MODELOS DE INTELIGÊNCIA COMPUTACIONAL PARA APOIO À TRIAGEM DE PACIENTES E DIAGNÓSTICO CLINICO DE TUBERCULOSE PULMONAR

Luís Victor Coelho Cascão

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientadores: José Manoel de Seixas
Afrânio Lineu Kritski

Rio de Janeiro
Outubro de 2011

MODELOS DE INTELIGÊNCIA COMPUTACIONAL PARA APOIO À TRIAGEM DE PACIENTES E DIAGNÓSTICO CLINICO DE TUBERCULOSE PULMONAR

Luís Victor Coelho Cascão

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. José Manoel de Seixas, D.Sc.

Prof. Afrânio Lineu Kritski, D.Sc

Prof. Mariane Rembold Petraglia, Ph.D.

Dr. Paulo Albuquerque da Costa, D.Sc.

Prof. Alcione Miranda dos Santos, D.sc.

Prof. Marley Maria Bernardes Rebuzzi Vellasco, Ph.D.

RIO DE JANEIRO, RJ – BRASIL

OUTUBRO DE 2011

Cascão, Luís Victor Coelho

Modelos de Inteligência Computacional para Apoio à Triagem de Pacientes e Diagnóstico Clinico de Tuberculose Pulmonar/Luís Victor Coelho Cascão. – Rio de Janeiro: UFRJ/COPPE, 2011.

XIV, 109 p.: il.; 29, 7cm.

Orientadores: José Manoel de Seixas

Afrânio Lineu Kritski

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2011.

Referências Bibliográficas: p. 99 – 104.

1. Sistema de Apoio a Diagnostico. 2. Mapas Auto-Organizáveis. 3. Tuberculose. 4. Escore de Triagem e Diagnóstico. I. Seixas, José Manoel de *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Aos meus pais Luiz Augusto e

Eliane, ao meu irmão Luís

Bernardo e à toda minha família.

Agradecimentos

- A Deus, pela saúde e disposição que me permitiram a realização deste trabalho.
- À minha família pela educação que me deram e pela infra-estrutura que me permitiu mais esta conquista. Além disto, agradeço pelo contínuo incentivo, pelo investimento e pela paciência.
- À minha namorada Juliana Freixo e a sua família pelo apoio, compreensão, carinho e ajuda incondicional.
- Ao meu orientador José Manoel de Seixas pela confiança e toda a sua ajuda que foram fundamentais para realização desse trabalho.
- A todos do Programa Acadêmico de Tuberculose da Faculdade de Medicina da UFRJ, em especial ao meu orientador Dr. Afrânio Kritski e ao Dr. Paulo Albuquerque pelas discussões, sugestões e críticas realizadas ao longo deste trabalho
- Aos amigos de longa data pelo apoio, pela paciência e pelos momentos de descontração.
- Aos colegas que ajudaram na realização do mestrado em especial: Rodrigo Torres, Felipe Grael, Fernando Ferreira, João Baptista, Thiago Ciodaro, Diego Rodrigues, Eduardo Simas, José Marcio Faier, Natanael Moura, Moura Jr. e Andressa Sivolella pelas dicas e sugestões dadas neste trabalho.
- Aos funcionários do LPS por estarem sempre dispostos a ajudar e pela infra-estrutura disponibilizada no laboratório.

- Ao Prof. Antônio Carlos Fernandes, Ivan Falcão, Fábio Moreira Coelho, Luiz Antônio Ferreira, Anderson Araújo do Santos e Anderson Ricardo Soares e todos os alunos do Laboratório de Ondas e Correntes pelo apoio e companheirismo dado ao longo desses últimos três anos.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

MODELOS DE INTELIGÊNCIA COMPUTACIONAL PARA APOIO À TRIAGEM DE PACIENTES E DIAGNÓSTICO CLINICO DE TUBERCULOSE PULMONAR

Luís Victor Coelho Cascão

Outubro/2011

Orientadores: José Manoel de Seixas
Afrânio Lineu Kritski

Programa: Engenharia Elétrica

A tuberculose (TB) é uma das principais enfermidades que acomete a humanidade. Um dos principais desafios para o controle da TB é a baixa taxa de detecção dos casos. Como a tosse crônica é o sintoma mais comum da TB pulmonar, não tem sido eficaz a recomendação de que todos os pacientes que apresentem tosse por 3 semanas ou mais devem fazer o exame de escarro. Portanto, novas tecnologias e abordagens que sejam apropriados para o uso em países em desenvolvimento são necessárias para que se tenha um diagnóstico rápido dos casos de TB. Neste trabalho é proposto a utilização de métodos de otimização e processamento neural para o apoio à triagem dos pacientes e ao diagnóstico médico da TB, utilizando um conjunto de pacientes de referência atendidos na Policlínica Augusto Amaral Peixoto, situado no bairro de Guadalupe, no Rio de Janeiro. Baseado num questionário de sintomas é identificado o grupo de risco e calculado o escore de triagem, que dá a chance do paciente ser portador de TB pulmonar, com 81,4% de sensibilidade e 61,3% de especificidade.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

MODELS OF COMPUTATIONAL INTELLIGENCE TO SUPPORT PATIENT SCREENING AND CLINICAL DIAGNOSIS OF PULMONARY TUBERCULOSIS

Luís Victor Coelho Cascão

October/2011

Advisors: José Manoel de Seixas

Afrânio Lineu Kritski

Department: Electrical Engineering

Tuberculosis (TB) is one of the main deseases affecting mankind. One of the main challenge for TB control is the low rate of detection of cases. As chronic cough is the most common symptom of pulmonary TB, has not been effective the recommendation that all patients suffer from cough for three weeks or more should do the sputum smear examination. Therefore, new diagnostic technologies and approaches that are appropriate for usage in developing countries are necessary in order to have a rapid diagnosis of TB cases. This work proposes the use of optimization methods and neural processing to support patient screening and medical diagnosis of TB, using a reference set of patients treated at the Health Center Augusto Amaral Peixoto,located on the neighborhood of Guadalupe at Rio de Janeiro. Based on a set of symptoms is identified the patient's risk group and calculated the screening score, which gives the patient's chance of having contracted tuberculosis, with 81.4% of sensitivity and 61.3% of specificity.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Devido aos avanços da informática temos que lidar com volume de dados cada vez maiores e de maior complexidade, devido a diferentes possibilidades de usos que possam existir nos sistemas informatizados existentes. Portanto, cada vez mais se faz necessária a utilização de sistemas de apoio à decisão (SAD). Os SADs são sistemas que ajudam o homem em tarefas que envolvem tomadas de decisão, compilando uma grande quantidade de dados a serem analisados, documentos, conhecimento prévio sobre o tema ou modelos matemáticos para identificar e achar uma solução que auxilia a decisão requerida para o problema.

Os SAD são amplamente utilizados na área de finanças, na análise de crédito; marketing, no auxilio à definição do público-alvo para as campanhas; engenharia, no suporte ao gerenciamento de custos de projetos e, em particular, na área médica.

Na área médica, o principal objetivo do SAD é auxiliar no serviço médico de diagnóstico e prognóstico nas unidades de saúde que prestam a atenção primária, fazendo com que o profissional de saúde interaja com o sistema, que o pode apoiar na triagem, diagnóstico ou acompanhamento de um determinado paciente.

Em geral, é desejável que os sistemas de apoio à triagem e diagnóstico médico apresentem as seguintes características: alta eficiência na detecção da doença em questão, aliada a uma baixa incidência de falsos alarmes (falsos positivos); fácil implementação e uso; propriedades que agilizem o serviço médico. E com os avanços nas técnicas de inteligência computacional e aprendizado de maquinas, almeja-se

que os sistemas aprendam com as experiências passadas e/ou reconheçam padrões ou características relevantes nos dados clínicos dos pacientes.

O crescente aumento de informação que pode estar disponível sobre o paciente para o profissional de saúde pode dificultar o julgamento clínico, principalmente em reconhecer padrões recorrentes, devido à dificuldade inerente do ser humano em tratar de dados em alta-dimensionalidade.

Portanto, é possível usar métodos baseados no conhecimento especialista sobre o problema, juntamente com métodos estatísticos lineares e não-lineares de reconhecimento de padrões e extração de característica , de modo que se possa desenvolver um sistema de apoio a decisão clínica.

1.1 Motivação

A potencialidade de utilização de sistemas de apoio a diagnóstico, baseados em estatísticas, na área médica, deve-se a diversos fatores econômicos-socias, bem como, pelo fato de algumas doenças apresentarem testes diagnósticos de sensibilidade limitada, como é o caso da tuberculose (TB).

A TB é umas das principais enfermidades que acometem a humanidade e constitui um sério problema de saúde pública. Segundo a Organização Mundial de Saúde (OMS), aproximadamente um terço da população mundial esta infectada por *Mycobacterium tuberculosis*, agente causador da tuberculose. No ano de 2009 foram registrados 9,4 milhões de novos casos e houve 1,7 milhões de mortes devido à enfermidade, apesar de existir tratamento medicamentoso com elevada eficácia.

No Brasil, segundo o Ministério da Saúde, em 2009, foram registrados 73.598 novos casos de TB e 70.601 novos casos em 2010. O Rio de Janeiro é o estado com a maior taxa de incidência da doença de todo o país, com 71,79 novos casos por cem mil habitantes no ano de 2010.

Um dos fatores que mais favoreceu o aumento da incidência da TB foi a co-infecção pelo vírus da imunodeficiência humana (HIV). A associação (HIV/TB) constitui um sério problema de saúde pública, podendo levar ao aumento da morta-

lidade pela tuberculose.

Outros fatores sócio-econômicos, como a falta de sistemas públicos de saúde eficientes a desigualdade social, o crescimento da população marginalizada rural e urbana são relacionados com o aumento da incidência de TB.

A TB pulmonar é uma doença infecto-contagiosa, que é transmitida pelo ar. A importância da doença pulmonar decorre da forma de transmissão da TB, que ocorre por via aérea através da inalação do agente causador, eliminado no meio ambiente pelo doente com TB pulmonar. Entretanto, a TB também pode ocorrer em qualquer área do corpo humano, TB extrapulmonar, sendo mais frequentes na pleura e nos linfonodos. Também pode atingir o sistema urogenital, ossos, articulações, fígado, baço, sistema nervoso central e pele.

O diagnóstico da TB pulmonar é feito com base nos sinais e sintomas relatados pelo paciente, associados ao uso de testes diagnósticos. A baciloscopia e a cultura para micobactéria têm sido indicados como dois testes fundamentais para diagnóstico da tuberculose pulmonar.

A baciloscopia direta do escarro é o exame prioritário para os casos suspeitos de TB pulmonar, porque permite descobrir a fonte mais importante de infecção, que é o paciente bacilífero. Por ser um método simples e seguro, é praticado em todos os serviços de saúde que disponham de laboratório. Entretanto, a baciloscopia possui uma baixa sensibilidade, podendo apresentar resultados falsos-negativos de 30 a 40% dos casos.

A cultura é o teste mais sensível para o diagnóstico da TB pulmonar e considerada padrão ouro. Sendo este teste indicado para suspeitos de tuberculose pulmonar persistentemente negativos ao exame direto do escarro. Entretanto, poucas unidades de saúde primárias ou hospitalares têm acesso à cultura. Como a cultura, em meio sólido, leva de 4-8 semanas para ser interpretada, impossibilita seu uso como primeira linha de diagnóstico. Existem métodos automatizados de cultura em meio liquido, logo, mais rápidos e mais caros; porém, ficam restritos a laboratórios de referência.

Nos casos suspeitos de tuberculose pulmonar paucibacilar, sem expectoração ou

com baciloscopia negativa, constituem cerca de 29% dos casos notificados de tuberculose pulmonar no Brasil em 2009. Portanto, na ausência da cultura, muitos casos paucibacilares são diagnosticados com base nos sintomas clínicos, exames radiológicos e outros testes laboratoriais, tendo a chance diminuída de um diagnóstico correto usando técnicas rotineiras.

O diagnóstico da TB, da forma como usualmente é feito, tende a produzir um atraso na identificação do doente, retardando-lhe o tratamento e permitindo que a transmissão do bacilo ocorra, contaminando outras pessoas.

1.2 Objetivos

Os exames rotineiros, utilizados para o diagnóstico de TB pulmonar, nem sempre são suficientes para a detecção dos pacientes com TB pulmonar ativa. A falha na detecção do paciente portador do bacilo facilita a transmissão da doença. Segundo a OMS, a melhoria na detecção dos casos implica diretamente na diminuição da incidência da doença.

Logo, obter um sistema que, alimentado com dados clinícos que tenham qualidade e que sejam representativos de determinada realidade, seja capaz de identificar um paciente com TB pulmonar ativa, pode colaborar significativamente na prática clínica e na tomada de decisão clínica, assim como, entender o relacionamento dos sintomas com a doença.

Este trabalho objetiva produzir um escore cliníco para TB pulmonar de fácil utilização pela equipe de enfermagem em sítios com recursos limitados. Para tal, serão utilizadas redes neurais não-supervisionadas do tipo Self Organizing Map (SOM), procurando obter agrupamentos entre pacientes, com base nos sintomas e sinais declarados, e a relação entre os sintomas clínicos e a presença de TB pulmonar na população em estudo, e métodos de otimização e classificação para elaboração do escore em si.

O escore a ser obtido deve ser de fácil utilização em postos de saúde que não disponham de recursos computacionais, onde não possam ser utilizados sistemas

de apoio ao diagnósticos mais refinados. Partindo destas premissas, tal sistema de ponderação, escore, se restringirá a uma ponderação composta somente por número inteiros, visando um rápido cálculo pelo profissional de saúde, agilizando a triagem dos pacientes e, eventualmente, orientando o tratamento da doença em situações onde os recursos humanos sejam bastante limitados.

Objetiva-se como desdobramento desse trabalho, a utilização do escore em parceria com a Faculdade de Medicina na triagem dos pacientes atendidos no Hospital Universitário Clementino Fraga Filho e na Policlínica Augusto do Amaral Peixoto, onde os dados para este trabalho foram coletados.

1.3 Apresentação do Trabalho

No próximo capítulo, é apresentada uma breve introdução da doença em estudo, definindo os principais sintomas da doença e o seu processo de diagnóstico. Assim como uma revisão bibliográfica sobre diferentes trabalhos desenvolvidos para sistemas de apoio a decisão na área médica e técnicas utilizadas no diagnóstico da TB, por meios de inteligência computacional.

No capítulo 3, é apresentada a base de dados a ser utilizada e o escore que hoje em dia é utilizado na Policlínica Augusto do Amaral Peixoto que será a referência de eficiência deste trabalho.

No capítulo 4 é realizada a investigação inicial sobre a existência de agrupamentos nos casos da base de dados, e o relacionamento entre os sintomas clínicos e os agrupamentos, utilizando redes neurais não-supervisionadas do tipo SOM. Também são descritas as diversas técnicas que foram utilizadas no projeto dos diferentes escores de triagem de TB pulmonar desenvolvido.

Já no capítulo 5, são apresentados os resultados obtidos para os diferentes escores produzidos neste trabalho. As conclusões e discussões sobre o uso do escore no diagnóstico da TB pulmonar são apresentadas no capítulo 6, assim como as perspectivas futuras para a continuidade dos trabalhos de pesquisa.

Capítulo 2

Tuberculose Pulmonar

Neste capítulo, será apresentada a doença em estudo, trazendo na seção 2.1 um panorama sobre a TB no mundo. Na seção 2.2 serão abordados os fatores de risco associados com a tuberculose. Já na seção 2.3, serão definidos os sinais e sintomas da doença. Por fim, a seção 2.4 mostra o processo de diagnóstico da tuberculose pulmonar e o sistemas de apoio a essa tarefa existentes .

2.1 Introdução

A tuberculose é, certamente, uma das mais antigas doenças que afligem a humanidade[1]. No cenário brasileiro, vem se firmando como uma das principais causas de morbi/mortalidade, atingindo indistintamente diversas faixas etárias e classes sociais [3].

A principal fonte de infecção é o homem, e raramente algumas regiões, o gado bovino [4]. Entende-se por fonte de infecção qualquer vetor capaz de transmitir o bacilo da tuberculose. Em geral, a fonte de infecção é o indivíduo com a forma pulmonar da doença, eliminando bacilos para o exterior. Calcula-se que durante um ano, numa comunidade, uma fonte de infecção poderá infectar, em média, de 10 a 15 pessoas que com ela tenham tido contato [5].

Estima-se que cerca de 2 bilhões de indivíduos em todo o mundo estejam infectados por *Mycobacterium tuberculosis*, correspondendo a 30% da população mundial,

sendo que novas infecções ocorrem a uma taxa de uma por segundo [1]. A proporção de pessoas que contraem TB a cada ano está estável ou decaindo mundialmente porém, devido ao crescimento populacional , os números absolutos de novos casos continua crescendo.

Segundo a OMS, em 2009, estima-se a ocorrência de 9,4 milhões de novos casos e 1,3 milhões de mortes [1]. A distribuição dos casos de tuberculose não é uniforme pelo mundo, como pode ser visto na figura 2.1. Um total de 22 países, principalmente da Ásia e África, é responsável por 80% dos casos de tuberculose. No Brasil, 19º país em numero de casos de TB, no mesmo período de tempo, foram notificados 94 mil casos de doentes crônicos, 87 mil novos casos detectados e ocorreram 4 mil mortes devido a tuberculose [1] [6].

