treino e obteve-se uma matriz A ou $\mathbf{W^{-1}}$, com 14 funções bases. Após isto, projetou-se as amostras de treino e teste neste subespaço. Este sinal projetado foi utilizado como a entrada do classificador SVM para uma classe.

A biblioteca para máquinas de vetores de suporte, chamada LIBSVM [36], foi utilizada para o treino e para o teste. A função kernel utilizada foi a RBF (radial basis function), com um valor de gamma de 0.00781. A primeira linha da Tabela 3.3 mostra os resultados encontrados para sensibilidade, especificidade e acurácia, respectivamente, para a classificação não invasiva entre diabéticos e não-diabéticos. Para verificar a influência de cada característica clínica nos resultados finais, foram feitos mais alguns testes diminuindo o número de características do vetor de entrada ou de observação, testando cada possibilidade de combinação.

Tabela 3.3: Resultados do teste não invasivo

Testes	Sensibilidade	Especificidade	Acurácia
Teste não invasivo	100%	100%	100%

3.5 Teste com todas as combinações de 14 características clínicas tomadas (14-n)

Dessa forma, foram feitos testes com as 14 características não invasivas tomadas (14-n), sendo n de 1 a 8. No final, foram feitos 12.910 possíveis testes, detalhados na tabela 3.4. Neste último teste foi observado o comportamento do framework e a influência de cada característica no resultado final. Para este teste, a

decomposição de ICA no espaço de (14-n) dimensões é feita para cada combinação das características para criar um apropriado espaço de projeções que segue para o classificador one-class. Os resultados podem ser vistos na Figura 3.6.

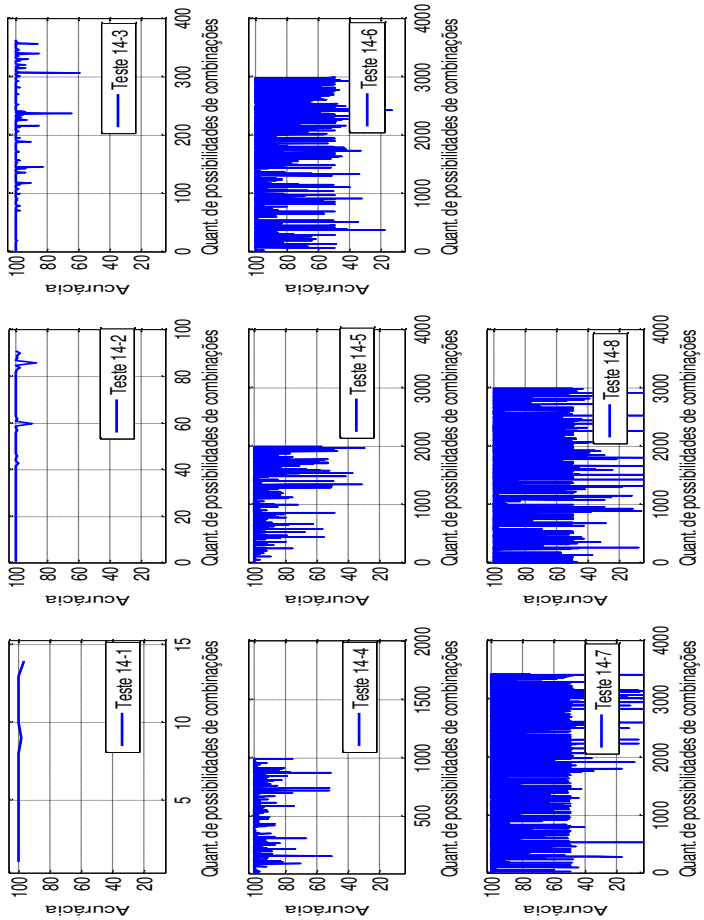

Figura 3.6: A Figura mostra graficamente os resultados em acurácia alcançados com todas as combinações das características. Com o objetivo de verificar quais características melhor influenciavam no resultado final fez-se a diminuição progressiva da quantidade de características por teste, utilizando a quantidade de 6 a 12 características por teste e em cada teste combinando as 14 características. Dessa forma, no primeiro teste foram utilizadas 13 características (o teste foi nomeado de 14-1) e todas as possibilidades de combinações das 14 caracetrísticas tomadas 13 a 13 foram testadas gerando 14 possibilidades de combinações. No último teste foram utilizadas 6 características(nomeado 14-6) e as combinações das 14 caracerísticas tomadas de 6 a 6 foram testadas, gerando a possibilidade de 3003 combinações.

Tabela 3.4: Quantidade de combinações possíveis de 14 características clínicas tomadas (14-n) e as descrições de cada teste e funções bases obtidas.

Testes	**Quant. de combinações possíveis**	**Quant. de funções bases**
1. **Teste (14-1)**	14	13
2. **Teste (14-2)**	91	12
3. **Teste (14-3)**	364	11
4. **Teste (14-4)**	1001	10
5. **Teste (14-5)**	2002	9
6. **Teste (14-6)**	3003	8
7. **Teste (14-7)**	3432	7
8. **Teste (14-8)**	3003	6

3.5.1 Influência de cada marcador no resultado final

Fez-se a análise da influência de cada característica clínica ou marcador não invasivo no resultado final de cada teste. Os marcadores que na falta fizeram o resultado decair mais ou aumentar o desvio padrão foram rastreados e encontrou-se a seguinte importância para cada grupo de marcadores em cada teste, conforme demostra a tabela 3.5:

Tabela 3.5: Quantidade de combinações possíveis de 14 características clínicas tomadas (14-n) e as descrições de cada teste. Cada teste utiliza um número de entrada de caracerísticas de 12 a 6 e em cada é feita todas as combinações das 14 caracetristicas, o que gerou 12910 possibilidades de testes.

Combinações	**Quantidade de testes**	**Média**	**Desvio Padrão**	**Ordem de importância**
$C_{14,6}$	3003	93.24%	16.13	1>14>6>4>3>9>5>7>8>2>10>13>11>12
$C_{14,7}$	3432	92.79%	16.47	1>13>10>9>8>2>4>5>7>6>3>12>11>14
$C_{14,8}$	3003	95.16%	12.40	1>7>9>4>5>3>8>10>2>6>11>12>14>13
$C_{14,9}$	2002	97.69%	7.5	1>10>7>3>8=9>2>4>13>6>12>5>11>14
$C_{14,10}$	1001	98.67%	4.79	1>5=8=12>4=11=13>2=3=10>9>6>7
$C_{14,11}$	364	99.24%	3.40	1=2=7=8=14>3=4=5=6=9=10=11=12=13
$C_{14,12}$	91	99.65%	1.74	1=5>2=3=4=6=7=8=9=10=11=12=13=14

3.6 Testes de generalidade do método com outras bases

O método foi aplicado a outras bases de dados de livre acesso e comumente utilizadas na literatura. Aqui utilizou-se elas para teste de generalidade do método proposto. A mais utilizada é a base de dados de índios Pima, ela tem a problemática de ser homogênea e composta somente de mulheres. A segunda base de dados encontrada na literatura é uma base de dados de americanos de origem africana, também é uma base de dados de uma população homogênea como a base Pima, mas são as únicas encontradas para teste em diabetes tipo 2 neste tempo.