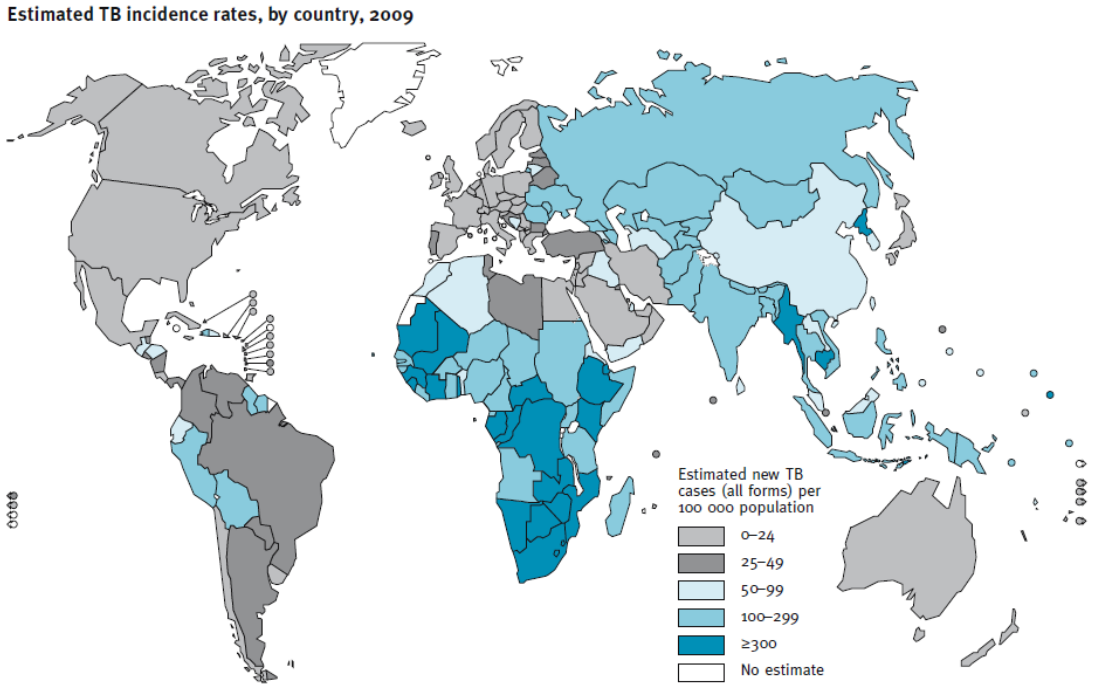


Figura 2.1: Estimação da taxa de incidência de TB em 2009. Fonte OMS [1]

Analisando-se mais a fundo os casos de tuberculose no Brasil, as regiões Norte e Centro-Oeste são as regiões que têm a maior e a menor taxa de incidência do *Mycobacterium tuberculosis* respectivamente. Se tratando dos estados individualmente, o estado do Rio de Janeiro tem a maior incidência de novos casos da doença 74,06 por 100.000 habitantes enquanto a média nacional foi de 38,41 novos casos por 100.000

habitantes, no ano de 2009[6]. Já no município do Rio de Janeiro, onde foram coletados os dados deste trabalho, em 2009 a taxa de incidência da tuberculose foi de 66,4 novos casos por 100.000 habitantes [2].

Vários eventos contribuíram para o atual panorama da TB no mundo: o aumento de casos de infecção por HIV [7], a deterioração das condições sócio-econômicas de parte da população mundial, alto índice de abandono do tratamento anti-tuberculose, o aparecimento da multi-resistência e a falta de interesse da comunidade científica e dos governos em desenvolver políticas públicas em relação à TB, ao não incentivar o desenvolvimento de novos instrumentos para o controle da TB [8]. A concentração dos casos, nas áreas urbanas, em particular nas Unidades Hospitalares e Prisões, locais de elevada concentração de pessoas e às vezes com prevalência de pacientes com co-morbidades, também propiciou um aumento do risco de transmissão da infecção e de adoecimento por TB.

Desde que a OMS, em 1993, declarou a tuberculose em estado de emergência mundial, o Brasil sinalizou, com marcos pontuais, sua posição frente às novas perspectivas do problema. O Ministério da Saúde lançou, em 1998, o Plano Emergencial para o controle da doença, recomendando a implantação da Estratégia do Tratamento Diretamente Observado (DOTS) para o controle da Tuberculose no Brasil. Dada a permanência do problema, com altas taxas de abandono do tratamento, com baixo percentual de cura e de detecção dos casos, em 1999 foi lançado o Plano Nacional de Controle da Tuberculose [3].

A estratégia DOTS tem como objetivo detectar 70% dos casos que apresentaram baciloscopia positiva, tratar corretamente 85% dos casos detectados e reduzir para 5% o abandono ao tratamento [1], sendo constituída de cinco componentes:

- Comprometimento governamental

  Colocando como prioridade entre as políticas publicas de saúde o desenvolvimento de sistemas de controle, monitoramento e treinamento em TB.

- Detecção de novos casos

Detecção de casos por baciloscopia entre os casos atendidos nas unidades de saúde.

- Tratamento padronizado

  Regime de tratamento diretamente observado por um profissional de saúde por no mínimo dois meses.

- Fornecimento regular de fármacos

- Registro das Informações

  Sistema de registro de informação que assegure a avaliação do tratamento.

Baseado no sucesso da estratégia DOTS, porém reconhecendo a necessidade de expansão do seu escopo, a OMS lançou em 2006 o programa Stop TB que da as diretrizes de implementação da estratégia DOTS entre outras. Os principais objetivos do Stop TB é conseguir, até 2015, detectar 84% dos casos de TB e conseguir tratar corretamente 87% dos casos detectados. Com isso se espera reduzir à metade as mortes por TB em 2015 em comparação ao ano de 1990 [9].

2.2 Fatores de Risco

A TB pulmonar está associada com a situação sócio-econômica, a desnutrição, má condições de higiene e saúde pública deficitária [10]. O risco de se contrair TB aumenta em decorrência do contato frequente com portadores da forma pulmonar da doença, presença em locais de grandes aglomerações ou de condições de vida insalubres e alimentação precária.

Pessoas com o sistema imunológico debilitado, como acontece com os portadores da Síndrome da Imunodeficiência Adquirida (SIDA), aquelas fazendo tratamento de quimioterapia e receptores de transplantes, que tomam medicamentos contra rejeição, apresentam risco maior de contrair a doença. Também são mais susceptíveis a doença crianças e idosos, indivíduos com má nutrição, profissionais de saúde, mineiros portadores de silicose, indivíduos dependentes do álcool, aqueles que fazem uso

de medicamentos como corticóides, portadores de outras doenças como o diabetes mellitus e neoplasias malignas[11].

2.3 Sinais e Sintomas

A doença costuma aparecer algumas semanas após a infecção primária, podendo, ainda, ficar latente durante anos antes de causar a doença. Se o sistema de defesa do organismo estiver em condições normais, na maioria dos casos, a bactéria não causará a doença, ficando latente. Se, em algum momento da vida, o sistema imunológico ficar debilitado, a bactéria que estava latente poderá entrar em atividade e vir a causar a doença, chamada de TB pós-primária. Mas, também há a possibilidade da doença se desenvolver no primeiro contato da pessoa com o bacilo, chamada TB primária.

No início, a TB pulmonar apresenta-se assintomática, isto é, sem sinais ou sintomas. Em geral, esses só aparecem quando a lesão torna-se visível em exames radiológicos, ou quando é possível se observar o bacilo infiltrado no pulmão. Com o agravamento da doença, surgem novos sintomas. Os principais sintomas e sinais da forma pulmonar são: tosse há mais de 15 dias, geralmente com expectoração, cansaço, febre vespertina, sudorese noturna, perda de apetite, emagrecimento, hemoptóicos (escarro com estrias de sangue) e hemoptise (escarro de sangue) [5] .

A evolução da tuberculose é muito variável, apresentando-se com sintomas leves ou ausentes. Frequentemente, na fase inicial da doença, a suspeita baseia-se nas avaliações obtidas mediante radiografias de tórax. Em adultos, a grande maioria dos casos de TB pulmonar inicia-se com uma lesão no pulmão.

2.4 Diagnóstico

O diagnóstico clínico é a identificação de uma doença por meio da anamnese e exame físico do paciente [12]. De um modo geral, o processo de diagnóstico pode ser visto como uma tarefa de decisão, que é realizada com base nos sinais, nos sintomas e

outros indícios laboratoriais. Frequentemente, esse processo envolve incertezas dos exames, devido às variações entre os pacientes, erros na observações dos sintomas entre outros motivos.

Os testes diagnósticos, quantitativos ou qualitativos, podem ser utilizados para a identificação de fatores de risco específicos, ou no processo de diagnóstico de uma doença. Portanto, eles podem úteis no tratamento do doente, avaliando a gravidade da doença, no estabelecimento do seu prognóstico e na monitorização da evolução clínica do paciente.

O teste de diagnóstico pode resultar em quatro possibilidades: o teste é positivo e o paciente tem a doença (deteção ou sensibilidade); o teste é positivo, porém o paciente não tem a doença (falso-positivo); o teste é negativo, mas o paciente tem a doença (falso-negativo) e por último o teste é negativo e o paciente não tem a doença (especificidade do teste).

Um teste muito sensível é útil para detectar a presença da doença em indivíduos doentes, podendo ser utilizado numa fase inicial de triagem dos pacientes. Por outro lado, um teste muito específico serve para excluir a presença da doença em indivíduos sadios, sendo útil na fase de diagnóstico, após a triagem dos pacientes.

O diagnóstico de TB pulmonar pode ser feito pela identificação dos sinais e sintomas, mas são necessários exames para se confirmar a presença do bacilo da TB. Para isso, os principais exames são: radiografia do tórax e exames de escarro (baciloscopia e cultura)

A baciloscopia do escarro é o método utilizado rotineiramente para a identificação do bacilo causador da TB por ser um exame simples, barato e de fácil execução. Entretanto, este exame tem baixa sensibilidade, em média de 60%, em pacientes com cultura positiva, não sendo capaz de discriminar a espécie da micobactéria [13].

Já a cultura para o bacilo da TB é um método mais sensível, pois detecta 70% a 89% dos casos, em média 80%, e permite a identificação da espécie da micobactéria, através de testes bioquímicos ou genéticos [14]. Porém, esse método é muito demorado, visto que o resultado da cultura fica disponível entre 15 a 60 dias após a

coleta do material respiratório. Portanto, nos pacientes com baciloscopia negativa no escarro, o diagnóstico da TB é geralmente tardio. Logo, o doente estará disseminando a micobactéria na comunidade e, em pacientes com HIV, este retardo pode ser fatal.

O exame radiológico do tórax é auxiliar no diagnóstico da tuberculose, justificando-se sua utilização, se possível, nos casos suspeitos. Este exame permite a identificação de pessoas portadoras de imagens sugestivas de tuberculose ou de outras enfermidades. O exame radiológico, em pacientes com baciloscopia positiva, tem como função principal a exclusão de outra doença pulmonar associada que necessite de tratamento concomitante, além de permitir avaliação da evolução radiológica dos pacientes [8].

Entretanto, cada vez mais se tem verificado que o diagnóstico de certas doenças, inclusive os diversos tipos de TB, pode ser auxiliado ou melhorado por meio da combinação de testes clínicos e modelos estatísticos. Quando formulados de uma forma sistemática e com uma base de dados consistente, esses modelos podem representar o problema clínico em questão, atuando como sistema de apoio ao diagnóstico. Ajudando os profissionais de saúde nas suas rotinas clinicas, assim como na administração de políticas públicas de saúde [15].

Hoje em dia, os modelos estatísticos mais utilizados no apoio ao diagnóstico são:

- Regressão Logística;
- Redes Bayesianas;
- Árvores de Decisão;
- Redes Neurais Artificiais.

A regressão logística é uma técnica bastante utilizada que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias [16].

Esta técnica é amplamente usada na área clínica para identificar os fatores de risco relacionados com uma doença de interesse. Entretanto, existem limitações no uso da regressão logística, principalmente na área médica, quando não se tem uma grande quantidade de dados para serem utilizados nos modelos [17].

Já as redes bayesianas consistem em uma forma gráfica de modelagem que representa um conjunto de variáveis aleatórias e suas dependências condicionais em forma de grafos, podendo, assim, representar a relação probabilística entre doenças e sintomas. Portanto, através das redes bayesianas, podem-se representar as relações de causa e efeito entre as variáveis em estudo [18], podendo ainda ser utilizada como ferramenta de apoio ao diagnóstico médico. Como, por exemplo, para diagnóstico de doenças do coração [19], câncer de ovários [20], pneumonia associada à ventilação mecânica [21], entre outras doenças. Entretanto, a especificação completa de uma rede bayesiana é uma tarefa complexa, uma vez que requer os valores de todas as probabilidades condicionais e as probabilidades a priori de todas as variáveis explicativas, o que dificulta o seu uso em aplicações com grande número de variáveis.

Outro método muito utilizado no apoio ao diagnóstico é o referente às árvores de decisão [22] [23] [24] [25]. Esta técnica visa particionar recursivamente um conjunto de dados, até que cada subconjunto do particionamento contenha casos iguais, o que forma um modelo graficamente estruturado, compacto e de forte apelo intuitivo para a classificação. Entretanto, problemas mais complicados tendem a necessitar de árvores de decisões grandes, na qual podem surgir sub-árvores idênticas em diferentes caminhos, além do fato de quanto maior o numero de decisões tem se tomar na árvore, mais nós a se percorrer, menos acurado serão os resultados obtidos.

O uso as redes neurais artificiais [26] já é uma realidade na área médica, se mostrando extremamente eficiente em inúmeras áreas da medicina, principalmente no auxilio ao diagnóstico [27], prognóstico e reconhecimento de padrões em sinais biológicos [28], exames clínicos e imagens médicas [29].

Os modelos estatísticos descritos acima podem auxiliar na triagem de pacientes, no apoio ao diagnóstico, na escolha terapêutica e no prognóstico, facilitando a prática

clínica.

Várias metodologias estatísticas, relacionadas ao diagnóstico da TB pulmonar, são sugeridas na literatura. A seguir, são apresentadas algumas dessas metodologias.

Bock et al. (1996) [30] utilizando um modelo logístico multivariado, identificaram que achados radiológicos no lóbulo superior dos pulmões, exposição ao bacilo, prova tuberculínica positiva e a não utilização da terapia preventiva com isoniazida são fatores associados ao diagnóstico de TB.

Samb et al. (1997) [31] identificaram quatro variáveis clínicas para o diagnóstico de tuberculose pulmonar paucibacilar, através de um modelo logístico multivariado. As quatro variáveis são: tosse por mais de 21 dias, dor torácica por mais de 15 dias, expectoração e dispnéia. O diagnóstico TB, usando duas das quatro variáveis, tem alta sensibilidade (85%) e baixa especificidade (67%). Quando utilizadas três das quatro variáveis, a especificidade aumenta levemente (86%), mas reduz a sensibilidade (49%).

El-Solh et al. (1999) [32] estabeleceram um modelo para identificação de TB pulmonar com uso das redes neurais artificiais. Diferentes variáveis foram incluídas no modelo, entre elas: idade, diabetes mellitus, SIDA, dor torácica, emagrecimento, tosse, sudorese, febre, dispnéia e achados radiográficos.

Kanaya et al. (2001) [33] criaram escores, baseado em regressão logística, para a predição de TB pulmonar paucibacilar usando variáveis clínicas como expectoração, resultado positivo na prova tuberculínica, achados radiológicos e HIV positivo. Aris et al. (1999) [34], através de um estudo prospectivo, propuseram um sistema de escores para discriminar TB pulmonar paucibacilar baseado na presença de resultado positivo na prova tuberculínica, derrame pleural, sarcoma de Kaposi, linfonodos aumentados na região cervical, linfonodo emaranhados e achados radiológicos no tórax.

Mello (2001) [8] desenvolveu um modelo de regressão logística multivariado e uma árvore de classificação, utilizando as informações sobre achados radiológicos, presença de escarro, emagrecimento e idade de pacientes atendidos em regime am-

bulatorial na Rede de Saúde do Município do Rio de Janeiro, com suspeita de tuberculose pulmonar paucibacilar. Obtendo 65,9% de sensibilidade e 60,1% de especificidade para o modelo de regressão logística e 64,2% de sensibilidade e 60,6% de especificidade para a árvore de classificação.

Santos (2003) [35] e Santos et al.(2007) [36] trazem um modelo de redes neurais artificiais para um conjunto de pacientes do Hospital Universitário Clementino Fraga Filho da Universidade Federal do Rio de Janeiro. A amostra tinha 59 casos com TB pulmonar e 77 casos sem TB, onde foram identificadas as características demográficas (sexo, idade e renda familiar), sintomas clínicos (tosse, febre, sudorese, emagrecimento, anorexia e outros) e fatores de riscos (diabetes, alcoolismo, SIDA e outros), totalizando 26 variáveis, tipicamente considerados relevantes e de fácil obtenção por anamnese para o diagnóstico da tuberculose. O modelo obteve uma especificidade de 83% para uma sensibilidade de 71% na classificação dos casos de TB paucibacilar.

Benfu et al. (2009) [37] desenvolveram um modelo neural para diagnóstico de TB paucibacilar que utiliza 29 variáveis no modelo, sendo estas relativas às informações pessoais (idade, sexo, estado civil, ocupação), histórico médico (alcoolismo, presença de doenças crônicas, febre, tosse, sudorese, emagrecimento e outros) achados radiográficos e exames laboratoriais. Num conjunto de pacientes com 291 casos de TB positivo e 298 casos de TB negativo, o modelo obteu especificidade de 100% e sensibilidade de 88,9%.

Ucar et al. (2010) [38] utilizam um modelo de redes bayesianas para diagnosticar a presença de TB pulmonar entre 503 pacientes utilizando **trinta variáveis clínicas** para a classificação.

Asha et al. (2010) [39] propõem o uso de *esemble methods* [40] para o diagnóstico da TB pulmonar, pois esse método melhora a acurácia da classificação devido a votação entre os classificadores. Para um conjunto de 250 pacientes, usando a idade, semanas de tosse, emagrecimento, febre, sudorese noturna, hemoptise, dor torácica, SIDA, achados radiológicos no tórax, exame de escarro e chiado no tórax

como variáveis clínicas, obtendo 80% de sensibilidade e 100% de especificidade para classificação da TB pulmonar e 100% de sensibilidade e 80% de especificidade nos casos de TB pulmonar retroviral, quando o paciente é HIV positivo, com o método de AdaBoost [41] .

Os modelos foram formulados com diferentes metodologias e aplicados em diferentes populações. Além disso, os modelos formulados incluem diferentes variáveis, sinalizando que cada população demanda a elaboração e a validação de modelos apropriados às suas características sócio-epidemiológicas.

Segundo as diretrizes da estratégia DOTS, o exame prioritário para os casos suspeitos de TB pulmonar é a pesquisa direta do bacilo álcool-ácido resistente (BAAR) em amostras de escarro espontâneo, devido ao baixo custo dessa técnica. Contudo, em pacientes com suspeita de TB pulmonar, apresentando sintomas respiratórios e com achados radiográficos compatíveis com TB pulmonar, cuja pesquisa do BAAR falham em revelar a micobactéria, ou quando não se obtém escarro, surge um problema de difícil solução. O médico se depara com a decisão de iniciar o tratamento de prova anti-TB, ou utilizar técnicas mais invasivas para documentar a TB e excluir outras enfermidades, ou ainda, permite-se aguardar por 15 a 60 dias os resultados de culturas, disponíveis em poucos centros [35]. Devido a esses fatores, vemos o crescente número de sistemas de apoio a decisão no diagnóstico da TB paucibacilar.