3.6.1 Base Pima

A base de dados dos índios Pima foi obtida do repositório de aprendizado de máquina da Universidade da Califórnia, Irvine (UCI). Esta base de dados foi selecionada de uma grande base de dados cedida pelo Instituto Nacional de Diabetes e doenças digestivas dos EUA. Todas são pacientes da tribo Pima, do sexo feminino, com idade maior ou igual a 21 anos. As características presentes nesta base são:

1. Número de vezes que esteve grávida
2. Concentração de glicose no sangue a uma sobrecarga de glicose em 2 hs.
3. Pressão Diastólica (mmHg)
4. Medida do triceps (mm)
5. Insulina (mu U/ml)
6. Indíce de massa corporal
7. Função diabetes(fator hereditário)
8. Idade

Retirando-se as características ou marcadores invasivos da base de dados Pima (marcadores 2 e 5), obteve-se 6 marcadores não invasivos, onde o método proposto anteriormente foi aplicado. Obteve-se um resultado em acurácia de 98.28%. Os resultados para este teste em sensibilidade, especificidade e acurácia podem ser vistos na primeira linha da tabela 3.6.

Existem vários trabalhos na literatura utilizando esta base, mas com todos os 8 marcadores da base Pima, levam em consideração as características clínicas invasivas e não invasivas dos pacientes. Nesta perspectiva o método proposto obteve um resultado em acurácia de 98.47%, que foi um resultado acima dos obtidos por outros trabalhos da literatura, que também utilizaram todos os marcadores clínicos da base, como mostra a tabela 3.7. Pode-se ver na Tabela 3.6 os resultados obtidos em sensibilidade, especificidade e acurácia com todos os marcadores invasivos e não invasivos na segunda linha desta.

Tabela 3.6: Resultados do método com todos os marcadores (**Com 8 marcadores**) e somente com os marcadores não invasivos (**Com 6 marcadores**.)

Teste	**Sensibilidade**	**Especificidade**	**Acurácia**
Com 6 marcadores	99.05%	98.19%	98.28%
Com 8 marcadores	99.81%	98.34%	98.47%

3.6.2 Base de americanos de origem africana

A base de dados dos americanos com origem africana foi obtida da escola de medicina da Universidade de Virgínia. Os dados consistem de 19 marcadores clínicos de 403 indivíduos. Foi obtida através de um estudo sobre prevalência de obesidade, diabetes e outros fatores de riscos cardiovasculares em afro descendentes no estado de Virgínia, EUA. 60 Indivíduos diabéticos e 343 não diabéticos. Os marcadores invasivos foram retirados. Somente 7 características clínicas foram utilizadas:

1. Idade

Tabela 3.7: Trabalhos correlatos na literatura com outra base de dados(Pima, 1975). Nestes trabalhos foram utilizados marcadores invasivos e não invasivos

Método	**Acurácia**	**Autor e ano**
Método proposto com base Pima	**98.47%**	
GA and Prot. Selec.[37]	92.60 %	Byeon et al. (2008)
HPM[38]	92.38%	Patil et al.(2010)
Fuzzy[39]	91.20%	Lee and Wang (2011)
LDA-Wavelet SVM[40]	89.74%	Calisir and Dogantekin (2011)
IFE-CF[41]	89.48%	M.Reddy and L.Reddy (2010)
PCA-ANFIS[42]	89.47%	Polat and Günes (2007)
MAIRS2[43]	89.10%	Chikh and Saidi (2011)
LDA-ANFIS[44]	84.61 %	Dogantekin et al. (2010)
ANN-FNN[45]	84.24%	Kahramanli and Allahverdi (2008)
GDA-LS-SVM[46]	82.05%	Polat et al. (2008)
C-HMLP[47]	81.74%	Isa and Mamat (2011)
Fuzzy[49]	79.37%	Lekkas and Mikhailov (2008)
ANFIS[50]	77.65%	Ghazavi and Liao (2008)
Fuzzy[48]	77.8%	Luukka (2011)
ANN[51]	76.62%	Jeatrakul et al. (2010)
OP-ELM [52]	76.3%	Miche et al. (2010)
IP-LSSVM [53]	76.1%	Carvalho and Braga (2010)
FSM-FuzzyEM[54]	75.97%	Luukka (2011)
SVM[55]	75.15%	Li and Liu(2010)

2. Peso

3. Altura

4. Primeira medida de pressão sistólica (mmHg)

5. Primeira medida de pressão diastólica (mmHg)

6. Waist (inches)

7. Medida da cintura

O método proposto foi aplicado novamente a esta base e obteve-se os seguintes resultados em sensibilidade, especificidade e acurácia, demonstrados na tabela 3.8:

Tabela 3.8: Resultados do método utilizando a base de dados de americanos de origem africana

Teste	**Sensibilidade**	**Especificidade**	**Acurácia**
Base African-American	99.88%	96.68%	97.01%

4 Discussões

A diabetes é uma doença que está crescendo no mundo significativamente, até 2030 o número de diabéticos no mundo irá dobrar. Incluindo as perdas sociais como incapacitações e mortalidade prematura, a doença causa um alto custo com seu tratamento e com suas complicações. Formas de preveni-la são investigadas no mundo todo e um grande número de pessoas no momento do diagnóstico da doença já apresenta algum tipo de complicação. Um ponto que complica o rastreamento da doença são os altos custos envolvidos, já que o procedimento requer algum teste invasivo como glicemia plasmática em jejum ou glicose capilar na primeira fase do rastreamento, como foi feito no último rastreamento no Brasil em 2001. [17]

Não haviam estudos comprovando a eficácia do rastreamento e o quanto seria econômico, mesmo somando-se os custos com a classe dos não diabéticos encontrados no rastreamento. O problema era que para estes pacientes haviam os gastos com os exames de primeira fase do rastreamento e estes não representam o população objetivo do rastreamento. Atualmente, sabe-se que é um passo eficaz para combater esta doença silenciosa. Foram orçados todos os gastos, benefícios e lucros advindos com o rastreio e percebeu-se várias vantagens, principalmente na vinculação mais rápida dos pacientes com o tratamento e consequente diminuição das complicações com a doença. Fora isso, cria uma conscientização na população que é eficaz como medida preventiva. O estudo mais detalhado está presente no Anexo I.

Métodos simples, baratos e seguros podem ser utilizados para estudos epidemiológicos para a detecção precoce da doença, esta medida pode prevenir as complicações do diabetes, dar margem e visualização para traçar planos e metas no combate da doença. Uma das contribuições deste trabalho é a possibilidade, no Brasil, da utilização de dados simples da doença para a obtenção de informação útil sobre seus padrões, com o fim de utiliza-los para medida preventiva do diabetes tipo

2.

Ao longo da pesquisa, encontrou-se outras fontes na literatura que propunham o mesmo objetivo, mas com a utilização de métodos que utilizam estatística de segunda ordem ou classificadores lineares. Estes trabalhos foram validados para a população dos EUA e China. A ferramenta validada nos Estados Unidos utilizou árvore de regressão, com 75.36% de sensibilidade e 64.69% de especificidade[18]. A outra ferramenta validada para este fim, foi a da China [19] que utilizada regressão linear e tem os seguintes resultados: 83.3% e 66.5%, para sensibilidade e especificidade, respectivamente. Ver tabela 4.1.

Os primeiros testes foram feitos com a base de índios PIMA, que tem os problemas do desbalanceamento e homogeneidade. A primeira tarefa foi a busca de uma base heterogênea e representativa para a população brasileira, o que foi sanada com a disponibilização da base brasileira pelo DATASUS. A hipótese principal foi o trabalho com ênfase na independência estatística, na fase de extração de características, para a simplificação do trabalho de classificação.

Com primeiros testes para a base brasileira, alcançou-se os resultados apresentados na tabela 4.1, última linha. Foram feitos outros testes diminuindo o número de marcadores clínicos em busca dos mais importantes nos resultados finais.

Tabela 4.1: Métodos não invasivos validados para rastreamento em diabetes

Testes	**Sensibilidade**	**Especificidade**
EUA	75.36%	64.69%
China	83.3%	66.5%
Método proposto	**100%**	**100%**

Testou-se todas as combinações possíveis das variáveis, retirando-se uma a uma. Das 14 características, foram retiradas progressivamente uma característica por vez (14-n) para verificar se sua falta afetava o resultado final. Todas as combinações possíveis podem ser visualizadas nas colunas 1 e 2 da tabela 3.5. Os resultados mostram que mesmo com a diminuição do número de características, o método apresenta resultados acima ou iguais a 75 % de acurácia quando os testes

foram feitos com 6 características, como visto na primeira linha da tabela 3.5, mas com o aumento de marcadores não invasivos da doença a acurácia aumenta para números perto dos 97 %, como mostra a última linha da tabela.

Na última coluna da tabela 3.5 para cada teste estão apresentados a ordem de influência para cada marcador não invasivo. No teste com 12 características a importância de ordem dos marcadores assemelha-se ao mundo real. 1-Idade e 5-Peso tem igual influência nos resultados e são mais importantes que todos os outros marcadores no resultado final: 2-Pressão arterial sistólica 3-Pressão arterial diastólica, 4-Circunferência da cintura, 6- Altura, 7-Histórico familiar, 8-Fumante,9-Praticante de atividade física, 10- Sobrepeso, 11- AVC, 12- Cegueira, 13- Outras coronariopatias, 14- Amputação.

Nos testes com 6 marcadores a importância de cada uma foi dada por : 1>14>6>4>3>9>5>7>8>2>10>13>11>12. Em todos os testes a idade teve maior influência, nos testes com combinações de idade com características relacionadas com a gordura corporal obtiveram-se melhores resultados. na retirada destas o método deu maior importância a outras características, mas com maior desvio padrão nos resultados.

Dessa forma nota-se que estas características são chaves para o projeto de um bom rastreador da diabetes tipo 2, como pode ser observado na figura 3.6 e tabela 3.5. Pode-se perceber que a precisão do método irá aumentar conforme o maior número de características representativas da doença.

Os testes confirmam que a minimização da dependência estatística dos dados por meio da independência é um critério válido a ser utilizado na implementação da fase de extração de características de dados nao invasivos para diabetes tipo 2, criando dessa forma componentes independentes no espaço, por serem independentes as classes estão separadas umas das outras, como mostra a Figura 3.5.

Na Figura 3.3, que mostra a dispersão dos dados originais, percebe-se que é um pouco difícil utilizar um método linear para alcançar a separabilidade das

classes, isto justifica os resultados encontrados na literatura por outros autores, ver tabela 4.1. Mesmo utilizando-se na fase de extração de características um método baseado em estatística de segunda ordem como PCA, Figura 3.4, a separabilidade das classes melhora [42], mas não como os resultados adquiridos com a utilização de ICA no pré processamento. As componentes independentes são facilmente separáveis, isto por que ICA encontra informação escondida nos momentos de alta ordem que PCA e outros métodos lineares não conseguem tomar.

O método de extração de características proposto aumentou a performance do classificador na base de dados do HIPERDIA, isto porque reduziu a dependência estatística nos dados coletados, que incrementou a habilidade do classificador em encontrar resultados precisos nas limitações das classes ou grupos de diabéticos e não diabéticos. Entretanto, não há referencial teórico que garanta que ICA irá sempre incrementar a performance do classificador, entretanto a metodologia proposta pode ser sempre validada na prática.

Nos casos da aplicação do método em bases com poucas amostras e características e desbalanceadas (mais amostras de uma classe do que de outra) como é o caso da base de americanos de origem africana e índias Pima, o método alcançou resultados menos significantes em relação aos resultados encontrados com a utilização da base brasileira. Para que se encontrem resultados satisfatórios é importante que a base de dados utilizada seja balanceada e com o maior número de características possíveis que representem a doença. Isto é natural pois no mundo real para encontrar um determinado ente, é preciso o maior número de informações deste.

A principal meta deste trabalho a longo prazo é usar somente testes não invasivos para rastrear pacientes com potencial a diabetes, isto pode aumentar a escala de diagnóstico do diabetes em áreas rurais, por exemplo, diminuindo os custos no rastreamento com a aplicação de recursos tecnológicos simples. Os primeiros testes foram realizados na base de dados de índios PIMA, mesmo com a diminuição de características invasivas, a diminuição do desempenho do método foi mínimo, com a validação do método para uma base de dados brasileira (HIPERDIA), fica

mais fácil de visualisar as potencialidades da técnica. O quadro epidemiológico da doença poderia ser mais real, com o poder de alcance em uma população maior, as medidas de prevenção para as complicações poderiam ser realmente planejadas e efetivadas, diminuindo assim os custos e males que o diabetes tipo 2 acarreta. O rastreamento como falado anteriormente, não descarta a realização do teste invasivo de glicemia em jejum para a confirmação do diagnóstico, mas pode levar a este teste as possíveis pessoas diabéticas, descartando os custos com as pessoas que não apresentam a doença.

Para os trabalhos futuros a ênfase no relacionamento do porquê do funcionamento da técnica para bases de diabéticos tipo 2 deve ser estudada, a priori, a hipotese que mais se encaixa é que o processo de codificação eficiente parece ser parecido com o de formação e processamento da diabetes tipo 2 no corpo humano: Causas escondidas, $\mathbf{A}$, estavam presentes no processo de formação da doença e estimularam independentemente, $\mathbf{s}$, para a formação do estado final da doença, $\mathbf{x}$. Encontrando-se as causas, pode-se encontrar as componentes independentes que descrevem a doença. No espaço parece que as componentes independentes dos diabéticos encontram-se separadas das componentes independentes dos não diabéticos, o que facilita a separação das classes até por meio de um classificador linear. É bom que isto, primeiramente, seje validado para que o método possa ser utilizado com rigor científico.

Outro ponto a ser estudado futuramente são os padrões de diabetes tipo 1 e diabetes gestacional para comparação com os padrões encontrados para a diabetes tipo 2, como estariam dispostas no espaço as componentes independentes detes tipos? e a relação com as componentes observadas da diabetes tipo 2.