Os modelos que apresentaram resultados melhores necessitam de poder computacional para fazer a classificação, já que os mesmos usam não linearidades, como as redes neurais e o AdaBoost, ou necessitem de contas que não são facilmente contabilizadas, como a regressão logística. Já nos modelos de escore explicitados, idéia semelhante a ser desenvolvida neste trabalho, se fez presente o uso de achados radiológicos e exames laboratoriais como o teste de escarro (BAAR) e a prova tuberculínica (PPD), exames estes que não serão utilizados durante o desenvolvimento dos escores para detecção da TB pulmonar deste trabalho.

Convém ressaltar que devido aos recentes avanços da engenharia genética, a possibilidade do diagnóstico de TB se basear na técnica de Reação de Polimerase em

Cadeia (PCR), onde identifica-se a existência de sequencias do gene de *Mycobacterium tuberculosis* numa amostra de escarro. Uma dessas técnicas é chamada de GeneXpert, onde este método tem a vantagem de alta sensibilidade, para os casos com teste de escarro positivo e negativo, 98% e 72%, respectivamente, alta especificidade, 99,2%, e o resultado final do teste fica pronta em 90 minutos [42]. Entretanto, esse exame ainda tem um custo muito elevado, inviabilizando o seu uso como um exame de rotina, tornando seu uso restrito a alguns centros de pesquisa e hospitais referências.

Capítulo 3

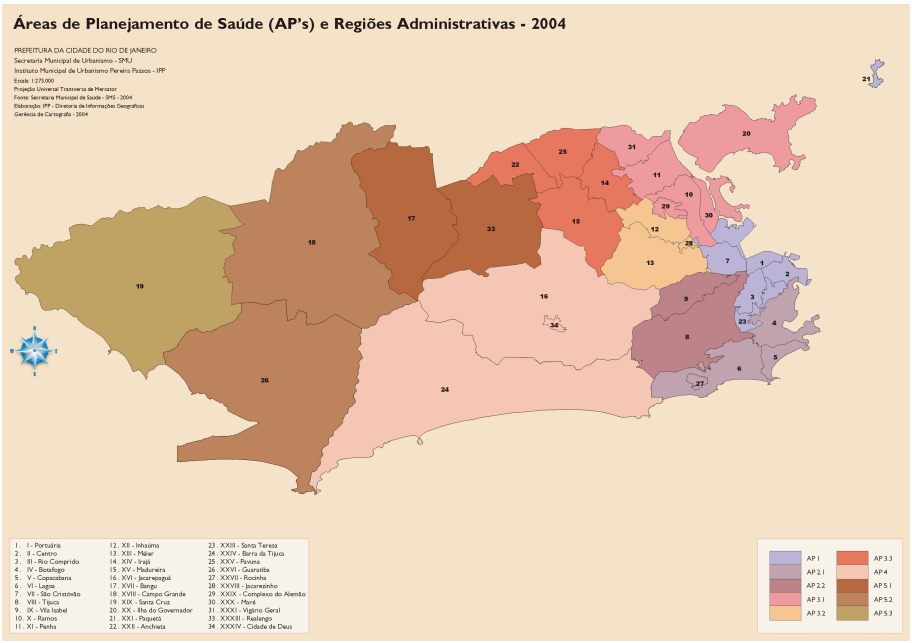
Base de Dados

O banco de dados a ser utilizado neste trabalho refere-se as informações clinícas dos pacientes atendidos no setor de Pneumologia da Policlínica Augusto Amaral Peixoto (PAAP) na Área Programática (AP)3.3 da Secretaria Municipal de Saúde do Rio de Janeiro (SMS-RJ), no período compreendido entre 26/09/06 a 31/07/07.

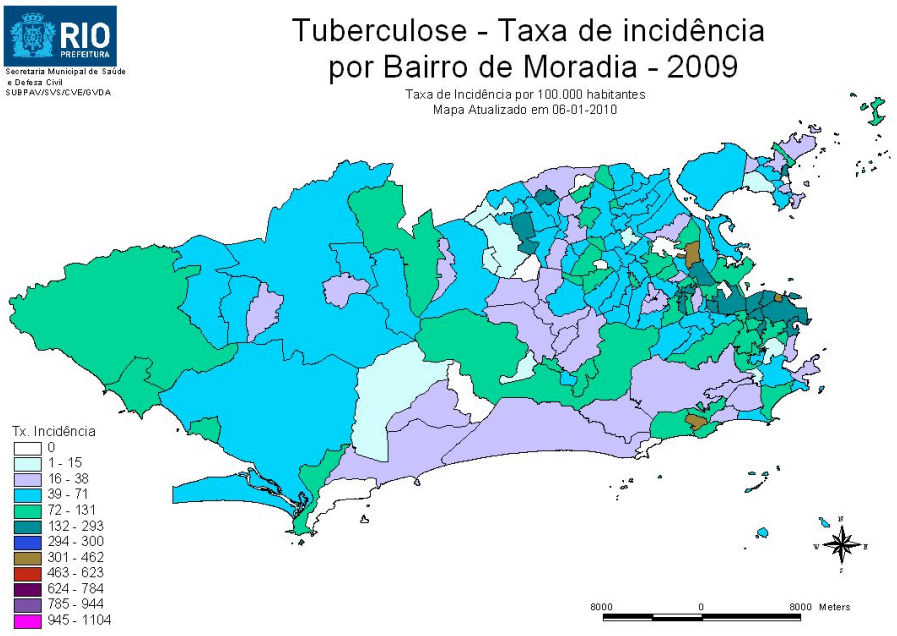
A Policlínica Augusto do Amaral Peixoto está localizada na AP3.3 da cidade do Rio de Janeiro, onde residem cerca de 1.100 mil habitantes. Esta unidade é responsável pela assistência médica ambulatorial dos bairros de Guadalupe, Deodoro, Costa Barros, Pavuna, Acari, Anchieta, Ricardo de Albuquerque, onde residem cerca de 330 mil habitantes. Na AP 3.3, nos anos de 2006 e 2007, a incidência de casos de TB notificados foi de 78,3 e 80,2 por 100.000 habitantes, respectivamente [2]. Na figura 3.1, podemos ver o panorama geral da taxa de incidência da TB no município do Rio de Janeiro e suas respectivas áreas programáticas.

A população de referência foi formada por pacientes que procuraram pela primeira vez atendimento no setor de pneumologia da PAAP, ou que retornaram após abandono de tratamento de TB no período do estudo. Foram coletadas informações clínicas com os pacientes sobre: os sinais e sintomas da doença, fatores de risco para tuberculose, história de tuberculose e de outras doenças prévias e dados sócio-demográficos, através de questionário padrão.

O diagnóstico de TB pulmonar ativa foi estabelecido nas seguintes situações [8]:



(a)



(b)

Figura 3.1: (a)Áreas Programática do Município do Rio de Janeiro. (b)Estimação da taxa de incidência de TB no município do Rio de Janeiro. Fonte Secretária Municipal de Saúde e Defesa Civil do Rio de Janeiro [2]

- Presença de baciloscopia positiva, associada à resposta clínica ao tratamento anti-TB instituído;

- Cultura para micobactérias positiva, associada à resposta clínica ao tratamento anti-TB instituído;

- Pacientes com resultados negativos à baciloscopia e cultura para a micobactéria e com melhora clínico-radiológica nos 2 meses iniciais de tratamento anti-TB, adotado de forma empírica, desde que não houvesse o emprego de outra terapêutica associada que justificasse uma resposta favorável ao tratamento.

Para esse trabalho, foram utilizados os dados de 1.144 pacientes submetidos a entrevista de triagem da enfermagem do setor de pneumologia da PAAP/SMS-RJ que aceitaram em participar do estudo. A amostra em estudo foi gentilmente cedida pelo Programa Acadêmico de TB da Faculdade de Medicina da Universidade Federal do Rio de Janeiro, onde há uma proporção igualitária entre homens e mulheres, a média de idade é aproximadamente de 46 anos e 21% dos pacientes tiveram diagnóstico final de TB pulmonar.

O projeto foi aprovado pela Comissão de Ética em Pesquisa (CEP) do Hospital Universitário Clementino Fraga Filho/Universidade Federal do Rio de Janeiro (HUCFF/UFRJ) com parecer número 067/06 em 24/04/06 (Apêndice A).

Os dados obtidos nos questionários foram armazenados de forma confidencial e somente o pessoal autorizado do grupo de pesquisa do Programa Acadêmico de Tuberculose da Faculdade de Medicina da UFRJ teve acesso a estas informações. Todos os pacientes que participaram da pesquisa concordaram e assinaram o Termo de Consentimento Livre e Esclarecido (TCLE) (Apêndice B).

Para essa mesma amostra de pacientes foi elaborado um questionário de triagem que permitiu a elaboração de um escore clínico, que foi elaborado por um painel de especialistas em pneumologia [43], e será a referência de desempenho deste trabalho. O escore clínico foi baseado nos sintomas mais relevantes à TB pulmonar ,identificados pela análise multivariada através de redes neurais nos pacientes atendidos

no Hospital Universitário Clementino Fraga Filho no Rio de Janeiro [35] [44] [45]. Sendo este formado por um conjunto de sintomas no qual o paciente não sofresse constrangimentos numa entrevista de triagem junto a equipe de enfermagem e que fosse de fácil contabilização. Portanto, o escore, proposto por especialistas, é formado com uma pontuação para os sintomas variando de 0 a 6 conforme mostrado na tabela 3.1.

SINTOMAS	PONTOS	
	SIM	NÃO
Sexo masculino	1	0
Idade até 59 anos	1	0
Dor torácica	2	0
Tosse até 2 semanas	1	0
Tosse $>$ 2 semanas	2	0
Expectoração até 2 semanas	1	0
Expectoração $>$ 2 semanas	2	0
Hemoptise	6	0
Sudorese noturna	2	0
Febre	2	0
Emagrecimento	2	0
Total	Suspeito de TB $\geq$ 8 pontos	Não TB $<$ 8 pontos

Tabela 3.1: Escore de Referência

A premissa de que os escores desenvolvidos neste trabalho devem ser computados sem o auxilio de poder computacional veio do escore de referência que hoje em dia é utilizado na triagem de pacientes no Posto de Saúde Augusto do Amaral Peixoto. Estando a sua respectiva curva ROC representada na figura 3.2. No ponto de corte escolhido pelo painel de especialista para a triagem de pacientes o escore obtêm sensibilidade de 83,06% e especificidade de 52,00%.

Ao analisarmos o histograma da saída do escore, figura 3.3, podemos ver que as classes são sobrepostas mostrando que a tarefa de classificação dos casos com TB é muito difícil, por isso de uma baixa especificidade para sensibilidade obtida, justificando o desenvolvimento de outros escores com diferentes variáveis a serem consideradas para a classificação dos casos com TB pulmonar.

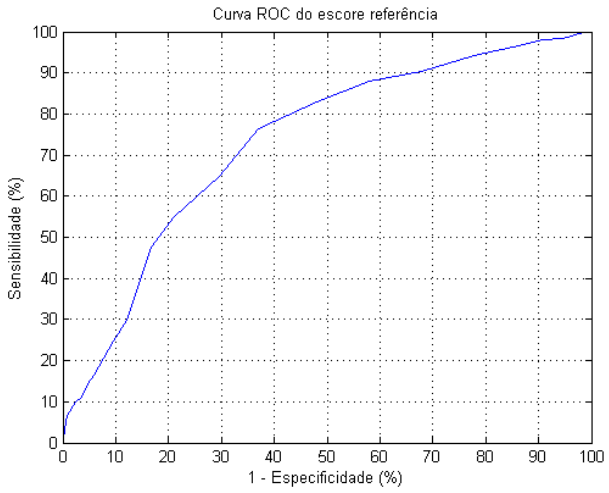
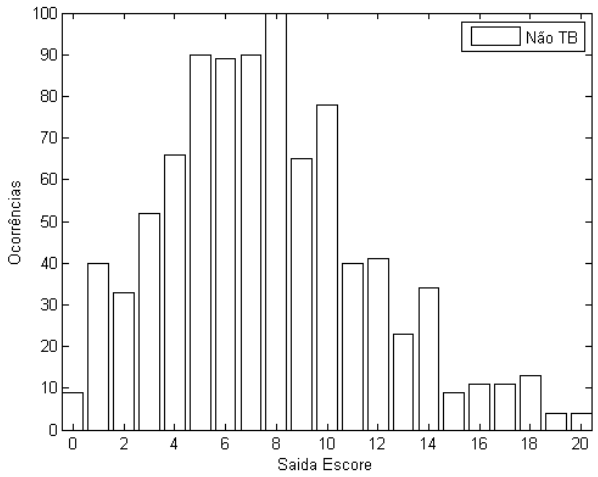
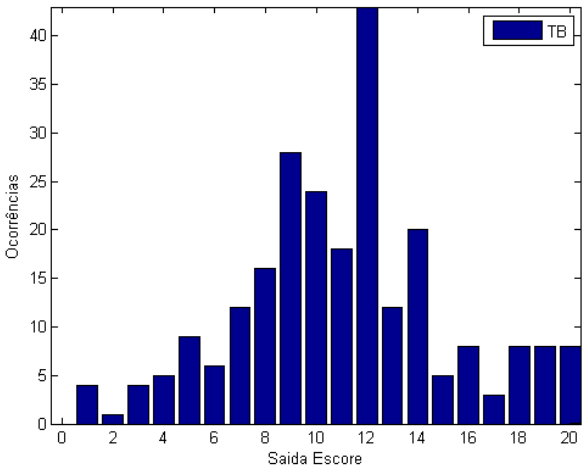


Figura 3.2: Curva ROC do escore de referência



(a) Pacientes sem TB



(b) Pacientes com TB

Figura 3.3: Histogramas da saída do escore de referência

3.1 Variáveis Utilizadas

As variáveis utilizadas neste trabalho, descritas na tabela 3.2, foram escolhidas de acordo com a dependência com a TB pulmonar, tendo sido identificadas da mesma forma que o trabalho de referência, e também foi levado em consideração a opinião de dois pneumologistas sobre variáveis relevantes, segundo a literatura médica, para o diagnóstico da TB pulmonar.

Variável	Codificação
Idade	anos
Tosse Hemoptise Sudorese Febre Emagrecimento Dispnéia Tabagismo Internação hospitalar Dor Torácica SIDA	ausência = -1, presença = 1, ignorado = 0
Sexo	homem = -1, mulher =1

Tabela 3.2: Sinais e Sintomas utilizados

Capítulo 4

Metodologia

Como o objetivo deste trabalho é desenvolver um escore para o auxilio na triagem e no diagnóstico da tuberculose pulmonar, se faz necessária uma análise prévia dos casos contidos na base de dados. Uma das formas de se realizar essa investigação inicial é explorar a distribuição dos sintomas para ver se existem variáveis com algum viés.

Após o estudo das distribuições das variáveis, os casos da base de dados passam por um processo de clusterização que buscará pela existência de agrupamentos naturais nos dados, ajudando num melhor entendimento da informação produzida por esses agrupamentos.

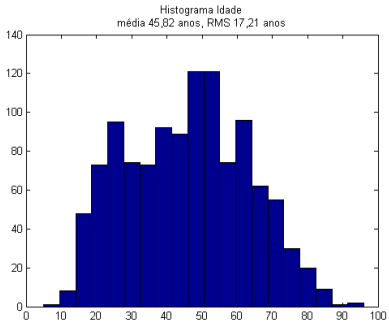
Por fim, serão discutidos os métodos de desenvolvimento do escore clínico de forma que este seja simples, de fácil entendimento e que não necessite de recursos computacionais para seu cálculo.

4.1 Análise dos sintomas e condições

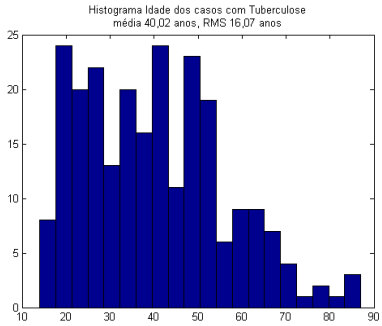
A base de dados em estudo é composta de 1.144 pacientes, sendo 242 pacientes, com TB pulmonar ativa. Foram observadas 12 variáveis explicativas, incluindo uma variável numérica e variáveis qualitativas, conforme mostra a tabela 3.2.

Para a variável com codificação numérica, idade, foram feitos os histogramas, apresentados na figura 4.1. A mediana da idade da população em estudo é de 47

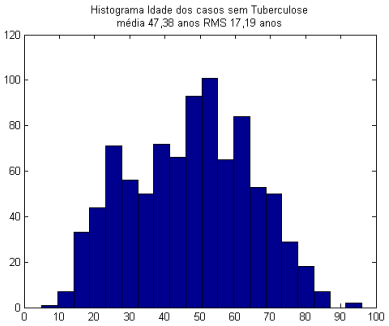
anos, os casos sem TB seguem a tendência de idade população, com mediana da idade de 48 anos. Porém, podemos ver uma predominância de pessoas um pouco mais jovens que foram diagnosticadas com tuberculose, uma vez que a mediana da idade desses casos é de 39 anos.



(a)



(b)



(c)

Figura 4.1: (a)Histograma da idade da população em estudo (b)Histograma da idade dos casos com diagnóstico de TB (c)Histograma da idade dos casos sem TB

Já para as variáveis dicotômicas, foi verificada a quantidade de ignorados presentes na base de dados e foi feito um estudo retrospectivo, calculando-se a razão de chance [46], do inglês *odds ratio*, das mesmas, conforme a tabela 4.1.

O *odds ratio* (OR) é uma forma simples e de fácil interpretação do relacionamento de um sintoma com uma doença.Uma razão de chances de 1 indica que o sintoma sob estudo é igualmente provável de ocorrer nos dois grupos, com ou sem TB. Uma razão de chances maior do que 1 indica que a condição ou evento tem maior probabilidade de ocorrer nos casos com TB. Finalmente, uma razão de chances menor do que 1 indica que a probabilidade é menor nos casos com TB do que nos sem TB.

Ao analisarmos a tabela 4.1, podemos observar que as quatro variáveis que apre-

	Tosse			Hemoptise			Sudorese Noturna		
	OR = 3,71			OR = 2,32			**OR = 2,72**		
	Não	Sim	Ignorado	Não	Sim	Ignorado	Não	Sim	Ignorado
TB -	157	745	0	831	71	0	623	279	0
TB +	13	229	0	202	40	0	109	133	0
	Febre			Emagrecimento			Dispnéia		
	OR = 3,56			**OR = 4,14**			OR = 0,79		
	Não	Sim	Ignorado	Não	Sim	Ignorado	Não	Sim	Ignorado
TB -	567	335	0	729	173	0	302	600	0
TB +	78	164	0	122	120	0	94	148	0
	Tabagismo			Internação Hospitalar			Sexo		
	OR = 1,13			OR = 1,01			OR = 0,47		
	Não	Sim	Ignorado	Não	Sim	Ignorado	Homens	Mulheres	Ignorado
TB -	373	513	16	756	122	24	419	483	0
TB +	93	145	4	202	33	7	157	85	0
	Dor Torácica			SIDA					
	OR = 1,23			OR = 0,49					
	Não	Sim	Ignorado	Não	Sim	Ignorado			
TB -	323	579	0	24	6	872			
TB +	76	166	0	82	10	150			

Tabela 4.1: Distribuição das variáveis dicotômicas

sentaram os maiores *odds ratio* na amostra de população em estudo vai de acordo com os sinais e sintomas que descrevem a TB pulmonar explicitados na seção 2.3. Portanto, baseado na análise dos *odds ratio* podemos inferir que pacientes que apresentem Tosse, Febre, Emagrecimento e Sudorese Noturna terão mais chances de serem TB positivos que outros pacientes.

Também podemos observar que somente três variáveis apresentam casos com respostas ignoradas. Esse tipo de resposta pode acontecer pelo fato de o paciente se sentir constrangido de responder a pergunta na frente de outras pessoas (perguntas sobre tabagismo e de cunho sexuais) ou por não se lembrarem, ou não saber em definir ao certo, se já sofreram internação hospitalar.

A variável SIDA apresenta 89% de ignorados e, pela sua distribuição, assim como pelo seu OR de 0,49, sugere que o paciente portador do virus HIV seja menos propenso a contração de tuberculose, que vai contra as crescentes taxas de co-infecção TB/HIV [8]. Fazendo que essa variavel tenha um vies estatístico, uma vez que pacientes que não forem HIV positivo tenderam a ser classificados como TB negativo.

Portanto, ela será descartada no desenvolvimento do escore proposto neste trabalho. Entretanto, para os sintomas de Tabagismo e Internação Hospitalar o baixo percentual de ignorados, 1,8% e 2,7% respectivamente, na base de dados não acarretará problemas futuros no desenvolvimento do escore.

Após a definição das variáveis a serem utilizadas, a seguir será abordada a construção dos agrupamentos, de forma a ajudar o entendimento do problema e verificar se existem agrupamentos naturais dos pacientes que auxiliam a distinguir TB e não TB.

4.2 Agrupamento dos dados

De forma instintiva, os seres humanos tendem a visualizar conjuntos em grupos discretos. É uma das formas mais naturais e rápidas de impor alguma ordem a um grande volume de objetos apresentado sem maiores informações, de modo a produzir um entendimento direto por parte do observador. Porém, perceber padrões e agrupar objetos tendo por origem uma base de dados numérica não é tarefa simples, ou que possa ser feita manualmente em tempo hábil. São necessários algoritmos de segmentação, especializados na descoberta e formação de grupos de dados.

Agrupamento de dados, ou clusterização, [47] [48] é uma área de pesquisa fundamental em análise de dados. De uma forma genérica, o objetivo da clusterização é a descoberta de estruturas implícitas em um conjunto de dados, denominados de *clusters*. Esses *clusters* contêm exemplos de dados similares entre si, de acordo com alguma métrica de semelhança pré-definida.

Na grande maioria dos casos, para variáveis reais, esta similaridade pode ser medida com base na distância geométrica. Portanto, acredita-se que as similaridades existentes entre os dados sejam apropriadamente representadas no espaço dos vetores de características. Logo, quão mais similares dois indivíduos de uma população mais próximos, segundo uma métrica, seriam seus vetores de características, podendo ser geometricamente próximos quando utilizado uma métrica geométrica, ou estatisticamente próximos quando utilizado uma métrica divergente. A avalia-

ção de similaridade exige, portanto, medir a distância ou a divergência entre dois vetores. Um critério de distância bastante geral, quando se utilizado uma métrica geométrica, deve-se a Minkowski [49], o qual, para dois vetores $\mathbf{x}$ e $\mathbf{y}$ de dimensão l, é definido como:

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \qquad \in \Re^l \tag{4.1}$$

onde x_i e y_i são componentes dos vetores $\mathbf{x}$ e $\mathbf{y}$, respectivamente, e p é um número inteiro qualquer maior ou igual a 1. Quando $p = 2$, temos a distância euclidiana, comumente utilizada como métrica de similaridade. Também existem métricas que consideram pesos diferentes para cada par de componentes, como por exemplo, a distância euclidiana ponderada e a distância de Mahalanobis [50].

No contexto da inteligência computacional, as técnicas de agrupamento são classificadas como métodos de aprendizado não-supervisionado [26], uma vez que não há conhecimento prévio sobre as classes dos dados, apenas sobre seus atributos. Existem diversas técnicas de clusterização disponíveis na literatura, cada qual explorando uma estratégia particular. Entretanto, essa escolha deve ser feita de forma cuidadosa, de modo que os agrupamentos formados, ao final do processo, reflitam a real estrutura real dos dados.

Essas técnicas podem ser divididas de acordo com vários critérios. Como por exemplo, os algoritmos hierárquicos e os particionais [49], sendo somente os últimos utilizados neste trabalho.

Os métodos de agrupamento particionais utilizam o conceito de protótipos de *clusters*, que são pontos no espaço multidimensional dos dados que representam o centro de cada um dos *clusters*. Os protótipos podem ser representados por meio de centróides, onde estes podem assumir qualquer posição no espaço dos dados, ou por meio de medóides, onde os protótipos fazem parte, necessariamente, do conjunto de dados.

Entre os métodos particionais, o que melhor exemplifica o método é o algoritmo k-means [51]. Este algoritmo, que é amplamente utilizado pela comunidade

cientifica, emprega o conceito de centróides . Dados os N centróides espalhados aleatoriamente no espaço dos dados, sendo N o número de *clusters* pré-definidos, o algoritmo particiona os eventos nos N *clusters*, de acordo com a distância entre o evento e o centróide, formando um diagrama de Voronoi [47]. De uma forma geral utiliza-se no k-means como métrica a distância euclidiana quadrática (eq. 4.2)

$$d_{ki}^2 = \|x_k - c_i\|^2 \tag{4.2}$$

onde x_k são os exemplos do conjunto de dados e c_i são os centróides dos *clusters*

Em seguida, os centróides são recalculados como o baricentro dos eventos associados aos seus respectivos *clusters*, logo, redefinindo o diagrama de Voronoi. Esse processo é repetido até que os centróides não mudem mais ou que um determinado número de iterações no algoritmo seja atingido.

4.2.1 Mapas Auto-Organizáveis

Como, neste trabalho, a clusterização destina-se ao entendimento dos sintomas e fatores associados da tuberculose e da busca de associações entre os pacientes atendidos e os *clusters* formados, buscam-se agrupamentos que representem grupos de baixo, médio e alto risco do paciente estar ou não com tuberculose, análogo ao agrupamento dos pacientes atendidos com suspeita de tuberculose paucibacilar utilizando redes ART [49], no Hospital Universitário Clementino Fraga Filho [45].

Os Mapas Auto-Organizáveis representam um tipo de rede neural artificial que, assim como a rede ART, utiliza o conceito de treinamento não supervisionado para produzir um mapeamento não-linear do espaço de entrada, numa representação discreta de baixa dimensionalidade (geralmente bi-dimensional), possibilitando uma análise mais lúdica dos resultados obtidos, já que os mesmos podem ser representados de forma gráfica e pictórica.

O SOM, do inglês *Self-Organizing Map*, tem o diferencial, em relação às outras formas de redes neurais por usar uma função de vizinhança que preserva as propri-

edades topológicas do espaço de entrada, o que faz desta técnica uma forma muito útil para visualização em baixa-dimensão de dados e de ampla utilização [52].O mapa, em si, é formado por neurônios que estão conectados entre si, com uma forma regular,um grid, como mostrado na figura 4.2.

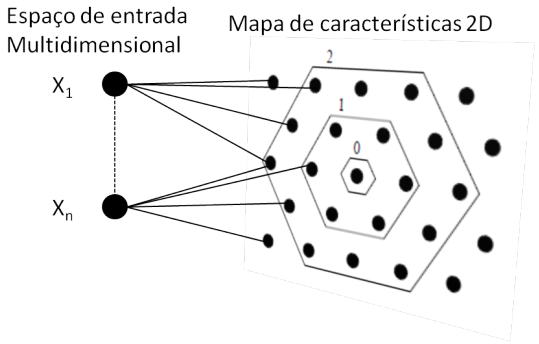


Figura 4.2: Diagrama de um mapa auto-organizável

O objetivo do aprendizado do SOM é fazer com que diferentes partes do mapa respondam de forma semelhante a certos padrões de entrada. Isto é em parte motivado pela forma de como as informações visuais, auditivas e sensoriais são tratadas em partes separadas do córtex cerebral no cérebro humano [26].

O processo de aprendizado é dividido em três partes: competição, cooperação e adaptação. O treinamento utiliza o aprendizado competitivo (*winner takes all*); ou seja, para cada vetor de entrada há apenas um neurônio vencedor, o que no contexto do SOM, é chamado de **BMU** (*Best Matching Unit*). No entorno do **BMU**, haverá uma cooperação topológica de neurônios, que serão excitados conforme uma função de vizinhança. Por fim, os pesos sinápticos do neurônio vencedor e de seus vizinhos são adaptados conforme o padrão de entrada.

Considerando os vetores de entrada $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, e como todos os neurônios são totalmente conectados com as entradas, os pesos sinapticos dos neurônios podem ser definidos por $\mathbf{w}_i = [w_{1i}, w_{2i}, \ldots, w_{ni}]^T$. A atualização do vetor de pesos do **BMU** é feita sequencialmente, através da equação 4.3:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \eta(t)h_{ij}(t)(\mathbf{x}(t) - \mathbf{w}_i(t)) \tag{4.3}$$

onde $\eta(t)$ é uma taxa de aprendizagem monotonicamente decrescente e $hij(t)$ é a função de vizinhança, que é escolhida de forma a ter seu valor máximo no **BMU** decrescendo à medida que se afasta dele e tendo uma largura (número de neurônios abrangidos por ela) que decresça com o tempo.

Os neurônios que formam o mapa também podem ser treinados utilizando o conceito de batelada [53]. O treinamento em batelada também é iterativo; porém, ao invés de usar um vetor de dados por vez, um conjunto de vetores da base de dados (eventualmente toda a base de dados) é apresentada ao mapa antes dos pesos serem ajustados.

A cada passo de iteração, a base de dados é dividida conforme o diagrama de Voronoi formado pelos neurônios do mapa. Assim, os vetores da base de dados são associados ao seu **BMU**, fazendo com que cada neurônio do mapa tenha um conjunto dos vetores da base de dados mais similares. Para cada neurônio, é calculado o somatório desse conjunto:

$$\mathbf{s}_i(t) = \sum_{j=1}^{n_{Vi}} \mathbf{x}_j$$

onde n_{Vi} é o número de eventos de cada conjunto do diagrama de Voronoi da unidade i. Após a essa etapa, os pesos sinápticos dos neurônios do mapa são atualizados conforme a equação 4.4:

$$\mathbf{w}_i(t+1) = \frac{\sum_{j=1}^{m} h_{ij}(t)\mathbf{s}_j}{\sum_{j=1}^{m} n_{Vi} h_{ij}(t)} \tag{4.4}$$

sendo m o número de neurônios utilizados no mapa. Portanto, no treinamento por batelada, os pesos sinápticos dos neurônios são médias ponderadas, obtidas pela função de vizinhança dos eventos contidos na base de dados.

4.2.1.1 Parâmetros do treinamento

No SOM, o número de neurônios e sua relação topológica são parâmetros que devem ser configurados antes do mapa ser treinado. Existem quatro escolhas que devem ser decididas a priori: o número de neurônios, a dimensão, o formato e o tipo de

treliça do mapa.

O número de neurônios, que define a granularidade do mapa, deve ser o maior possível, deixando que o tamanho da vizinhança controle a suavização e generalização do mapa. Entretanto, um mapa muito grande pode prejudicar o treinamento, pois podem existir muitos neurônios que nunca são ativados, e que torna o treinamento uma tarefa computacional muito pesada.

Existem regras heurísticas para o cálculo do número de neurônios e as dimensões do mapa, que levam em consideração o número de eventos a serem utilizados e a relação entre os autovalores da primeira e segunda componentes principais da base de treinamento [52].Porém, como a base de dados utilizada é composta por variáveis numéricas e dicotômicas, a estimação da Análise de Componentes Principais (PCA) pode ser falha [54], inviabilizando o uso desta técnica na estimação do número de neurônios e da forma do mapa. Portanto, baseando-se em experiências anteriores com a técnica clusterização por SOM, optou-se pela a escolha de um mapa e que tivesse um número de neurônios suficiente para formar um *codebook* que pudesse representar os possíveis padrões dos casos que poderiam existir entre os pacientes atendidos.

A função de vizinhança determina o quão forte é a ligação entre neurônios. A função de vizinhança pode ser definida de formas diferentes. Desde funções mais simples, como um degrau com valores constantes na vizinhança do neurônio vencedor **BMU**, até funções baseadas em distribuições gaussianas,

$$h_{ij}(t) = \exp\left(-d_{ij}^2/2\sigma^2(t)\right) \tag{4.5}$$

onde d_{ij} é a distância euclidiana entre o neurônio j e o **BMU**, e $\sigma(t)$ é a largura da vizinhança dos neurônios na iteração t. No mapa contido na figura 4.2, estão representadas larguras de vizinhança igual a 0, 1 e 2.

O número de neurônios e a função de vizinhança determinam a granularidade do mapa resultante. Quanto maior a área em que a função de vizinhança tem valores significativos, mais rígido será o mapa. Um mapa maior o torna mais flexível,

entretando mais sujeito a ter neurônios não ativados. Essa relação determina a acurácia e a habilidade de generalização do SOM.

Portanto neste trabalho, o mapa a ser treinado terá os seguintes parâmetros:

- Mapa bidimensional com 64 neurônios, com dimensões 8x8 neurônios
- Função de vizinhança gaussiana descrita na equação 4.5
- Treliça hexagonal (que faz a distância, no mapa, entre neurônios vizinhos seja sempre igual)
- Treinamento por batelada

Antes do treinamento do SOM, a idade dos pacientes foi normalizadas para ter variância unitária. Após essa etapa de pré-processamento, os pesos sinápticos dos neurônios do mapa são inicializados de forma aleatória, com os pesos escolhidos a partir de uma distribuição uniforme com valores máximos e mínimos dados pelos respectivos valores da base de dados pré-processada.

O treinamento foi dividido em duas partes, que utilizaram o treinamento por batelada. A primeira parte realiza um treinamento mais abrangente, com uma largura de vizinhança maior ($\sigma = 2$). Já a segunda parte, utiliza o mapa treinado pela fase anterior como condição inicial de treinamento e reduz a largura de vizinhança para $\sigma = 1$, suavizando, portanto, o mapa.

4.2.1.2 Clusterização por SOM

Existem duas principais formas de interpretar o SOM. A primeira é considerar que os pesos sinápticos dos neurônios, enquanto ponteiros para o espaço de entrada, formam, portanto, uma aproximação da distribuição dos eventos utilizados no treinamento. Mais neurônios irão apontar para regiões com alta concentração dos dados e menos para regiões onde há poucos eventos.

A outra forma de se interpretar vem do fato de que, durante o treinamento, os neurônios de certa vizinhança irão se mover para uma mesma direção, pois eventos similares tendem a ativar neurônios adjacentes. Portanto, o SOM forma um mapa

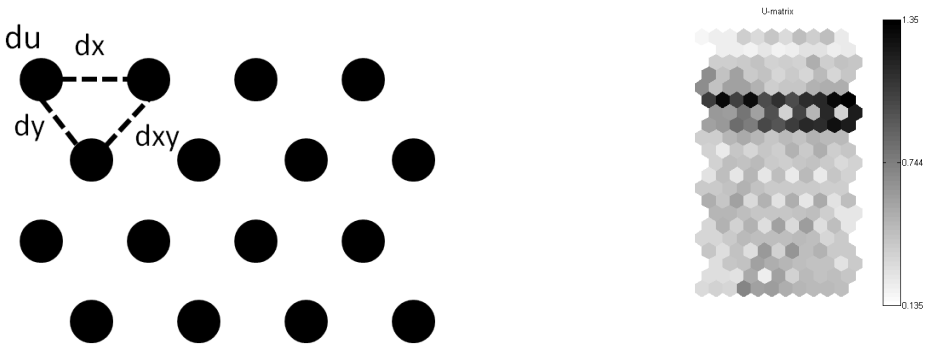
semântico onde eventos semelhantes são mapeados conjuntamente e os dissimilares são separados. Esse mapeamento pode ser visualizado através da U-Matrix do SOM [55].