Além destes pontos a serem estudados futuramente, algumas relações estatísticas que poderiam transforma-se em métodos de aprendizado de máquina para cálculo entre padrões aceitáveis de perfil glicêmico e taxa de hemoglobina glicada para acompanhamento clínico de pacientes diabéticos, conforme recentes trabalhos publicados em revistas de referência da ADA. A relação entre estes dois entes ainda é discutida por médicos ainda como não aceitável. No Brasil a

sociedade brasileira de diabetes recomenda um limite diferente do recomendado pela Associação de diabetes americana. Estudos clínicos demonstram que o padrão da ADA foi mais aceitável, mas ainda com necessidades de comprovação estatística da relação[56]. Informação obtida por meio de momentos de alta ordem para esta questão podem evoluir resultados aceitáveis e clinicamente direcionados.

5 Publicações

Os estudos descritos nesse trabalho geraram as seguintes produções:

Congresso:

Tracking Type 2 Diabetes Using Sparse Coding Áurea Ribeiro, Allan K. Barros, Ewaldo Santana e Roseane Diniz, American Diabetes Association's (ADA) 75 Scientific Sessions, Boston.

Periódico aceitos:

Feature Extraction in Diabetes Diagnosis Using Effcient Coding,Qualis B1, Áurea Ribeiro, Allan K. Barros, José Carlos Príncipe, Ewaldo Santana, Revista Brasileira de Engenharia Biomédica (Research on Biomedical Engineering).

Tracking Type 2 Diabetes Using Sparse Coding,Qualis A2, Áurea Ribeiro, Allan K. Barros, Ewaldo Santana e Roseane Diniz, Diabetes(ADA).

Periódico a ser enviado a revista:

Non-invasive automated screening of Diabetes, Qualis A2, Áurea Ribeiro, Allan K. Barros, Ewaldo Santana, Daniel Costa. Diabetes Care(ADA).

Referências Bibliográficas

[1] Malerbi D A, Franco L J. The Brazilian Cooperative Group on the Study of Diabetes Prevalence. Multicenter Study of the Prevalence of diabetes mellitus and Impaired Glucose Tolerance in the urban Brazilian population aged 30-69 years. Diabetes Care, 15,11, 1509-16, 1992.

[2] Sociedade Brasileira de Diabetes, "Complicações do diabetes e principais co-morbidades"no livro publicado "Diabetes na prática clínica e-book 2011", 2013, modulo 2 [Online].Disponível em: http://http://www.diabetesebook.org.br/

[3] L. BAHIA ."Os Custos do Diabetes Mellitus". [Online]. Disponível em: http://www.diabetes.org.br/educacao-continuada/491-os-custos-do-diabetes-mellitus>. Acesso em: 19 de abr.2013.

[4] American Diabetes Association. "Economic Costs of Diabetes in the U.S. in 2007". Diabetes Care,vol. 31,n.3,p.596-615, Mar, 2008.

[5] American Diabetes Association. "Economic Costs of Diabetes in the U.S. in 2012". Diabetes Care,vol. 36,n.4,p.1033-1046, April, 2013.

[6] M. Marinho, E. Cesse et al. "Análise de custos da assistência à saúde aos portadores de diabetes melito e hipertensão arterial em uma unidade de saúde pública de referência em Recife-Brasil". Arq Bras Endocrinol Metab. vol. 55, n.06, p.406-411, 2011.

[7] Brasil-Ministério da Saúde-Secretaria-Executiva. "Área de Economia da Saúde e Desenvolvimento. Avaliação econômica em saúde: desafios para gestão no Sistema Único de Saúde". Brasília: MS; 2008.

[8] C.L. Blake, C.J. Merz, UJI Repository of Machine Learning Databases [Online], availabe at http://www.ics.uci.edu/ mlearn/MLRepository.html, accessed in August 2010.

[9] N.V. Chawla, K.W. Bowyer, L.O Hall, W.P Kegelmeyer. "SMOTE:Synthetic Minority Over-sampling Technique". arXiv preprint,1106, 2011.

[10] DATASUS, HIPERDIA, 2014.

[11] DATASUS, "Sistema de gestão clínica de hipertensão arterial e diabetes mellitus da atenção básica", [Online], disponível em: http://hiperdia.datasus.gov.br/, acessado em : Janeiro de 2013

[12] C. M. Toscano. "As campanhas nacionais para detecção das doenças crônicas não-transmissíveis: diabetes e hipertensão arterial. Ciênc Saúde Coletiva, 9.4, pp.885-95, 2004.

[13] J.M.G. Wilson, G. Jungner. "Principles and Practice of Screening for disease". Geneva: World Health Organization, 1968, pp. 26-27.

[14] American Diabetes Association. "Screening for Diabetes". Diabetes Care, Estados Unidos, vol.25, suppl. 1,2002, p.521-524.

[15] R. Kahn, P. Alperin et al. "Age at initiation and frequency of screening to detect type 2 diabetes: a cost-effectiveness analysis", v.375, ,p. 1365-2010,Abr.2010.

[16] R. Chatterjee, K.M. Narayan et al. "Screening for Diabetes and Prediabetes should be cost-saving in patients at high risk", Diabetes Care, v.36, n.1,p. 1981-1987,Jul.2013.

[17] A. George, B. Duncan et al. "Análise econômica de programa para rastreamento do diabetes mellitus no Brasil".Revista de Saúde Pública, vol. 39, n. 03, p.452, 2005.

[18] E. KENNETH, M. DAVID et al, "Diabetes Risk Calculator", Diabetes Care, 5, pp.1040-1045, 2008.

[19] J. J. Dong et al., “Evaluation of a risk factor scoring model in screening for undiagnosed diabetes in China population". Journal of Zhejiang Univ Sci B, 10, pp.846-852,2011.

[20] A.D. Kulkarni. “Feature Extraction", “Artificial neural networks for image understanding", New York, USA, 1993, ch.3,sec.3.1,pp.49-55.

[21] A.K. Barros, A. Chichocki, “Neural Coding by Redundancy Reduction and Correlation", in Proc. of the VII Brazilian Symposium on Neural Networks (SBRN) (IEEE), 2002, pp. 223-226.

[22] E.P. Simoncelli, B.A. Olshausen, “Natural Image statistics and Neural Representation". Annu. Rev. Neurosci, pp.1193-216, vol. 24, 2001.

[23] R. Baddeley, L.F. Abbott, M.C. Booth, F. Sengpiel, T. Freeman, E. A. Wakeman, E.T. Rolls. "Responses of neurons in primary and inferior temporal visual cortices to natural scenes". Proc R Soc Lond B Biol Sci 264, 1775-1783, 1998.

[24] M. Deweese, M. Wehr, A. Zador. “Binary spiking in auditory cortex", J Neurosci 23, 7940-7949, 2003.

[25] D.H. Hubel, T.N. eWiesel, “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex", J. Physiol, 160, 106-154 , 1962.

[26] P. Comon. “Independent component analysis, A new concept?"Signal Processing, 36, pp.287-314, 1994.

[27] A. Hyvarinen, A. Karhunen, J. Oja. “Independent Component Analysis. John Wiley and Sons", New York (2001)

[28] A. Papoulis , S. Pillai .“Probability, Random Variables and Sthocastic Processes". 4. ed. New York : McGraw- Hill, 2002.

[29] C.J.C.Burges, “A Tutorial on Support Vector Machines for Pattern Recognition". Kluwer Academic Publishers, 1998.

[30] L. Zhuang et al. “Parameter Optimization of Kernel-based One-class Classifier on Imbalance Learning .journal of computers", vol. 1, no. 7, october/november 2006.

[31] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. Smola, R. Williamson. “Estimating the Support of a High-Dimensional Distribution". Neural Computation 13 (7)(2001) 1443- 1471.

[32] L. Manevitz, M. Yousef. “One-class SVMs for document classification", Journal of Machine Learning Research 2 (2) (2001) 139154.

[33] A. Hyvärinen, E. Oja, “Independent Component Analysis: Algorithms and Applications". Neural Networks, 13(4-5):411-430, 2000.

[34] K. Polat and S. Günes, “An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease", Digital Signal Processing, 2007, pp.702-710.

[35] Y. Liu et al., “Understanding of internal clustering validation measures", presented: IEEE International Conference on Data Mining,Australia, 2010.

[36] C. C. Chang, C. J. Lin, “LIBSVM - A Library for Support Vector Machines"(2003), available at http://www.csie.ntu.edu.tw/cjlin/libsvm/ accessed in January 2008.

[37] B. Byeon, K. Rasheed, and P. Doshi, "Enhancing the Quality of Noisy Training Data Using a Genetic Algorithm and Prototype Selection", in Proc. IC-AI, 2008, pp.821-827.

[38] B.M. Patil, R.C. Joshi, and D. Toshniwal, "Hybrid prediction model for Type-2 diabetic patients", presented at Expert Syst. Appl., 2010, pp.8102-8108.

[39] C. Lee and M. Wang, "A Fuzzy Expert System for Diabetes Decision Support Application", presented at IEEE Transactions on Systems, Man, and Cybernetics, Part B, 2011, pp.139-153.

[40] D. Calisir, E. Dogantekin, "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier", Expert Syst. Appl.,2011, pp. 8311-8315

[41] M.B. Reddy and L.S.S. Reddy, "Dimensionality Reduction: An Empirical Study on the Usability of IFE-CF (Independent Feature Elimination- by C-Correlation and F-Correlation) Measures", presented at CoRR, 2010.

[42] K. Polat and S. Günes, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease", presented at Digital Signal Processing, 2007, pp.702-710.

[43] M. Chikh, M. Saidi and N. Settouti, "Diagnosis of Diabetes Diseases Using an Artificial Immune Recognition System2 (AIRS2) with Fuzzy K-nearest Neighbor", J. Of Medical Systems, 2011, pp.1-9.

[44] E. Dogantekin, A. Dogantekin and D. Avci, "An intelligent diagnosis system for diabetes on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System: LDA-ANFIS", presented at Digital Signal Processing, 2010, pp.1248-1255.

[45] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases", presented at Expert Syst. Appl., 2008, pp.82-89.

[46] K. Polat , S. Günes , A. Arslan, "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine", presented at Expert Systems with Applications: An International Journal, January, 2008, pp.482-487.

[47] N.A.M. Isa and W.M.F.W. Mamat, "Clustered-Hybrid Multilayer Perceptron network for pattern recognition application", presented at Appl. Soft Comput., 2011, pp.1457-1466.

[48] P. Luukka, "Fuzzy beans in classification", presented at Expert Syst. Appl., 2011, pp.4798-4801.

[49] S. Lekkas and L. Mikhailov, "Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases", presented at Artificial Intelligence in Medicine, 2010, pp.117-126.

[50] S.N. Ghazavi and T.W. Liao, "Medical data mining by fuzzy modeling with selected features", presented at Artificial Intelligence in Medicine, 2008, pp.195-206.

[51] P. Jeatrakul, K.W. Wong, and C.C. Fung, "Data Cleaning for Classification Using Misclassification Analysis", presented at JACIII, 2010, pp.297-302.

[52] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse, "OP-ELM: optimally pruned extreme learning machine", presented at IEEE Transactions on Neural Networks, 2010, pp.158-162.

[53] B.P.R.D. Carvalho and A.D.P. Braga, "IP-LSSVM: A two-step sparse classifier", presented at Pattern Recognition Letters, 2009, pp.1507-1515.

[54] P. Luukka, "Feature selection using fuzzy entropy measures with similarity classifier", presented at Expert Syst. Appl., 2011, pp.4600-4607.

[55] D. Li and C. Liu, "A class possibility based kernel to increase classification accuracy for small data sets using support vector machines", ,Expert Syst. Appl., 2010, pp.3104-3110.

[56] L.A. Diehl, "Diabetes: hora de rever as metas?", Arq Bras Endocrinol Metab, 2013. (57)7, pp.545-549.

6 Anexo I

Este anexo contém as tabelas do estudo feito com 1573 pacientes sem diabetes disgnosticada que demonstra a economia nos custos financeiros e sociais advindos do diagnóstico de pacientes diabéticos e pré-diabéticos oriundos do rastreamento com cinco tipos de exames invasivos [16], a saber estes: [**1**]Glicemia plasmática de 1h depois do consumo de 50 g de glicose oral(GCTpl), [**2**]Glicemia capilar de 1 h depois da ingestão de 50 g de glicose oral(GCTcap), [**3**]Medição de glicemia plasmática aleatória (RPG), [**4**]Medição de glicemia capilar aleatória (RPG,) e [**5**]Hemoglobina Glicada (A1c).

Além da mensuração dos custos com o diagnóstico, também foram orçados os custos financeiros e sociais com o tratamento dos pacientes vindos do rastreamento com o diagnóstico de diabetes e pré-diabetes para os principais fatores que pesam no orçamento em saúde dos Estados Unidos durante 3 anos.

Os fatores foram subdivididos em : **1a**- Índice de Massa Corporal menor que $25kg/m^2$ ($IMC < 25kg/m^2$), **1b**- Índice de Massa Corporal entre $25kg/m^2$-$35kg/m^2$ ($IMC25 - 35kg/m^2$), **1c**- Índice de Massa Corporal maior que $35kg/m^2$ ($IMC > 35kg/m^2$), **2a**- Idade menor que 40 anos (Idade $<$ 40 anos, **2b**- Idade entre 40 e 55 anos (Idade 40-55 anos), **2c**- Idade maior que 55 anos (Idade $>$ 55 anos), **3a**- Pressão menor que 130 mmKg ($Pressão < 130mmKg$), **3b**- Pressão maior ou igual a 130 mmKg ($Pressão \geq 130mmKg$), **4a**- Triglicéridos baixo risco (TG baixo risco),**4b**- Triglicéridos alto risco (TG alto risco), **5a**- HDL baixo risco (HDL baixo risco), **5b**- HDL alto risco (HDL alto risco), **6a**- Cintura baixo risco (Cintura baixo risco), **6b**- Cintura alto risco (Cintura alto risco), **7a**- Histórico familiar negativo (Histórico familiar negativo), **7b**- Histórico familiar positivo (Histórico familiar positivo), sendo que os principais fatores de risco para doença observados foram: **1b**, **2b**,**3b**, **4b**, **5b**, **6b** e **7b**.