A idéia básica da U-matrix, equação 4.6, é usar a mesma métrica utilizada durante o treinamento para calcular distâncias entre pesos dos neurônios vizinhos. O resultado é uma matriz que pode ser interpretada como uma imagem, na qual as coordenadas de cada pixel (x, y) são derivadas das coordenadas dos neurônios no grid do mapa, e a intensidade de cada pixel na imagem f(x, y) corresponde a uma distância calculada. Um mapa bidimensional de tamanho $N \times M$ irá gerar uma imagem$(2N-1) \times (2M-1)$ pixels, onde $du(x, y)$ é o valor médio dos pesos sinápticos dos neurônios vizinhos.

$$\begin{bmatrix} du(0,0) & dx(0,0) & du(1,0) & \dots & du(N-1,0) \\ dy(0,0) & dxy(0,0) & dy(1,0) & \dots & dy(N-1,0) \\ du(0,1) & dx(0,1) & du(1,1) & \dots & du(N-1,1) \\ dy(0,1) & dxy(0,1) & dy(1,1) & \dots & dy(N-1,1) \\ \dots & \dots & \dots & \dots & \dots \\ du(0,M-1) & dx(0,M-1) & du(1,M-1) & \dots & du(N-1,M-1) \end{bmatrix} \tag{4.6}$$

Pode-se abstrair vales e montanhas, os primeiros correspondendo a regiões de neurônios similares, enquanto que montanhas refletem a dissimilaridade entre neurônios vizinhos e podem ser associadas a regiões de fronteiras de agrupamentos [56]. Tornando a U-Matrix uma ferramenta que pode ser facilmente analisada na procura por agrupamentos, tanto de forma matemática ou visualmente, como pode ser visto na figura 4.3.

A tarefa de descobrir os *clusters* formados pelo SOM pode ser feita de forma visual, através da projeção do mapa por meio da U-matrix e nos planos das componentes. Por exemplo, analisando a figura 4.3 (b), como a distância entre os neurônios está representado pela escala de intensidade de cinza, podemos inferir que a U-Matrix



(a) (b)

Figura 4.3: (a)Exemplo do calculo da distancia da U-Matrix (b)Exemplo da figura da U-Matrix

representada tem 2 dois *clusters* distintos, uma vez que existe um faixa mais escura entre duas regiões mais claras, representando a separação dos mesmos.Entretanto, a inspeção visual pode ser tornar difícil caso não existam agrupamentos claros na U-matrix.

De qualquer modo, a inspeção visual só pode ser utilizada para uma análise qualitativa. Para produzir descrições quantitativas dos dados, devem ser selecionados grupos de interesse dentro do mapa. Então, ao se utilizar das informações geradas pelo SOM e por outros métodos de clusterização, pode-se ter uma boa idéia dos agrupamentos presentes na base de dados, de uma forma rápida, robusta e com uma visualização dos resultados eficiente.

A clusterização do SOM se dá após o treinamento do mapa, com os neurônios do mapa sendo agrupados por outros métodos de clusterização . O uso desta técnica reduz o custo computacional para clusterizacao de base de dados volumosas e também reduz o número de eventos atípicos que possam impactar na clusterização, pois agora os eventos são os neurônios do mapa [57].

Neste trabalho, o mapa gerado busca encontrar grupos de pacientes de baixo, médio e alto risco para TB. Logo, o mapa foi clusterizado usando o algoritmo de k-means buscando a melhor forma de agrupar o mapa em 3 grupos. Para isso, o k-means foi inicializado 5 vezes e foi escolhida a melhor separação dos grupos, através do índice de Davies-Bouldin [58] dos agrupamentos gerados em cada inicialização.

O índice Davies-Bouldin é uma medida da similaridade entre agrupamentos, independente do número de agrupamentos e do método de partição dos dados utilizada, o que o torna indicado para a avaliação dos *clusters* formados. O índice é dado pela equação 4.7:

$$I_{DB} = \frac{1}{C} \sum_{k=1}^{C} \max_{l \neq k} \left\{ \frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Ql)} \right\} \tag{4.7}$$

onde Q é um *cluster*, C é o número de *clusters* e S_c, equação 4.8, é uma medida de similaridade intra *cluster*:

$$S_c = \frac{1}{N_k} \sum_{i=1}^{N_k} \|x_i - c_k\| \tag{4.8}$$

sendo, N_k o numero de eventos pertencentes ao *cluster* de centróide c_k. Já o termo d_{ce}, é a distância entre os *clusters*:

$$d_{ce} = \|c_k - c_l\| \tag{4.9}$$

Quanto menor for esse índice, um melhor agrupamento dos dados é obtido, ou seja, os *clusters* se encontram bem definidos e separados entre si.

Como o SOM tem como uma das suas principais virtudes o relacionamento topológico entre o mapa treinado e os dados utilizados, pode-se utilizar dessa propriedade para o entendimento da influência de cada variável no mapa gerado pelos Planos de Componentes. Tais planos são formados pelos valores médios de cada variável utilizada na classificação projetados na treliça de neurônios que forma o mapa [52], possibilitando, assim, a visualização da distribuição espacial no mapa de uma determinada variável.

A análise dos planos dos componentes, juntamente com a clusterização obtida no mapa, nos dá a relação topológica entre as variáveis e os *clusters* formados. Isto é útil na análise da influência dos sintomas nos grupos formados, auxiliando na analise das variáveis que possam ser mais influentes na elaboração dos escores clínicos.

4.3 Desenvolvimento do Escore

Após a fase de clusterização e análise dos agrupamentos formados, obtendo-se o relacionamento dos sintomas clínicos com os mesmos, segue-se a fase de projeto do sistema de apoio à decisão que possa ser utilizado na triagem dos pacientes e no auxílio ao diagnóstico da tuberculose, para aplicações em sítios com recursos bastante limitados.

Conforme indicado anteriormente, as premissas para o sistema a ser desenvolvido são: fácil implementação, utilização por médicos e enfermeiros, sem necessidade de poder computacional para operá-lo, ou seja, utilização sem a necessidade de um computador (no papel).

Considerando este contexto utilizaram-se métodos de classificação, linear como o Discriminante de Fisher [47],e aplicou-se a otimização natural como o recozimento simulado[59], sobre modelos inspirados em SOM e na informação especialista (médicos com elevada experiência no diagnóstico de TB).

4.3.1 Discriminante Linear de Fisher

A elaboração do escore pode ser encarada como um problema de classificação entre duas classes: pacientes portadores de tuberculose e não portadores de tuberculose.

O objetivo da classificação é separar os eventos da base de dados e associá-los a classes únicas, separando assim o espaço de entrada em regiões de decisão nas quais suas fronteiras são chamadas de superfície de decisão. A representação mais simples desse método é a de um classificador linear, que é um hiperplano linear, capaz de separar os eventos das classes distintas de forma que cada evento pertença a somente uma única classe.

Portanto, para realizar essa classificação linear, se faz necessário uma função discriminante que irá definir essa superfície de separação. A representação mais simples de uma função discriminante é utilizar uma função linear que projete o

vetor de entrada, no caso os sintomas dos pacientes, em uma única dimensão.

$$y(x) = \mathbf{w}^T\mathbf{x}$$

Se $y(x)$ for maior que um certo limiar de decisão, os sintomas e condições do paciente representado pelo vetor de entrada $\mathbf{x}$, será suspeito de ter tuberculose.

Em geral, a projeção dos dados em um único componente faz com que se perca muita informação, e classes que poderiam estar bem separadas no espaço original de entrada podem ficar sobrepostas em uma única dimensão. Portanto, o discriminante de Fisher tem como objetivo encontrar um vetor de pesos $\mathbf{w}$ que maximize a separação das classes da projeção dos dados em $\mathbf{w}$.

Para realização deste objetivo, a análise por discriminante de Fisher busca a direção ótima de discriminação de forma que minimize a distância intraclasses e maximize a distância interclasses [60]. Assim, é necessário encontrar a direção w_0 que maximiza a equação 4.10.

$$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S_B}\mathbf{w}}{\mathbf{w}^T\mathbf{S_W}\mathbf{w}} \tag{4.10}$$

Onde $\mathbf{S_B} = (m_1 - m_2)(m_1 - m_2)^T$ é a matriz de separação interclasses, sendo m_i a média da classe i, e $\mathbf{S_W} = \mathbf{S1} + \mathbf{S2}$ é a matriz de separação intraclasses, onde:

$$\mathbf{S_i} = \sum (\mathbf{x} - m_i)(\mathbf{x} - m_i)^T \tag{4.11}$$

Pode-se provar que a direção ótima que maximiza a equação 4.10 é dada por [61]:

$$\mathbf{w} = \frac{(m_1 - m_2)}{\mathbf{S_W}} \tag{4.12}$$

Para que a flutuação estatística inerente aos dados possa ser levada em consideração e para avaliar a capacidade de generalização do escore clínico, foi utilizada a técnica da validação cruzada [26] na estimação do discriminante de Fisher.

Neste trabalho, a validação cruzada foi realizada da seguinte forma:

1. Todo o conjunto de pacientes com presença e ausência de tuberculose em 12 blocos com número igual de pacientes em cada bloco.

2. A cada rodada de estimação do discriminante.

    - Sorteia-se, para cada classe, 8 blocos para compor o conjunto de treinamento e 4 para conjunto de teste.
    - A idade dos pacientes do conjunto de treinamento é normalizada para ficar entre 0 e 1, e os parâmetros de normalização são guardados.

$$\frac{Idade}{max(Idade)}$$

    - Estima-se o classificador baseado no conjunto de treinamento.
    - Os parâmetros de normalização são aplicados no conjunto de teste.
    - Cria-se um novo classificador composto por números inteiros [1]
    - Os resultados obtidos com os classificadores são armazenados.

Ao final das 100 rodadas que foram utilizadas na validação cruzada é possível estimar a flutuação do desempenho do escore desenvolvido segundo os critérios de avaliação de desempenho descritos na seção 4.3.2.

4.3.2 Avaliação de Desempenho

Ao longo deste trabalho, a avaliação de desempenho dos classificadores, escores clínicos, desenvolvidos foram baseados na análise de algumas propriedade da curva ROC [47] como a sensibilidade, especificidade, acuidade, valor preditivo positivo e negativo e o índice SP(soma produto). A curva ROC mostra como as probabilidades de detecção e falso alarme, respectivamente sensibilidade e (1- especificidade), variam com o patamar de decisão. A eficiência de um classificador binário também pode

[1] Os componentes do vetor **w** foram multiplicados por 10 e arredondados para o número inteiro mais próximo

ser estimada a partir da área sob a curva ROC. Quanto maior a área, mais eficiente é o classificador.

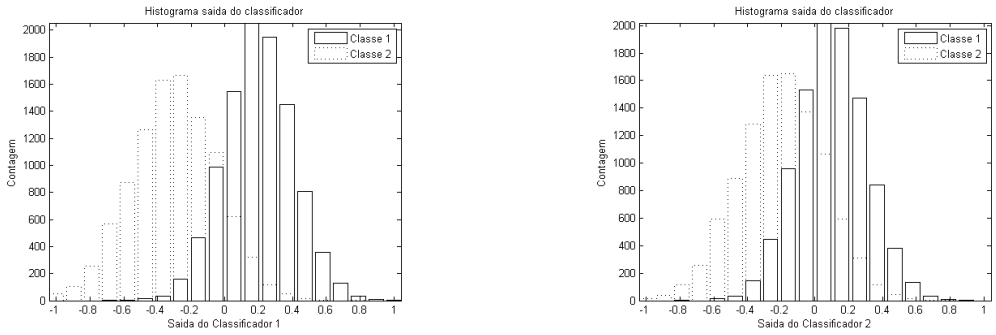
A sensibilidade (**S**) do classificador pode ser definida como a fração dos pacientes portadores de TB que o classificador é capaz de classificar como suspeitos de serem portadores da doença. A especificidade (**E**) é a proporção de pacientes sem TB que o classificador é capaz de classificar corretamente como um caso não suspeito.

Adicionalmente, o índice SP é definido por [62]:

$$SP = \sqrt{\sqrt{\mathbf{S} \times \mathbf{E}} \times \frac{\mathbf{S} + \mathbf{E}}{2}} \tag{4.13}$$

Este índice incorpora em um único valor o desempenho do classificador para duas classes e permite um projeto balanceado entre elas, haja vista que se o desempenho de especificidade ou sensibilidade cair sistematicamente, o índice SP diminuirá fortemente.

Neste trabalho, o índice SP é utilizado como parâmetro na escolha do limiar de decisão de um dado classificador. Para se encontrar o limiar ótimo, varia-se o limiar de decisão em toda sua faixa de excursão e calcula-se o índice SP correspondente. O valor máximo do índice SP indica o limiar de decisão que apresenta alta eficiência entra as duas classes, ou seja, a melhor relação entre sensibilidade e especificidade.



(a) (b)

Figura 4.4: Exemplo da saídas de dois classificadores distintos (a)Classificador 1 (b)Classificador 2.

Para exemplificar a escolha do limiar de decisão pelo índice SP, na figura 4.4 são mostradas as saídas de dois classificadores hipotéticos para duas classes distintas.

O classificador 1 apresenta uma menor superposição entre as classes em relação ao classificador 2, logo tem um melhor poder de discriminação, que se reflete numa maior área sobre a curva ROC e um índice SP com valor máximo maior. Para o classificador 1, o SP máximo ($SP_{max} = 0,87$) é atingindo no limiar de corte igual a $-0,035$, fazendo o classificador ter uma especificidade de $88,23\%$ e sensibilidade de $85,92\%$, enquanto que, para o classificador 2, tem-se $SP_{max} = 0,75$ para o limiar de corte igual a $-0,047$, com especificidade de $77,29\%$ e sensibilidade de $72,85\%$.

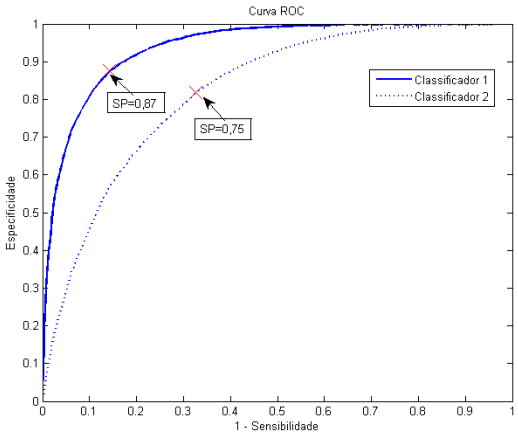


Figura 4.5: Curvas ROC dos dois classificadores com seus respectivos índice SP máximos.

Portanto, o valor máximo do índice SP serve como um índice de desempenho para comparação da eficiência de dois ou mais classificadores; pois, quanto maior for o seu valor máximo, mais eficiente é o classificador em termos de discriminação das duas classes, como pode ser visto na figura 4.5.

4.3.3 *Simulated Annealing*

Encontrar um classificador linear ótimo para os casos suspeito de TB pode ser considerado um problema de otimização combinatória quando todas as variáveis são dicotômicas, pois o espaço de busca consiste em todos os arranjos possíveis dos pesos que podem ser associados aos sintomas. Tal problema pode ser visto como o já conhecido Problema Quadrático de Alocação (PQA).

O PQA é um problema da classe NP-difícil [63], e possui aplicação em diversas

áreas como engenharia, economia, arquitetura e ergonometria. Consiste em alocar objetos de forma que cada um seja posicionado em um único local, com o objetivo de otimizar as distâncias ou fluxos de demanda entre cada par, ou então o custo associado ao posicionamento dos próprios objetos. Em Loiola [64], é apresentada uma revisão de várias abordagens para encontrar a solução de um PQA, que a técnica de recozimento simulado é amplamente utilizada para resolução deste tipo de problema [65] [66].

No caso deste trabalho, o problema de classificação dos casos entre TB e não TB pode ser visto como um problema de alocação de objetos, tendo como custo o posicionamento das projeções dos dados dos pacientes nos escores em regiões que maximizem a separação entre os casos de TB e não TB.

Simulated annealing é um algoritmo de otimização estocástico, inspirado num fenômeno físico conhecido como recozimento, do inglês (*Annealing*). O recozimento é um método utilizado na metalurgia que visa a formação de uma rede cristalina de átomos. Esse processo ocorre quando um sólido é aquecido até o seu ponto de fusão, fazendo com que os átomos do material fiquem livres para se movimentarem, devido ao elevado grau de agitação térmica.

Se o material fundido for resfriado muito rapidamente, processo conhecido como tempera, os átomos não terão tempo suficiente para se rearranjarem de forma regular e organizada. Assim, o sólido apresentará um arranjo irregular de átomos. Por outro lado, se o material fundido for resfriado lentamente, os átomos terão tempo para encontrar a melhor forma de se rearranjarem e restabelecerem suas ligações químicas. Esse arranjo ótimo forma uma estrutura cristalina que representa a condição de mínima energia potencial dos átomos.

Inspirado nesse fenômeno físico, foi desenvolvido o algoritmo de simulated annealing [59], que tem como objetivo encontrar mínimos globais de problemas de otimização bastante complexos.

O algoritmo inicia a partir de um ponto inicial p, escolhido aleatoriamente no espaço de busca, com uma temperatura inicial t. Uma perturbação é aplicada ao

ponto inicial de forma a obter um novo ponto p' nas vizinhanças de p. Então a diferença entre as energias, em relação à função custo, de ambos os pontos é calculada.

$$\Delta J = J(p') - J(p)$$

Caso o novo ponto tenha uma energia menor que o ponto inicial, $\Delta J < 0$, o ponto p é descartado e p' passa a ser o ponto atual da busca ($p \leftarrow p'$). Por outro lado, se p' apresentar uma energia maior, $\Delta J > 0$, ou seja um custo maior, a aceitação do novo ponto p' ocorre de acordo com uma probabilidade, que neste trabalho, foi dada pela lei de Boltzmann:

$$P(\Delta J) = \exp(-\Delta J/kt) \tag{4.14}$$

onde k é a constante de Boltzmann e t é a temperatura absoluta. Assim, uma vez definido se a nova solução será aceita ou não, o algoritmo é repetido, aplicando-se novamente uma perturbação ao ponto atual e decidindo se o novo ponto deve ser aceito como a nova solução atual. Esse processo é, então, repetido até que um critério de parada seja satisfeito.

O fato do algoritmo permitir que uma solução pior, com maior energia, seja aceita como solução atual dá ao simulated annealing a capacidade de fugir de mínimos locais durante a busca. Caso toda solução pior fosse descartada, a busca convergiria rapidamente para um mínimo local.

De acordo com a equação 4.14, a probabilidade de aceitação de uma solução pior é maior quanto mais alta for a temperatura t. Dessa forma, a temperatura inicial deve ser alta para que o método possa explorar bem o espaço de busca. A temperatura t deve ser lentamente reduzida ao longo das iterações do algoritmo, seguindo um plano de resfriamento pré-estabelecido. Assim, a probabilidade de aceitação de soluções piores vai progressivamente diminuindo, permitindo a convergência do simulated anneling para o mínimo global.