Logo a seguir são apresentadas 4 tabelas, as duas primeiras

correspondentes aos custos financeiros (tabela 6.1)e sociais (tabela 6.2) para os diabéticos e as duas últimas correspondentes aos custos financeiros (tabela 6.3) e sociais (tabela 6.4) com os pré-diabéticos.

Tabelas de vantagens financeiras do rastreamento para diabéticos

6.1 Para os diabéticos:

Tabela 6.1: Custos financeiros do sistema de saúde nos EUA com o rastreamento e tratamento de diabetesde 1573 participantes da pesquisa, na tabela pode-se observar vários grupos de risco de risco e a diminuição com os custos do paciente rastreado em relação ao não rastreamento

		Quantidade	GCTpl[1]	GCTcap[2]	RPG[3]	RCG[4]	A1c[5]	Média(SEM)	Sem rastreamento
	Custo total	1573	**U$66,8786**	U$68,375	U$67,838	U$70,888	U$81,467		U$ 95,710
	%percentagem diferença		**-30.12**	-28.56	-29.12	-25.93	-14.88	-25.72(2.80)	
1a	$IMC < 25kg/m^2$	349	U$9,292	U$8,502	U$8,582	**U$8,171**	U$11,836		U$9,305
	%percentagem de diferença		-0.15	-8.63	-7.77	**-12.19**	27.20	-0.31(7.15)	
1b	IMC entre $25 - 35kg/m^2$	888	U$33,483	U$33,173	**U$32,713**	U$34,699	U$43,908		U$ 43,867
	%percentagem de diferença		-23.67	-24.38	**-25.43**	-20.90	0.09	-18.86(4.80)	
1c	$IMC > 35kg/m^2$	336	**U$24,103;6**	U$26,700	U$26,543	U$28,018	U$25,723		U$ 42,538
	%percentagem de diferença		**-43.34**	-37,23	-37.60	-34.13	-39.53	-38.37(1.51)	
2a	$Idade < 40$	377	U$9,167	U$8,837	U$7,853	**U$6,604**	U$11,620		U$ 9,305
	%percentagem de diferença		-1.48	-5.03	-15.61	**-29.03**	-24,88	-5.26(8.93)	
2b	Idade entre $40 - 55$	744	**U$28,546**	U$29,337	U$30,736	U$33,906	U$35,125		U$38,550
	%percentagem de diferença		**-25.95**	-23.90	-20.27	-12.05	-8.88	-18.21(3.33)	
2c	$Idade > 55$	452	**U$29,165**	U$30,201	U$29,249	U$30,378	U$34,722		U$47,855
	%percentagem de diferença		**-39.05**	-36.89	-38.88	-36.52	-27.44	-35.76(2.14)	
3a	$Pressão < 130mmHg$	1171	U$40,359	U$39,435	**U$37,947**	U$38,857	U$48,924		U$46,526
	%percentagem de diferença		-13.25	-15.24	**-18.44**	-16.48	5.15	-11.65(4.29)	
3b	$Pressão \geq 130mmHg$	402	**U$26,519**	U$28,940	**U$29,890**	U$32,031	U$32,543		U$49,184
	%percentagem de diferença		**-46.08**	-41.16	-39.23	-34.87	-33.83	-39.04(2.22)	
4a	Triglicéridos baixo risco	1377	U$53,718	U$54,909	**U$52,205**	U$55,093	U$66,810		U$74,441
	%percentagem de diferença		-27.84	-26.24	-29.87	-25.99	-10.25	-24.04(3.52)	
4b	Triglicéridos alto risco	196	**U$13,161**	U$13,466	U$15,636	U$15,795	U$14,657		U$21,269
	%percentagem de diferença		**-38.12**	-36.69	-26.50	-25.73	-31.09	-31.63(2.54)	
5a	HDL baixo risco	833	U$28,052	U$29,762	**U$27,659**	U$30,776	U$35,670		U$37,220
	%percentagem de diferença		-24.63	-20.04	**-25.69**	-17.32	-4.16	-18.37(3.86)	
5b	HDL alto risco	740	U$38,826	**U$38,613**	U$40,179	U$40,112	U$45,797		U$58,489
	%percentagem de diferença		-33.62	**-33.98**	-31.31	-31.42	-21.70	-30.41(2.24)	
6a	Cintura baixo risco	808	U$25,275	**U$22,430**	U$25,065	U$23,206	U$32,513		U$29,245
	%percentagem de diferença		-13.57	**-23.30**	-14.29	-20.65	11.18	-12.13(6.11)	
6b	Cintura alto risco	765	**U$41,603**	U$45,944	U$42,773	U$47,682	U$48,954		U$66,465
	%percentagem de diferença		**-37.41**	-30.87	-35.65	-28.26	26.35	-31.71	
7a	Histórico familiar negativo	849	**U$27,406**	U$28,866	U$29,590	U$30,310	U$36,938		U$35,891
	%percentagem de diferença		**-23.64**	-19.57	-17.55	-15.55	2.92	-14.68(4.60)	
7b	Histórico Familiar positivo	724	U$39,472	U$39,509	U$38,247	U$40,578	U$44,529		U$59,819
	%percentagem de diferença		-34.01	-33.95	**-36.06**	-32.16	-25.56	-32.35(1.81)	

[**1**]GCTpl, Glicemia plasmática 1 h depois da ingestão de 50 g de glicose oral [**2**]GCTcap, glicemia capilar 1 h depois da ingestão de 50 g de glicose oral [**3**] RPG,Medição de glicemia aleatória [**4**]RCG,Medição de glicemia capilar aleatória [**5**]A1C, Hemoglobina A1c [**6**]* O menor custo para cada grupo de risco é indicado em negrito

Tabela 6.2: Custos sociais do rastreamento e tratamento do diabetes de 1573 participantes da pesquisa, observa-se como resultado uma diminuição também nos custos sociais com o rastreamento

		Quantidade	GCTpl[1]	GCTcap[2]	RPG[3]	RCG[4]	A1C[5]	Média(SEM)	Sem rastreamento
	Total	1573	**U$109,9466**	U$110,476	U$113,007	U$117,146	U$127,257		U$ 146,426
	%percentagem diferença		**-24.91**	-24.55	-22.82	-20.00	-13.09	-21.08(2.18)	
1a	$IMC < 25kg/m^2$	349	U$15,126	**U$13,995**	U$14,570	U$14,223	U$17,618		U$14,236
	%percentagem de diferença		6.26	**-1.69**	2.35	-0.09	23.76	6.12(4.61)	
1b	IMC entre $25 - 35kg/m^2$	888	U$55,833	**U$54,198**	U$55,324	U$57,917	U$68,643		U$ 67,112
	%percentagem de diferença		-16.81	**-19.24**	-17.57	-13.70	2.28	-13.01(3.93)	
1c	$IMC > 35kg/m^2$	336	**U$38,987;6**	U$42,284	U$43,114	U$45,005	U$40,996		U$ 65,078
	%percentagem de diferença		**40.09**	-35.03	-33.75	-30.84	-37.00	-35.34(1.51)	
2a	$Idade < 40$	377	U$14,486	U$13,656	U$13,347	**U$11,743**	U$17,398		U$ 14,236
	%percentagem de diferença		1.76	-4.07	-6.24	**-17.51**	-22.22	-0.77(6.54)	
2b	Idade entre $40 - 55$	744	**U$47,457**	U$47,832	U$51,557	U$56,090	U$55,072		U$58,977
	%percentagem de diferença		**-19.53**	-18.90	-12.58	-4.90	-6.62	-12.51(3.02)	
2c	$Idade > 55$	452	**U$48,003**	U$48,989	U$48,103	U$49,314	U$54,787		U$73,213
	%percentagem de diferença		**-34.43**	-33.09	-34.30	-32.64	-25,17	-31.93(1.72)	
3a	$Pressão < 130mmHg$	1171	U$66,333	**U$64,029**	U$64,381	U$65,628	U$76,047		U$71,180
	%percentagem de diferença		-6.81	-10.05	**-9.55**	-7.80	6.84	-5.47(3.13)	
3b	$Pressão \geq 130mmHg$	402	**U$43,613**	U$46,447	**U$48,627**	U$51,518	U$51,210		U$75,247
	%percentagem de diferença		**-42.04**	-38.27	-35.38	-31.54	-31.94	-35.83(1.98)	
4a	Triglicéridos baixo risco	1377	U$88,232	U$88,848	**U$87,624**	U$91,838	U$104,376		U$113,887
	%percentagem de diferença		-22.53	-21.99	**-23.06**	-19.36	-8.35	-19.06(2.75)	
4b	Triglicéridos alto risco	196	U$21,714	**U$21,628**	U$25,383	U$25,308	U$22,882		U$32,539
	%percentagem de diferença		-33.27	**-33.53**	-21.09	-22.22	-29.68	-28.14(2.56)	
5a	HDL baixo risco	833	**U$46,713**	U$48,641	U$47,011	U$51,596	U$55,315		U$56,944
	%percentagem de diferença		**-17.97**	-14.58	-17.44	-9.39	-2.86	-12.45(2.84)	
5b	HDL alto risco	740	U$63.232	**U$61.835**	U$65,997	U$65,550	U$71,943		U$89,483
	%percentagem de diferença		-29.34	**-30.90**	-26.25	-26.75	-19.60	-26.57(1.94)	
6a	Cintura baixo risco	808	U$41,280	**U$36,560**	U$41,975	U$39,665	U$50,044		U$44,741
	%percentagem de diferença		-7.74	**-18.29**	-6.18	-11.35	11.85	-6.34(5.00)	
6b	Cintura alto risco	765	**U$68,666**	U$73,916	U$71,033	U$77,481	U$77,214		U$101,685
	%percentagem de diferença		**-32.47**	-27.31	-30.14	-23.80	-24.07	-27.56(1.69)	
7a	Histórico familiar negativo	849	**U$45,410**	U$46,664	U$49,902	U$50,875	U$57,135		U$54,910
	%percentagem de diferença		**-17.30**	-15.02	-9.12	-7.35	4.05	-8.95(3.73)	
7b	Histórico Familiar positivo	724	U$64,536	U$63,812	U$63,106	U$66,271	U$70,122		U$91,517
	%percentagem de diferença		-29.48	-30.27	**-31.04**	-27.59	-23.38	-28.35(1.37)	

[1]GCTpl, Glicemia plasmática 1 h depois da ingestão de 50 g de glicose oral [2]GCTcap,glicemia capilar 1 h depois da ingestão de 50 g de glicose oral [3] RPG,Medição de glicemia plasmática aleatória [4]RCG,Medição de glicemia capilar aleatória [5]A1C, Hemoglobina A1c [6]* O menor custo para cada grupo de risco é indicado em negrito

6.2 Para os pré-diabéticos:

Tabela 6.3: Custos do sistema de saúde dos EUA com o rastreamento e tratamento de pré-diabéticos em 1573 participantes da pesquisa, como resultado observa-se uma redução com os custos da doença

		Quantidade	GCTpl[1]	GCTcap[2]	RPG[3]	RCG[4]	A1C[5]	Média (SEM)	Sem rastreamento
	Total	1573	**U$216,0076**	U$216,788	U$217,681	U$219,934	U$230,278		U$ 242,737
	%percentagem diferença		**-11.01**	-10.69	-10.32	-9.39	-5.13	-9.31(1.08)	
1a	$IMC < 25kg/m^2$	349	U$25,837	U$25,508	U$25,472	**U$24,765**	U$28,325		U$25,400
	%percentagem de diferença		1.72	0.42	0.28	**-2.50**	11.51	2.29(2.41)	
1b	IMC entre $25 - 35kg/m^2$	888	U$123,373	**U$122,629**	U$123,376	U$123,965	U$133,366		U$ 132,273
	%percentagem de diferença		-6.73	**-7.29**	-6.73	-6.28	0.83	-5.24(1.53)	
1c	$IMC > 35kg/m^2$	336	**U$66,798;6**	U$68,650	U$68,833	U$71,204	U$68,588		U$ 85,064
	%percentagem de diferença		**-21,47**	-19.30	-19.08	-16.29	-19.37	-19.10(0.83)	
2a	$Idade < 40$	377	U$25,150	U$24,750	U$23,634	**U$22,194**	U$26,956		U$ 24,430
	%percentagem de diferença		2.95	1.31	-3.26	**-9.15**	10.34	0.44(3.25)	
2b	Idade entre $40 - 55$	744	U$101,473	**U$101,468**	U$103,185	U$106,192	U$107,661		U$110,460
	%percentagem de diferença		-8.14	**-8.14**	-6.59	-3.86	-2.53	-5.85(1.14)	
2c	$Idade > 55$	452	**U$89,385**	U$90,569	U$90,863	U$91,548	U$95,662		U$107,847
	%percentagem de diferença		**-17,12**	-16.02	-15.75	-15.11	-11.30	-15.06(0.99)	
3a	$Pressão < 130mmHg$	1171	U$135,703	U$134,998	U$134,801	**U$134,329**	U$144,260		U$140,670
	%percentagem de diferença		-3.53	-4.03	-4.17	**-4.51**	2.55	-2.74(1.33)	
3b	$Pressão \geq 130mmHg$	402	**U$80,304**	U$81,790	U$82,881	U$85,605	U$86,018		U$102,068
	%percentagem de diferença		**-21.32**	-19.87	-18.80	-16.13	-15.72	18.37(1.08)	
4a	Triglicéridos baixo risco	1377	U$176,518	U$177,980	**U$176,295**	U$177,643	U$189,366		U$195,312
	%percentagem de diferença		-9.62	-8.87	**-9.74**	-9.05	-3.04	-8.06(1.27)	
4b	Triglicéridos alto risco	196	U$39,489	**U$38,808**	U$41,387	U$42,291	U$40,913	U$47,426	
	%percentagem de diferença		-16.73	**-18.17**	-12.73	-10.83	-13.73	14.44(1.34)	
5a	HDL baixo risco	833	U$92,633	U$92,798	**U$91,756**	U$93,546	U$98,661		U$99,385
	%percentagem de diferença		-6.79	-6.63	**-7.68**	-5.88	-0.73	5.54(1.24)	
5b	HDL alto risco	740	**U$123,375**	U$123,990	U$125,926	U$126,387	U$131,617		U$143,352
	%percentagem de diferença		**-13.94**	-13.51	-12.16	-11.83	-8.19	-11.92(1.01)	
6a	Cintura baixo risco	808	U$77,374	**U$75,158**	U$77,906	U$75,827	U$84,767		U$80,546
	%percentagem de diferença		-3.94	**-6.69**	-3.28	-5.86	5.24	-2.90(2.13)	
6b	Cintura alto risco	765	**U$138,633**	U$141,630	U$139,775	U$144,107	U$145,512		U$162,191
	%percentagem de diferença		**-14.52**	-12.68	-13.82	-11.15	-10.28	-12.49(0.79)	
7a	Histórico familiar negativo	849	**U$99,059**	U$99,741	U$100,078	U$100,072	U$107,167		U$105,102
	%percentagem de diferença		**-5.75**	-5.10	-4.78	-4.79	1.96	-3.69(0.142)	
7b	Histórico Familiar positivo	724	**U$116,949**	U$117,047	U$117,603	U$119,861	U$123,112		U$137,636
	%percentagem de diferença		**-15.03**	-14.96	-14.55	-12.91	-10.55	-13.60(0.85)	