A temperatura inicial t_0 e o plano de resfriamento são parâmetros que influenciam criticamente na convergência do algoritmo. Trabalhar com temperaturas exagera-

damente altas dificulta a convergência do método. Entretanto, temperaturas baixas ou resfriamentos muito rápidos fazem com que o algoritmo venha a convergir prematuramente para mínimos locais. A escolha desses parâmetros é bastante dependente da função a ser otimizada.

Neste trabalho, a energia a ser minimizada é função custo definida por:

$$J(\mathbf{w}) = 1 - SP_{max}(\mathbf{w}) \tag{4.15}$$

Onde $SP_{max}(\mathbf{w})$ é o valor máximo do índice soma produto, definido pela equação 4.13, para o vetor de pesos $\mathbf{w}$ que está contido no espaço de busca. Como o problema de classificação de casos suspeitos de TB é um problema muito complexo, o algoritmo de simulated annealing foi inicializado cinco vezes com estados inicial de partida escolhidos aleatoriamente dentro do espaço de busca, com temperatura inicial elevada $t_0 = 100$, função de resfriamento exponencial, onde $t_{i+1} = 0,95 \times t_i$, e critério de parada por variação do valor da função custo, fazendo o algoritmo parar caso a variação fosse menor que $1x10^{-6}$. Somente o resultado da inicialização que rendeu o menor valor da função custo, ao final do algoritmo, é armazenada.

Para facilitar o desenvolvimento e uso do escore encontrado pelo algoritmo de simulated anneling foi feita uma transformação na base de dados, de forma que todas as variáveis fossem dicotômicas. A idade foi categorizada em duas partes, a primeira parte contempla as pessoas jovens com até 35 anos de idade, uma vez que essa idade equivale ao início da segunda moda do histograma dos casos de TB, ver figura 4.1 c, e esta próxima da mediana da idade dos pacientes diagnosticados com TB é de 39. E foram criados diferentes modelos, conforme descritos a seguir, para estudo do desempenho do escore.

- Modelo 1 - Escore com pesos separados para a presença e ausência dos sintomas

  O modelo foi criado seguindo recomendações de médicos especialistas em pneumologia que recomendaram dar diferentes pontuações para a presença e a ausência dos sintomas

Idade categorizada em duas partes: Idade $\leq$ 35 anos e Idade $>$ 35 anos. O paciente que estiver numa certa faixa de idade assume valor 1 para a mesma e 0 para a outra.

Os outros sintomas serão separados entre presença e ausência do sintoma. Com isso o sintoma será representado por uma codificação binária de dois dígitos, sendo a presença do sintoma representada pela dupla (1,0), a ausência por (0,1) e quando ignorado por (0,0). Logo, o vetor que caracteriza um paciente será formado pela união da segmentação da idade com as duplas dos sintomas.

- Modelo 2 - Escore com pesos separados para a presença, ausência e para os casos ignorados dos sintomas

  Análogo ao modelo 1, mas será avaliado o fato de um determinado sintoma ser ignorado.

  Idade categorizada em três partes: Idade $\leq$ 35 anos, 35 $<$ Idade $\leq$ 65 anos e Idade $>$ 65 anos. O paciente que estiver numa certa faixa de idade assume valor 1 para a mesma e 0 para as outras.

  Os outros sintomas serão separados entre presentes, ausentes e ignorados. Com isso o sintoma será representado por uma codificação binária de três dígitos, sendo a presença do sintoma representada pela tripla (1,0,0), a ausência por (0,1,0) e quando ignorado por (0,0,1). Logo, o vetor que caracteriza um paciente será formado pela união da segmentação da idade com as duplas dos sintomas.

- Modelo 3 - Escore com pesos separados somente para a presença dos sintomas

  Modelo semelhante ao escore hoje utilizado na triagem de pacientes na PAAP

  Idade segmentada em duas partes: Idade $\leq$ 35 anos e Idade$>$35 anos. O paciente que estiver numa certa faixa de idade assume valor 1 para a mesma e 0 para a outra.

Os outros sintomas serão representados somente quando presentes. Com isso o sintoma será representado por uma codificação binária de hum dígito, sendo a presença do sintoma representada pelo dígito (1) e quando o sintoma for ausente ou ignorado pelo dígito (0). Logo, o vetor que caracteriza um paciente será formado pela união da segmentação da idade com a representação da presença dos sintomas.

Outro fator importante no desenvolvimento do escore é a definição do espaço de busca. Foram testados diferentes espaços de busca limitados por um valor máximo e mínimo que cada peso do escore poderia alcançar. Portanto, para cada modelo, foram desenvolvidos diferentes escores, de forma que os pesos dos mesmos tivesses valores reais com limites superiores e inferiores indo de 1 e -1 até 15 e -15 respectivamente. E, além disso, para o modelo 3 também foram desenvolvidos escores com pontuação com valores somente positivos com limites superiores entre 1 e 15.

Como o escore a ser utilizado tem a premissa de ser de fácil utilização e não requerer poder computacional para o seu uso, após a determinação do mesmo, pelo algoritmo de simulated annealing, seus pesos são arredondados para o número inteiro mais próximo e assim utilizados no cálculo do desempenho dos mesmos.

4.4 Cenários de Aplicação

Na prática clínica, quatro situações básicas estão sempre presentes: diagnóstico, tratamento, prognóstico e prevenção. Para se obter um bom resultado em cada uma dessas etapas, é necessária a identificação do problema e investigação cuidadosa das informações disponíveis.

Numa situação de recursos escassos, como em unidades de saúde básicas, ou por motivos de biossegurança, antes do diagnóstico se faz necessária uma etapa de seleção dos pacientes conhecida como triagem. Onde os profissionais de saúde têm que escolher qual o paciente que irá receber atendimento prioritário.

No caso da TB é desejável que a triagem consiga detectar o máximo possível de

pacientes suspeitos de serem TB positivos, indo de acordo com a estratégia StopTB, mas também , do ponto de vista da biossegurança, não deixar pessoas sadias expostas ao bacilo que pode ser transmitido por um paciente suspeito de ter TB. Portanto, quando o escore for analisado num cenário de triagem é desejável que o mesmo tenha a maior sensibilidade possível, para detectar os casos de TB, ao mesmo tempo que o escore não perca em especificidade, por motivos de biossegurança.

Já quando o escore for utilizado no apoio ao diagnostico os objetivos estão relacionados ao custo da oportunidade de iniciar corretamente o tratamento de um paciente suspeito de TB. Este cenário de aplicação é muito importante para casos onde os recursos humanos são escassos e se tem que tomar uma decisão rápida e confiável para agendamento de exames mais específicos e de inicialização do tratamento de anti-proliferação do bacilo. Portanto, quando o escore for analisado para o cenário de diagnóstico é desejável que o mesmo tenha a maior especificidade possível, excluindo do tratamento os casos que não tem TB, e um grande valor preditivo positivo, que dará a razão entre os casos que foram previstos como suspeito de TB e de fato serão diagnosticadas com TB positiva.

O rendimento dos escores desenvolvidos neste trabalho neste dois cenários guiarão a escolha do melhor escore e dos seus respectivos limiares de decisão que atendam as exigências impostas pelos cenários de aplicação.

Capítulo 5

Resultados

Este capítulo tem como objetivo mostrar os resultados obtidos pela metodologia de desenvolvimento deste trabalho, que foram descritas no capítulo anterior, no qual foram expostas as técnicas de clusterização e construção do escore clínico.

Todos os resultados apresentados neste capítulo são obtidos a partir do teste das técnicas utilizadas sobre todo o conjunto de dados. O uso da base toda é a melhor forma de compararmos os resultados dos escores desenvolvidos pelas diferentes técnicas, bem como o trabalho de referência [43], apresentado na seção 3.

A seguir, serão mostrados e discutidos os resultados do agrupamento dos casos existentes na base de dados por meio dos Self-Organizing Maps. Em seguida, na seção 5.2, serão apresentados os resultados dos diferentes escores desenvolvidos. Por fim, serão expostas as análises gerais sobre o relacionamento dos sintomas estudados com os agrupamentos formados, assim como a comparação entre o escore referência e o melhor escore obtido neste trabalho.

5.1 Clusterização

Nesta seção, serão mostrados os resultados obtidos na clusterização da base de dados com o uso do Self-Organizing Maps. Como a análise dos agrupamentos tem um caráter meramente qualitativo, serão feitas análises sobre os agrupamentos formados no mapa, assim como o relacionamento dos mesmos com os sintomas clínicos

representados nos planos dos componentes.

Ao longo deste trabalho, foi utilizado um mapa bidimensional, de tamanho 8x8 com treliça hexagonal, com função de vizinhança gaussiana e aprendizagem em batelada. A escolha do tamanho do mapa foi baseada de forma empírica. Foi levado em consideração que o mapa não tivesse uma grande quantidade de neurônios não ativos, casos de mapas muito grandes, e tivesse um número suficiente de neurônios de forma a não se perder a capacidade de generalização do mapa atendendo aos requisitos necessários para tratar convenientemente a tarefa de agrupar os casos numa ótica do diagnóstico de tuberculose.

O resultado do treinamento do mapa pode ser avaliado tanto de forma quantitativa, como qualitativa. A forma quantitativa usa o erro de quantização e o erro topográfico do mapa. O erro de quantização, Q_e, corresponde à média do erro, dos N casos usados no treinamento, correspondente à diferença entre o vetor de características $\mathbf{x_k}$ e o vetor de código,w_{BMU}, da sua respectiva **BMU**.

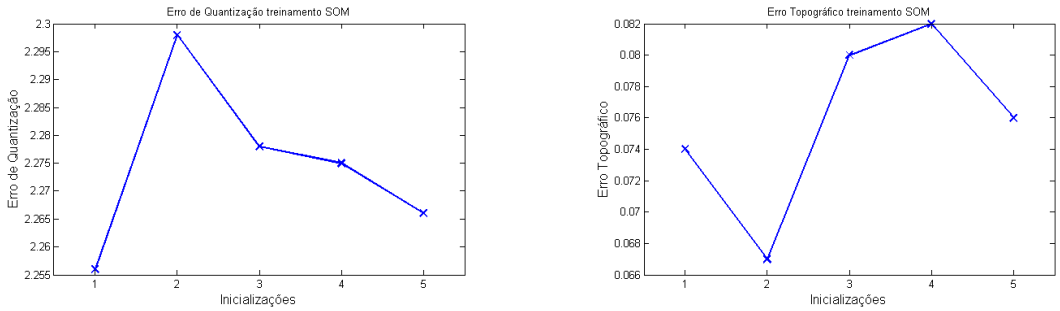
$$Q_e = \frac{1}{N}\sum_{k=1}^{N} \|\mathbf{x_k} - \mathbf{w_{BMU}}\|$$

Já o erro topográfico, T_e, avalia o quanto os neurônios do mapa se aproximam dos padrões existentes no espaço de entrada. Avaliando o quanto os dois neurônios mais próximos da entrada, $\mathbf{BMU}_1$ e $\mathbf{BMU}_2$ estão próximos entre si na estrutura do mapa,

$$T_e = \frac{1}{N}\sum_{k=1}^{N} u(x_k)$$

onde $u(x_k)$ é igual a 1 se $\mathbf{BMU}_1$ e $\mathbf{BMU}_2$ não são vizinhos, e 0, caso o contrário. Já na análise qualitativa do treinamento, é realizada uma inspeção visual dos planos dos componentes e da U-Matrix, que busca encontrar os agrupamentos formados e os seus relacionamentos com as variáveis.

Ao longo do desenvolvimento, os pesos sinápticos do SOM foram treinados com cinco inicializações diferentes. Conforme visto na figura 5.1, a variação de ambos os erros de quantização e topográfico, ao final do treinamento, é muito pequena.



(a) (b)

Figura 5.1: Análise quantitativa do treinamento do SOM (a)Erro de Quantização (b)Erro Topográfico

A escolha do mapa foi realizada de forma qualitativa, sendo escolhido o mapa que mostrou relacionamento entre as variáveis e os agrupamentos que foram visualizados na U-Matrix. Através da análise gráfica da U-Matrix, ilustrada na figura 5.2, pode ser observado um único agrupamento de dados bem definido no canto superior direito da imagem. Já na parte inferior central da imagem, há uma região candidata a agrupamento, mas não há uma separação bem definida entre os neurônios. A região central forma, aparentemente, uma região homogênea, ou seja, não se percebe um agrupamento explícito.

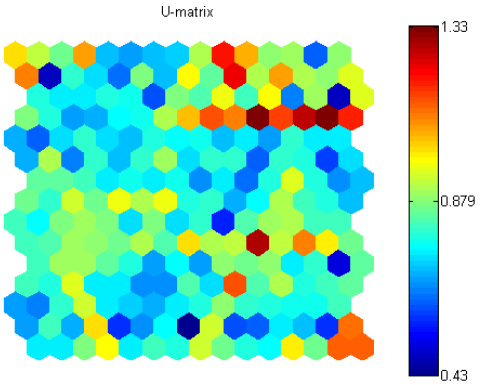
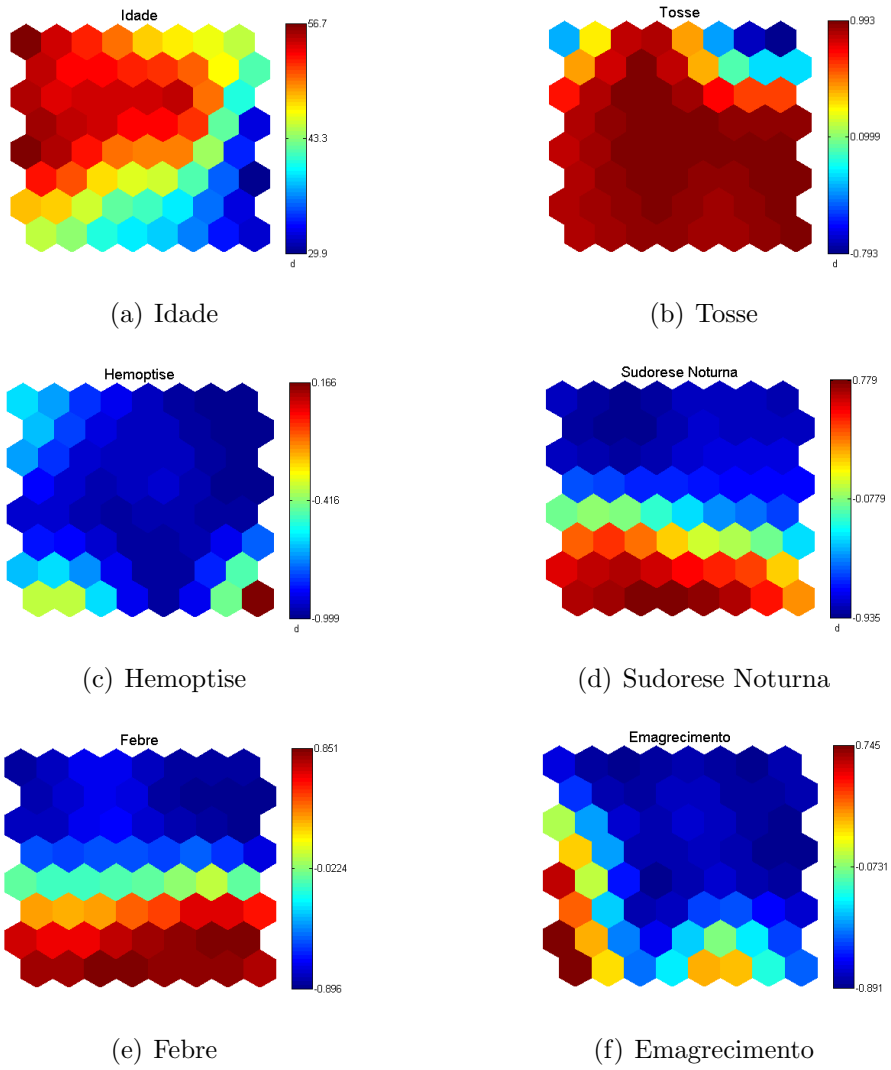


Figura 5.2: U-Matrix do Mapa Escolhido

Cada Plano de Componente, ilustrados nas figuras 5.3 e 5.4, representa a distribuição do respectivo sintoma em cada neurônio, mostrando a média dos valores do componente dos casos projetados em cada um dos neurônios do mapa. A média dos componentes é representada por uma paleta de cores, onde as cores quentes repre-

sentam os valores elevados e as cores frias valores reduzidos destas médias. Portanto, sob os planos de componentes podemos analisar o relacionamento entre as variáveis e com a U-Matrix, de forma espacial.



(a) Idade (b) Tosse

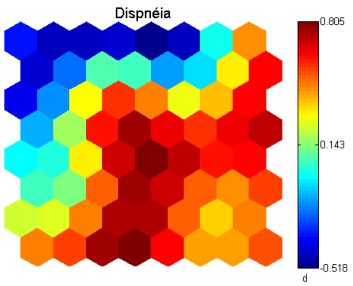
(c) Hemoptise (d) Sudorese Noturna

(e) Febre (f) Emagrecimento

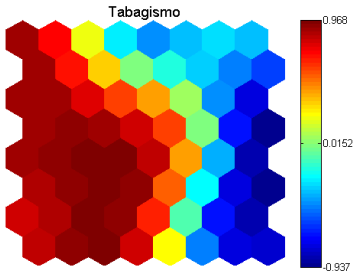
Figura 5.3: Planos de Componentes

Numa primeira etapa de análise dos resultados gerados no treinamento do SOM, podemos observar, por meio de inspeção visual dos mapas, algumas características que serão descritas a seguir.

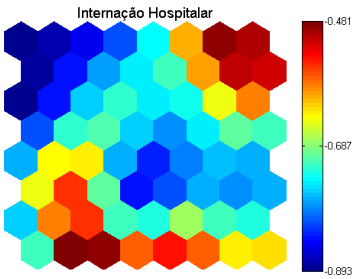
Quando analisada a distribuição da Idade no mapa, fig 5.3 (a), vemos uma forte concentração de pessoas jovens no canto inferior direito do mapa. Também podemos notar a existência de uma concentração de neurônios na parte central e superior esquerda do mapa, que tem valor associado perto da idade média dos casos da base



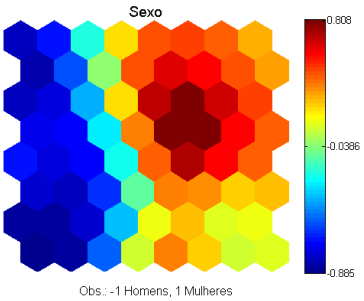
(a) Dispnéia



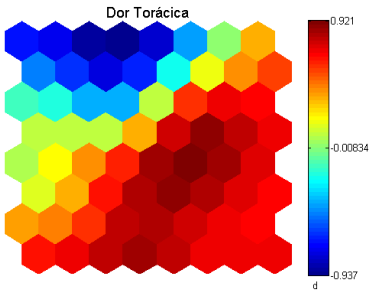
(b) Tabagismo



(c) Internação Hospitalar



(d) Sexo



(e) Dor Torácica

Figura 5.4: Continuação Planos de Componentes

de dados, que é de 45,8 anos.

Vale notar que os casos que não apresentam Tosse se concentram no canto superior direito do mapa, figura 5.3 (b), indicando que os pacientes que não apresentam o sintoma da Tosse são casos bem distintos em relação ao resto da população estudada.

Já o plano que representa o sintoma Hemoptise, figura 5.3 (c), não apresenta muitas informações relevantes, uma vez que a distribuição dos sintomas pelos neurônios é majoritariamente uniforme, com valor igual a -1, o que indica a ausência do sintoma, na parte central e superior do mapa. Somente alguns poucos neurônios na parte inferior do mapa acusaram a presença do sintoma. Entretanto, essa predominância dos neurônios, na cor azul, faz sentido, quando se considera a distribuição do sintoma entre os casos da base de dados. (ver tabela 4.1)

Ao compararmos os sintomas Sudorese Noturna e Febre, figuras 5.3 (d) e 5.3 (e), respectivamente, percebe-se que a distribuição de ambos os sintomas pelo mapa é muito similar, com uma grande concentração dos casos que apresentam esses sintomas na parte inferior do mapa, sendo esta característica um forte indício que ambos os sintomas podem ser estatisticamente correlacionados.

Os casos que apresentam Emagrecimento não chegam a formar agrupamentos bem distintos no mapa, figura 5.3 (f); entretanto, a presença do sintoma está mais concentrada na parte inferior e em toda a lateral esquerda do mapa, não dando informações muito relevantes nessa primeira análise.

No plano de componente referente à Internação Hospitalar, figura 5.4 (c), podemos ver uma pequena concentração dos casos que não declaram ter sofrido internação hospitalar no canto superior esquerdo do mapa e uma leve concentração de casos que declaram o sintoma na parte inferior e no canto superior direito do mapa. Apesar dessas pequenas zonas de concentração, podemos ver que existe um grande números de neurônios com valor próximo a $-0,68$, indicando que os casos com e sem o sintoma se distribuem uniformemente pelo resto do mapa. Efetivamente, a razão entre os casos com e sem o sintoma é de 16/100, e o mapa mostra o valor médio do sintoma por neurônio, sendo essa variável dicotômica, com média amostral dos

casos igual a $-0,68$. Portanto podemos concluir que a Internação Hospitalar não é uma variável muito discriminatória, o que condiz com o seu *Odds Ratio*, conforme mostrado na tabela 4.1.

Os mapas dos pares de sintomas Dispnéia e Dor Torácica, figuras 5.4 (a) e 5.4 (e), respectivamente, e Sexo e Tabagismo, figuras 5.4 (b) e 5.4 (d), respectivamente, têm distribuição espacial no mapa similar, o que dá indícios, assim como a Febre e a Sudorese Noturna, que esses sintomas podem estar estatisticamente correlacionados. No caso de Dispnéia e Dor Torácica, esses sintomas são clinicamente correlacionados, uma vez que, dores na região do tórax podem ser devidas à dificuldades na respiração e/ou podem causar dificuldade na respiração. Já no caso de Tabagismo e Sexo do paciente, essa correlação não tem sentido clínico. Porém, pela análise dos mapas, podemos inferir que, na população em estudo, há uma proporção maior de homens do que de mulheres fumantes, podendo indicar uma informação georeferenciada, característica da região de estudo.

5.1.1 Descoberta de agrupamentos no SOM e análise da dependência espacial

Após essa etapa de inspeção visual dos planos dos componentes e da U-Matrix, foi realizada a busca pelos agrupamentos existentes na base de dados. A partição da base de dados num número c de agrupamentos, ou *clusters*, foi realizada em duas fases, conforme descrito na seção 4.2.1.2 e exemplificado na figura 5.5. Primeiramente, os dados são utilizados no treinamento do SOM. Em seguida, os vetores de código, que são os pesos sinápticos dos neurônios do mapa, são particionados pelo método de k-means, formando os agrupamentos reconhecidos na base de dados.

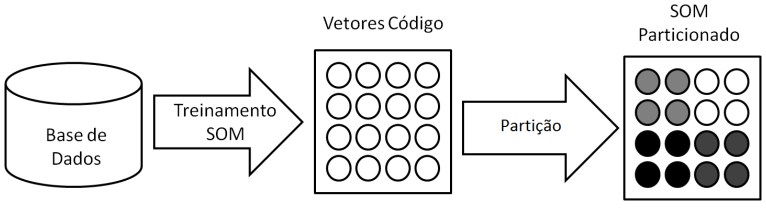


Figura 5.5: Processo de Particionamento dos Dados

Os agrupamentos formados podem ser validados pelo índice de Davies-Bouldin, conforme mostrado na figura 5.6. Nesse caso, vemos que, para o mapa escolhido, o menor índice de Davies-Bouldin, $I_{DB} = 0,8021$, se dá quando o mapa é dividido em 12 agrupamentos distintos.

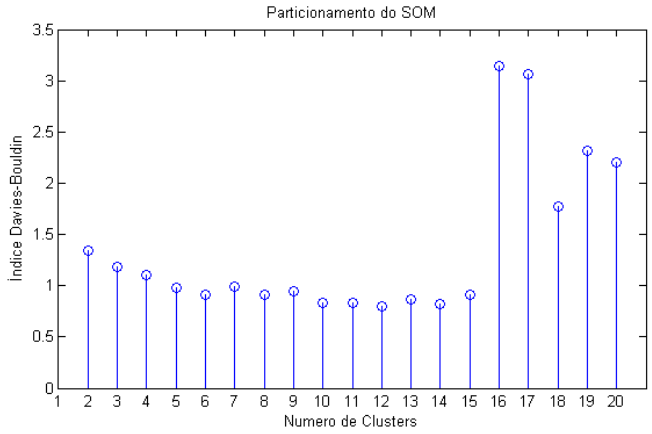


Figura 5.6: Indice Davies-Bouldin por número de *clusters* encontrados no SOM

O objetivo desta etapa de clusterização é o entendimento dos fatores de risco para a tuberculose pulmonar na população em estudo, através do relacionamento dos sinais e sintomas com os *clusters* formados. Apesar do número ideal de *clusters*, pelo índice de Davies-Bouldin, ser de 12 clusters, neste trabalho será feita uma analogia aos agrupamentos formados por rede ART [49] que representem grupos de baixo, médio e alto risco, com respeito à tuberculose paucibacilar, atendidos no Hospital Universitário Clementino Fraga Filho [45]. Portanto, foi utilizada a partição que encontrou somente 3 *clusters* no mapa do SOM, tendo este, $I_{DB} = 1,19$. A incidência dos pacientes diagnosticados com tuberculose e os pacientes sem a doença em cada *cluster* encontrado, conforme a tabela 5.1.1, foi o que determinou se os casos contidos em cada *cluster* apresentavam baixo, médio e alto risco de serem TB positivos.

	TB		Não TB		Pacientes	Risco
	Total	(%) do Cluster	Total	(%) do Cluster		
grupo 1	45	13,16%	297	86,84%	342	Médio
grupo 2	165	39,47%	253	60,53%	418	Alto
grupo 3	32	8,33%	352	91,67%	384	Baixo
Total	242		902		1144	

Tabela 5.1: Distribuição dos pacientes por *clusters*

Portanto, o mapa ficou segmentado conforme a figura 5.7, sendo que as cores verde, amarelo e vermelho fazem uma alusão a um semáforo de transito, demonstrando os *clusters* que são de baixo, médio e alto risco, respectivamente.

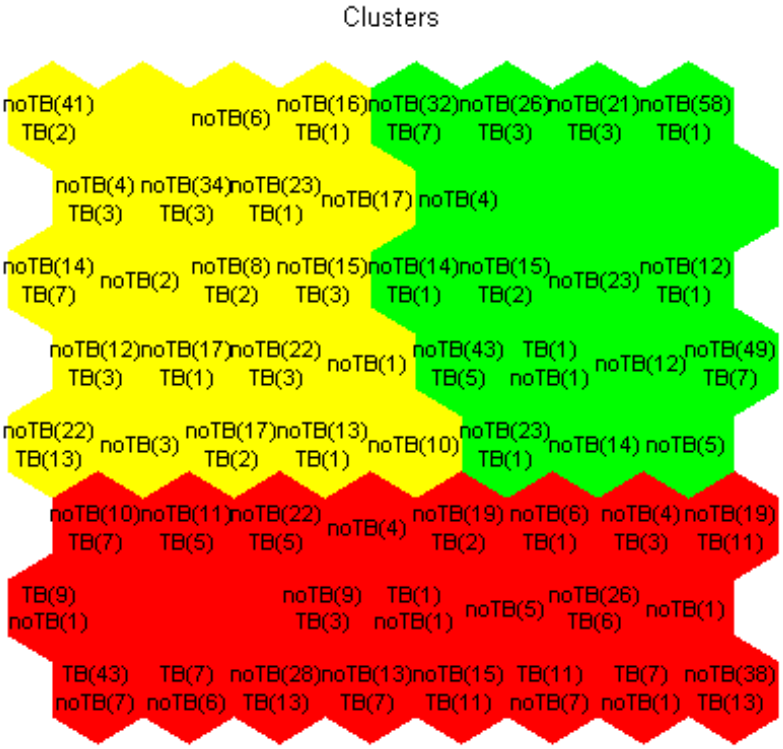


Figura 5.7: Representação do mapa de neurônios após sofrer clusterização

Após a clusterização dos dados, podemos analisar, de forma visual, através do relacionamento espacial dos Planos de Componentes com o mapa de neurônios segmentados por *cluster*, e, de forma quantitativa, através da distribuição dos sintomas por *clusters*, a influência dos sintomas nos grupos de risco.

Comparando o plano do componente Tosse, figura 5.3 (b), com o mapa segmentado, figura 5.7, e a tabela 5.2, podemos observar que a ausência do sintoma caracteriza o grupo de baixo risco, pois, numa inspeção visual, comparando espacialmente o Plano de Componente com o mapa segmentado, vemos a concentração dos casos sem tosse no canto superior direito do plano, o que condiz com a posição do *cluster* de baixo risco.

Especificamente, 66,87% dos casos que não têm TB e não apresentam o sintoma

Tosse				
	Não		Sim	
TB -	157		745	
TB +	13		229	
Cluster Baixo Risco				
	Não	%	Sim	%
TB -	105	66,88	247	33,15
TB +	7	53,85	25	10,92
Cluster Médio Risco				
	Não	%	Sim	%
TB -	44	28,03	253	33,96
TB +	3	23,08	42	18,34
Cluster Alto Risco				
	Não	%	Sim	%
TB -	8	5,10	245	32,89
TB +	3	23,08	162	70,74

Tabela 5.2: Distribuição da Tosse entre os *clusters*

Tosse		
Clusters		
Baixo Risco	Médio Risco	Alto Risco
9,19%	14,24%	39,80%

Tabela 5.3: Prevalência da Tosse nos casos TB positivos por *clusters*

Tosse se encontram no grupo de baixo risco. Os casos que apresentam o sintoma se distribuem de forma quase que igualitária entre os *clusters* para os casos de não TB, demonstrando que a tosse é um sintoma que se manifesta em diversas doenças pulmonares, e que, não necessariamente, é o melhor discriminante para TB. Entretanto, o *cluster* de alto risco concentra, aproximadamente, 41% de todos os casos que declaram ter tosse, sendo que 70% dos pacientes diagnosticados com TB e que apresentaram o sintoma estão nele, indicando que o sintoma é um dos sintomas principais da TB, o que vai de acordo com a literatura médica.

O Plano do Componente que representa a Hemoptise, figura 5.3 (c), não nos dá muita informação visual. Porém, ao analisarmos a tabela 5.4, podemos entender porque o plano do componente não dá informações claras sobre esse sintoma. Primeiramente, a presença do sintoma foi declarada por poucos pacientes, 9,72% (111/1142). Ademais, os casos que não apresentam o sintoma estão distribuídos entre os três *clusters*, o que justifica o plano de componente quase uniforme na cor

azul, o que representa a ausência do sintoma. No entanto, a Hemoptise é um sintoma clínico muito importante para o diagnóstico da TB, segundo a literatura médica, e podemos ver que, apesar da baixa presença do sintoma, ele é discriminante para TB, pois 77,50% (31/40) dos pacientes que apresentam o sintoma e são TB positivo estão no *cluster* de alto risco.

Hemoptise				
	Não		Sim	
TB -	831		71	
TB +	202		40	
Cluster Baixo Risco				
	Não	%	Sim	%
TB -	384	46,21	4	5,63
TB +	30	14,85	2	5,00
Cluster Medio Risco				
	Não	%	Sim	%
TB -	278	33,45	19	26,76
TB +	38	18,81	7	17,50
Cluster Alto Risco				
	Não	%	Sim	%
TB -	205	24,67	48	67,61
TB +	134	66,34	31	77,50

Tabela 5.4: Distribuição da Hemoptise entre os *clusters*

Hemoptise		
Clusters		
Baixo Risco	Médio Risco	Alto Risco
33,33%	26,92%	39,24%

Tabela 5.5: Prevalência da Hemoptise nos casos TB positivos por *clusters*

Vale notar, também, que outras doenças pulmonares graves, como pneumonia e câncer de pulmão, também têm esse sintoma em comum e o *cluster* de alto risco contém 67,61% (48/71) dos pacientes que apresentaram o sintoma e não foram diagnosticadas com tuberculose, demonstrando, portanto, a importância desse sintoma no diagnóstico de doenças pulmonares graves.

Assim como observado na relação espacial entre os Planos dos Componentes Sudorese Noturna e Febre, figuras 5.3 (d) e 5.3 (e), respectivamente, ao analisarmos esses planos em relação ao mapa de neurônios segmentado, figura 5.7, vemos que

ambos os sintomas têm como característica principal a relação da sua presença com o *cluster* de alto risco, haja vista que os casos que apresentam o sintoma se encontram concentrados na região inferior do Plano de Componente, o qual é caracterizado como o *cluster* de alto risco.

Sudorese Noturna				
	Não		Sim	
TB -	623		279	
TB +	109		133	
Cluster Baixo Risco				
	Não	%	Sim	%
TB -	322	51,69	30	10,75
TB +	27	24,77	5	3,76
Cluster Medio Risco				
	Não	%	Sim	%
TB -	255	40,93	42	15,05
TB +	37	33,94	8	6,02
Cluster Alto Risco				
	Não	%	Sim	%
TB -	46	7,38	207	74,19
TB +	45	41,28	120	90,23

Tabela 5.6: Distribuição da Sudorese Noturna entre os *clusters*

Sudorese Noturna		
Clusters		
Baixo Risco	Médio Risco	Alto Risco
14,29%	16,00%	36,70%

Tabela 5.7: Prevalência da Sudorese Noturna nos casos TB positivos por *clusters*

Tal comportamento também pode ser notado nas tabelas 5.6 e 5.8, que apresentam a distribuição desses sintomas pelos *clusters* encontrados. No caso da Sudorese Noturna, nos pacientes TB positivos, esse sintoma está presente em 54,95% dos pacientes; entretanto, aproximadamente 90% desses casos ficaram no *cluster* de alto risco. No caso da Febre, existe uma porcentagem maior de casos com presença do sintoma nos pacientes TB positivos, 67,77%, mas o padrão de distribuição dos pacientes TB positivos, que têm o sintoma em questão, é o mesmo da Sudorese Noturna. Ao analisarmos os casos que não apresentam o sintoma, vemos que a distribuição destes sintomas entre os *clusters*. No caso da Sudorese Noturna (47,67% - 349/732)

Febre				
	Não		Sim	
TB -	567		335	
TB +	78		164	
Cluster Baixo Risco				
	Não	%	Sim	%
TB -	297	52,38	55	16,42
TB +	30	38,46	2	1,22
Cluster Medio Risco				
	Não	%	Sim	%
TB -	246	43,39	51	15,22
TB +	31	39,74	14	8,54
Cluster Alto Risco				
	Não	%	Sim	%
TB -	24	4,23	229	68,36
TB +	17	21,79	148	90,24

Tabela 5.8: Distribuição da Febre entre os *clusters*

Febre		
Clusters		
Baixo Risco	Médio Risco	Alto Risco
3,51%	21,54%	39,26%

Tabela 5.9: Prevalência da Febre nos casos TB positivos por *clusters*

e Febre (50,69% - 327/645), os pacientes que não apresentaram o sintoma estão associados ao grupo de baixo risco, sendo que os casos que não apresentam o sintoma e são TB negativo nesse grupo são 51,68% e 52,38%, respectivamente.

Portanto, pela similaridade gráfica entre os Planos de Componentes e os *clusters* encontrados no mapa e pela distribuição dos sintomas pelos *clusters* encontrados, podemos inferir que Febre e Sudorese Noturna são variáveis discriminantes na clusterização encontrada.

Já no caso do emagrecimento, a presença do sintoma se distribui, conforme a figura 5.3 (e), em toda a lateral esquerda, com uma pequena concentração no canto inferior, onde se encontram os *clusters* de alto e médio risco, segundo o mapa segmentado. Esse comportamento da distribuição do sintoma pelos *clusters* é mais bem visto na tabela 5.10, na qual podemos notar uma leve concentração de pacientes que não apresentam o sintoma no *cluster* de baixo risco, sendo que 43,9% dos casos que não apresentam o sintoma e são TB negativo estão neste *cluster*. Também nota-se

uma concentração dos pacientes que apresentaram emagrecimento nos *cluster* de médio e alto risco, 27,99% (82/293) e 59,39%(174/293), respectivamente; no entanto, há uma grande concentração de pacientes TB positivos e que não declararam ter sofrido emagrecimento, 59,02% (79/122) no *cluster* de alto risco. Isto dá um indicio de que o paciente sofrer emagrecimento é discriminante para a clusterização dos casos na base de dados, mas não tão relevante como a hemoptise, febre e emagrecimento aparentam ser.