[1]GCTpl, Glicemia plasmática 1 h depois da ingestão de 50 g de glicose oral [2]glicemia capilar 1 h depois da ingestão de 50 g de glicose oral [3] RPG,Medição de glicemia plasmática aleatória [4]RCG,Medição de glicemia capilar aleatória [5]A1C, Hemoglobina glicada [6]* O menor custo para cada grupo de risco é indicado em negrito

Tabela 6.4: Custos sociais para o rastreamento e tratamento dos pré-diabéticos de 1573 participantes da pesquisa

		Quantidade	GCTpl[1]	GCTcap[2]	RPG[3]	RCG[4]	A1c[5]	Média (SEM)	Sem rastreamento
	Total	1573	U$325,735	U$317,906	U$319,483	**U$315,107**	U$327,261		U$ 242,737
	%percentagem diferença		2.82	0.35	0.85	**-0.53**	3.31	1.36(0.73)	
1a	$IMC < 25kg/m^2$	349	U$39,212	U$37,952	U$37,418	U$35,840	U$38,808		**U$32,642**
	%percentagem de diferença		20.13	16.27	14.63	9.80	18.89	15.94(1.81)	
1b	IMC entre $25 - 35kg/m^2$	888	U$186,676	U$178,817	U$180,106	U$175,343	U$187,860		U$ **169,482**
	%percentagem de diferença		-10.15	5.51	6.27	3.46	10.84	7.24(1.41)	
1c	$IMC > 35kg/m^2$	336	**U$99,847**	U$101,138	101,959	U$103,923	U$100,593		U$114,662
	%percentagem de diferença		**-12,92**	-11,79	-11.08	-9.37	-12.27	-11.49(0.61)	
2a	$Idade < 40$	377	U$36,334	U$35,174	U$34,794	U$32,386	U$36,638		U$ 31,536
	%percentagem de diferença		15.21	11.54	10.33	2.70	16.18	11.19(2.39)	
2b	Idade entre $40 - 55$	744	U$152,904	U$147,879	U$151,918	U$152,041	U$152,460		U$**142,536**
	%percentagem de diferença		7.27	3.75	6.58	6.67	6.96	6.25(0.64)	
2c	$Idade > 55$	452	**U$136,498**	U$134,854	U$132,771	**U$130,679**	U$138,162		U$142,715
	%percentagem de diferença		-4.36	-5.51	-6.97	**-8.43**	-3.19	-5.69(0.93)	
3a	$Pressão < 130mmHg$	1171	U$204,053	U$196,200	U$197,720	U$192,754	U$204,095		**U$180,041**
	%percentagem de diferença		13.34	8.98	9.82	7.06	13.36	10.51(1.24)	
3b	$Pressão \geq 130mmHg$	402	**U$121,683**	U$121,706	U$121,763	U$122,353	U$123,165		U$136,745
	%percentagem de diferença		**-11.02**	-11.00	-10.96	-10.53	-9.93	10.68(0.21)	
4a	Triglicéridos baixo risco	1377	U$265,986	U$260,969	U$258,605	U$254,709	U$270,144		**U$253,849**
	%percentagem de diferença		4.78	2.81	1.87	0.34	-6.42	-3.24(1.07)	
4b	Triglicéridos alto risco	196	U$59,749	**U$56,937**	U$60,878	U$60,398	U$57,117		U$62,938
	%percentagem de diferença		-5.07	**-9.53**	-3.27	-4.04	-9.25	6.23(1.32)	
5a	HDL baixo risco	833	U$140,664	U$136,541	U$135,779	U$135,255	U$138,356		U$**128,767**
	%percentagem de diferença		9.24	6.04	5.45	5.04	7.45	6.64(0.77)	
5b	HDL alto risco	740	**U$185,072**	U$181,365	U$183,703	**U$179,851**	U$188,905		U$188,019
	%percentagem de diferença		-1.57	-3.54	-2.30	**-4.34**	0.47	-2.25(0.84)	
6a	Cintura baixo risco	808	U$117,250	U$109,456	U$112,739	U$109,021	U$118,934		**U$103,762**
	%percentagem de diferença		13.00	5.49	8.65	5.07	14.62	9.37(1.93)	
6b	Cintura alto risco	765	U$208,485	U$208,450	U$206,744	**U$206,085**	U$208,327		U$213,024
	%percentagem de diferença		-2.13	-2.15	-2.95	**-3.26**	-2.20	-2.54(0.24)	
7a	Histórico familiar negativo	849	**U$149,740**	U$145,642	U$147,278	U$142,579	U$150,622		**U$134,924**
	%percentagem de diferença		10.98	7.94	9.16	5.67	11.64	9.08(1.07)	
7b	Histórico Familiar positivo	724	U$175,995	U$172,264	**U$172,205**	U$172,528	U$176,638		U$181,863
	%percentagem de diferença		-3.23	-5.28	**-5.31**	-5.13	2.87	-4.36(0.54)	

[1]GCTpl, Glicemia plasmática 1 h depois da ingestão de 50 g de glicose oral. **[2]**GCTcap, Glicemia capilar 1 h depois da ingestão de 50 g de glicose oral **[3]** RPG,Medição de glicemia plasmática aleatória **[4]**RCG,Medição de glicemia capilar aleatória **[5]**A1C, Hemoglobina glicada **[6]*** O menor custo para cada grupo de risco é indicado em negrito