Emagrecimento				
	Não		Sim	
TB -	729		173	
TB +	122		120	
Cluster Baixo Risco				
	Não	%	Sim	%
TB -	320	43,90	32	18,50
TB +	27	22,13	5	4,17
Cluster Medio Risco				
	Não	%	Sim	%
TB -	237	32,51	60	34,68
TB +	23	18,85	22	18,33
Cluster Alto Risco				
	Não	%	Sim	%
TB -	172	23,59	81	46,82
TB +	72	59,02	93	77,50

Tabela 5.10: Distribuição do Emagrecimento entre os *clusters*

Emagrecimento		
Clusters		
Baixo Risco	Médio Risco	Alto Risco
13,51%	26,83%	53,45%

Tabela 5.11: Prevalência da Emagrecimento nos casos TB positivos por *clusters*

Os Planos dos Componentes referentes à Dispnéia e a Dor Torácica , figuras 5.4 (a) e 5.4 (e), apresentam uma distribuição espacial muito parecida, sendo que a presença destes sintomas se distribui em toda parte inferior, central e na lateral direita superior. Se compararmos com o mapa da segmentação, vemos que essas áreas acabam por englobar todos os *clusters*. Esse padrão de distribuição dos sintomas pode ser visto nas tabelas 5.12 e 5.13, de modo que observa-se uma desproporção entre

os casos que apresentam e não apresentam os sintomas. A maioria dos pacientes que apresentam os sintomas foi diagnosticada sem TB, 80,21% (600/748) e 77,72% (579/745) para a Dispnéia e a Dor Torácica, respectivamente. Podemos, ainda ver que esses casos se distribuem quase que igualmente entre os *clusters*, no caso da Dispnéia, e com uma leve concentração nos *clusters* de baixo e alto risco, para Dor Torácica. Entretanto, há uma concentração dos casos que apresentaram o sintoma e foram diagnosticados com TB no *cluster* de alto risco, o que está de acordo com a literatura médica. Também se pode observar que os casos TB negativos e que não apresentam o sintoma estão concentrados no *cluster* de médio risco, enquanto que, para os outros sintomas discutidos anteriormente, esse padrão ocorre no *cluster* de baixo risco. Devido ao padrão de distribuição dos sintomas entre os *clusters* e do grande número de casos que apresenta o sintoma e não foram diagnosticados com TB, podemos inferir que Dispnéia e Dor Torácica são sintomas que confundem a clusterização.

Dispnéia				
	Não		Sim	
TB -	302		600	
TB +	94		148	
Cluster Baixo Risco				
	Não	%	Sim	%
TB -	119	39,40	233	38,83
TB +	17	18,09	15	10,14
Cluster Medio Risco				
	Não	%	Sim	%
TB -	122	40,40	175	29,17
TB +	26	27,66	19	12,84
Cluster Alto Risco				
	Não	%	Sim	%
TB -	61	20,20	192	32,00
TB +	51	54,26	114	77,03

Tabela 5.12: Distribuição da Dispnéia entre os *clusters*

Ao compararmos o Plano do Componente Sexo, figura 5.4 (d), com o mapa segmentado, podemos perceber que a concentração dos pacientes do sexo feminino, no canto superior direito, nos indica que as mulheres são menos propensas a contraírem TB na população em estudo, já que essa área do mapa corresponde ao *cluster* de

Dor Torácica				
	Não		Sim	
TB -	323		579	
TB +	76		166	
Cluster Baixo Risco				
	Não	%	Sim	%
TB -	98	30,34	254	43,87
TB +	12	15,79	20	12,05
Cluster Medio Risco				
	Não	%	Sim	%
TB -	188	58,20	109	18,83
TB +	31	40,79	14	8,43
Cluster Alto Risco				
	Não	%	Sim	%
TB -	37	11,46	216	37,31
TB +	33	43,42	132	79,52

Tabela 5.13: Distribuição da Dor Torácica entre os *clusters*

baixo risco. Já nos casos em que o paciente é do sexo masculino, existe uma concentração em toda lateral esquerda do mapa, que representa o *cluster* de médio risco e parte do *cluster* de alto risco.

Na tabela 5.14, pode-se observar, de forma quantitativa, essa relação entre o sexo do paciente e os *clusters*. Apesar da base de dados apresentar uma proporção entre homens e mulheres quase igualitária, nos pacientes que apresentaram TB essa relação praticamente dobra. Podemos ver que no *cluster* de baixo risco, há uma predominância de pacientes do sexo feminino, sendo que 57,56% (278/483) das mulheres que não apresentaram TB estão nesse *cluster*, e uma presença muito pequena de homens sem TB, apenas 17,66% (74/419). No *cluster* de médio risco, temos a predominância dos pacientes do sexo masculino, sendo que 48,45% (203/419) dos pacientes homens nesse *cluster* não apresentaram TB. Já no *cluster* de alto risco, existe uma leve concentração dos casos do sexo masculino; porém, entre os pacientes que foram diagnosticados com TB, a proporção desses casos, em relação ao sexo, é muito parecida para homens e mulheres 68,15% e 68,24%,respectivamente, e nos casos sem TB o percentual de homens é maior do que de mulheres o que justifica a coloração do plano da componente sexo na região referente a este *cluster*. Logo, após a comparação gráfica do plano da componente e da analise da tabela de distri-

buição do sexo pelos *clusters*, podemos perceber que os homens são mais propensos a contraírem TB, o que é uma característica importante para a clusterização.

Sexo				
	Homem		Mulher	
TB -	419		483	
TB +	157		85	
Cluster Baixo Risco				
	Homem	%	Mulher	%
TB -	74	17,66	278	57,56
TB +	11	7,01	21	24,71
Cluster Medio Risco				
	Homem	%	Mulher	%
TB -	203	48,45	94	19,46
TB +	39	24,84	6	7,06
Cluster Alto Risco				
	Homem	%	Mulher	%
TB -	142	33,89	111	22,98
TB +	107	68,15	58	68,24

Tabela 5.14: Distribuição dos Sexos entre os *clusters*

Sexo			
Clusters			
	Baixo Risco	Médio Risco	Alto Risco
Mulheres	7,02%	6,00%	34,32%
Homens	19,94%	16,12%	42,97%

Tabela 5.15: Prevalência por Sexo nos casos TB positivos por *clusters*

Ao analisarmos o Plano do Componente Tabagismo, figura 5.4 (b), vemos que sua distribuição é semelhante a do componente Sexo, com uma leve concentração de pacientes fumantes em toda a lateral esquerda do plano, dando indícios de que essas duas variáveis podem ser estatisticamente correlacionadas.

Entretanto, essa variável apresenta casos ignorados, quando o paciente não declarara sua posição em relação à pergunta feita na triagem, como pode ser visto na tabela 5.16. Entretanto, o número de casos com status ignorado é muito pequeno em relação ao total de casos na base de dados, fato que não interfere na clusterização. No *cluster* de baixo risco, encontram-se 61,93% dos casos sem TB, o que justifica a coloração azul no canto superior direito do plano de componente. Já no *cluster* de médio risco, vemos uma concentração de pacientes sem TB e que se declararam

fumantes, fazendo com que 45% dos pacientes que se declararam fumantes sejam atribuídos a esse *cluster*. No *cluster* de alto risco, existem duas concentrações de casos bem distintas: há uma concentração dos pacientes que se declararam fumantes e diagnosticados com TB contendo 64,83% desses casos, e outra concentração que contém 37,34% de todos os pacientes que se declararam não fumante. Essas duas concentrações distintas, no *cluster* de alto risco, justificam a divisão da região correspondente ao *cluster* de alto risco no Plano da Componente Tabagismo. Portanto, devido à grande concentração dos pacientes que se declararam fumantes nos *clusters* de médio e alto risco, 81,31% de todos os fumantes, pode-se inferir que o Tabagismo aumenta o risco de se contrair TB.

Tabagismo						
	Não		Ignorado		Sim	
TB -	373		16		513	
TB +	93		4		145	
Cluster Baixo Risco						
	Não	%	Ignorado	%	Sim	%
TB -	231	61,93	8	50	113	22,02
TB +	21	22,58	1	25	10	6,90
Cluster Médio Risco						
	Não	%	Ignorado	%	Sim	%
TB -	37	9,92	5	31,25	255	49,70
TB +	3	3,22	1	25	41	28,28
Cluster Alto Risco						
	Não	%	Ignorado	%	Sim	%
TB -	105	28,15	3	18,75	145	28,27
TB +	69	74,19	2	50	94	64,83

Tabela 5.16: Distribuição do Tabagismo entre os *clusters*

A análise gráfica Plano do Componente Internação Hospitalar, figura 5.4 (c), não traz muita informação do relacionamento entre essa condição clínica do paciente e os *clusters* referentes aos riscos de se ter TB. O Plano apresenta uma pequena concentração dos pacientes que declararam ter sofrido internação hospitalar no canto superior direito, região que correspondente ao *cluster* de baixo risco, e na parte inferior do plano correspondente ao *cluster* de alto risco. Já os pacientes que não sofreram internação estão, de uma forma geral, espalhados por todo o plano com uma pequena concentração no canto superior esquerdo, área correspondente ao *cluster* de

médio risco. Devido ao fato de não haver um padrão bem definido através da análise gráfica do relacionamento da variável com os *clusters*, tais padrões podem ser melhor estudado através da tabela 5.17.

Primeiramente, notamos a grande diferença que há na base de dados entre os pacientes que não declararam ter sofrido internação hospitalar em relação aos que sofreram, essa proporção é de aproximadamente seis vezes. Também nota-se pacientes que não declararam nada a respeito, mas o número de ignorados e muito pequeno em relação ao conjunto todo, o que não altera a clusterização. Ao analisarmos os casos que não declararam ter sofrido internação, vemos que os casos em que os pacientes que não foram diagnosticados com TB se espalham de maneira quase que uniforme entre os três *clusters* o que justifica a predominância das cores mais claras e frias, que representa a ausência do sintoma ou da condição, no Plano da Componente Internação. A concentração de casos que apresentam o sintoma no canto superior direito do mesmo plano, região do *cluster* de baixo risco, é devido à concentração dos 41,80% (51/122) dos casos em que o paciente declarou ter sofrido internação hospitalar e não tem TB. Já no *cluster* de alto risco, 65,70% (159/242) dos casos que foram diagnosticados com TB, e entre os pacientes que declaram ter sofrido internação e foram diagnosticados com TB, 69,7% destes casos, se encontram neste *cluster*. A informação sobre o paciente ter sofrido internação hospitalar é muito importante porque as pessoas que sofreram uma internação hospitalar têm mais chances de estar em contato com diversos agentes transmissores de doenças, principalmente o *Mycobacterium Tuberculosis*. Entretanto, colher essa informação é difícil; uma vez que, para considerar que uma pessoa sofreu internação hospitalar a mesma deveria ter ficado internada em alguma unidade de saúde por no mínimo 24 horas nos últimos dois anos.

Baseado na análise das distribuições dos sintomas e condições, na análise gráfica dos planos de componentes foi proposto um escore inicial de forma que representassem o relacionamento encontrado entre os sintomas com o diagnóstico da TB pulmonar.

Internação Hospitalar						
	Não		Ignorado		Sim	
TB -	756		24		122	
TB +	202		7		33	
Cluster Baixo Risco						
	Não	%	Ignorado	%	Sim	%
TB -	289	38,23	12	50,00	51	41,80
TB +	27	13,37	1	14,29	4	12,12
Cluster Medio Risco						
	Não	%	Ignorado	%	Sim	%
TB -	257	33,99	8	33,33	32	26,23
TB +	38	18,81	1	14,29	6	18,18
Cluster Alto Risco						
	Não	%	Ignorado	%	Sim	%
TB -	210	27,78	4	16,67	39	31,97
TB +	137	67,82	5	71,43	23	69,70

Tabela 5.17: Distribuição do Internação Hospitalar entre os *clusters*

O escore proposto, representado na tabela 5.18, pontua tanto na presença quanto na ausência dos sintomas ou condições. Apesar de este escore ter sido proposto de forma empírica, por pessoas sem experiência na área clínica, obteve uma curva ROC levemente abaixo da curva do escore de referência, conforme visto na figura 5.8, com sensibilidade de 79,75% e especificidade de 49,33%.

SINTOMAS	PONTOS	
	SIM	NÃO
Idade $\leq$ 35 anos	2	X
Idade $>$ 35 anos	0	X
Tosse	1	-2
Hemoptise	0	0
Sudorese	3	1
Febre	3	1
Emagrecimento	1	-1
Dispnéia	0	0
Tabagismo	0	0
Internação Hospitalar	1	0
Homens	0	X
Mulheres	0	X
Dor Torácica	2	0
TOTAL	Suspeito de TB $\geq$ 6 pontos	Não TB $<$ 6 pontos

Tabela 5.18: Escore baseado no SOM

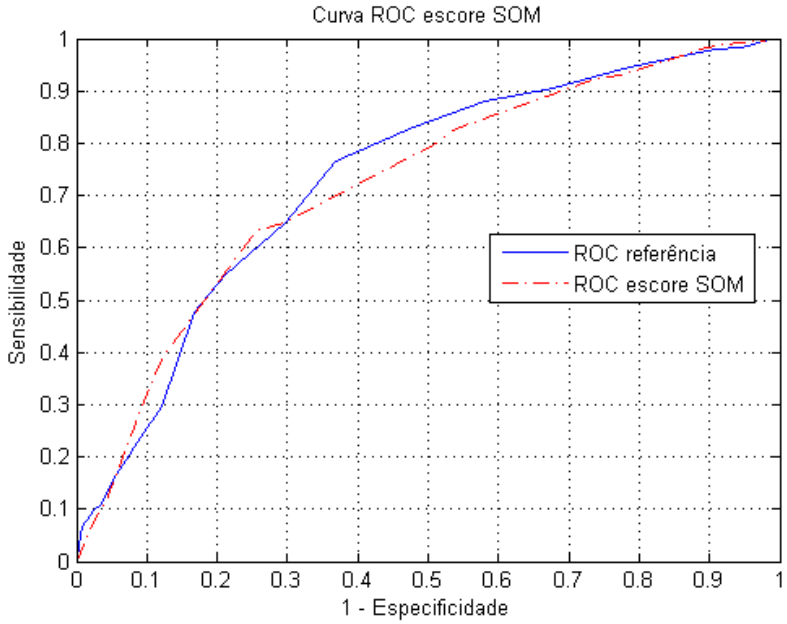


Figura 5.8: Curva ROC escore baseado no SOM

5.2 Escore

Nesta seção, serão apresentados os resultados obtidos durante o desenvolvimento do escore que é responsável por pontuar os sintomas dos pacientes atendidos na Policlínica Augusto Amaral Peixoto e, assim, atribuir uma pontuação que possa auxiliar o serviço médico na triagem e diagnóstico dos pacientes.

5.2.1 Discriminante de Fisher

Conforme descrito na seção 4.3.1, o escore por Discriminante de Fisher foi desenvolvido usando o método da validação cruzada, que utiliza um conjunto de treino e outro de teste do treinamento. Nas cem rodadas de treinamento realizadas, o índice SP no conjunto de teste variou conforme a tabela 5.19. O escore que rendeu o SP máximo no conjunto de teste, a princípio, deveria ser escolhido como o melhor escore.

SP_{min}	SP_{max}	SP_{medio}	SP_{RMS}
0,6587	0,7568	0,7107	0,0247

Tabela 5.19: Variação do índice SP no conjunto de teste

No entanto, como os conjuntos de treino e teste são formados por sorteio dos

pacientes que têm ou não TB, e devido ao fato de as classes serem sobrepostas, conforme visto na figura 3.3, o valor máximo do SP no conjunto de teste da validação cruzada pode ser fruto de um bom sorteio do conjunto em que o escore foi testado. Devido a essa característica do problema e o fato de o escore de referência ter sido desenvolvido usando todos os casos da base de dados, o critério de escolha do escore, foi o discriminante com valores arredondados para números inteiros que obteve o maior índice SP na classificação de todos os casos contidos na base de dados. O desempenho deste teste pode ser visto na tabela 5.20.

SP_{min}	SP_{max}	SP_{medio}	SP_{RMS}
0,6914	0,7288	0,7084	0,0075

Tabela 5.20: Variação do índice SP com pesos dos escore arredondados utilizando todos os casos da base de dados

Os pesos do escore para os sintomas clínicos que geraram o valor do SP máximo e o limiar de decisão estão representado na tabela 5.21. Ao analisarmos os pesos vemos que são atribuídos pesos negativos à Idade, Dispnéia e Sexo . No caso da Idade, como é utilizada a idade normalizada entre 0 e 1, as pessoas mais jovens tendem o escore a ficar mais próximo do zero , logo, com mais chances de serem TB positivas. Como o Sexo do paciente é representado como -1 para homens e 1 para mulheres, os pacientes homens irão somar mais pontos ao escore tendendo-o a ficar maior que o limiar de corte. Já no caso da Dispnéia, o peso negativo dado ao sintoma no escore pode ser atribuída ao arredondamento do peso no discriminante, uma vez que a Dispnéia é uma variável confundidora, assim como o Tabagismo, a Internação Hospitalar e a Dor Torácica, que foram atribuídas peso zero para esses sintomas.

Ao analisarmos a curva ROC desse escore, figura 5.9, vemos que a curva ROC do escore por Discriminante de Fisher é sempre superior a ROC do escore de referência indicando que, de forma geral, o resultado obtido por esse escore é melhor que o do escore de referência.

Analisando, no gráfico, a faixa de interesse para a comparação entre os escores, que foi escolhida segundo um painel de médicos especialistas em pneumologia, na

SINTOMAS	PONTOS	
	SIM	NÃO
Idade Normalizada	-9	0
Tosse	1	0
Hemoptise	1	0
Sudorese noturna	2	0
Febre	1	0
Emagrecimento	3	0
Dispnéia	-1	0
Tabagismo	0	0
Internação Hospitalar	0	0
Sexo	-1	0
Dor Torácica	0	0
TOTAL	Suspeito de TB ≥ -7 pontos	Não TB $<$ -7 pontos

Tabela 5.21: Escore por Discriminante de Fisher

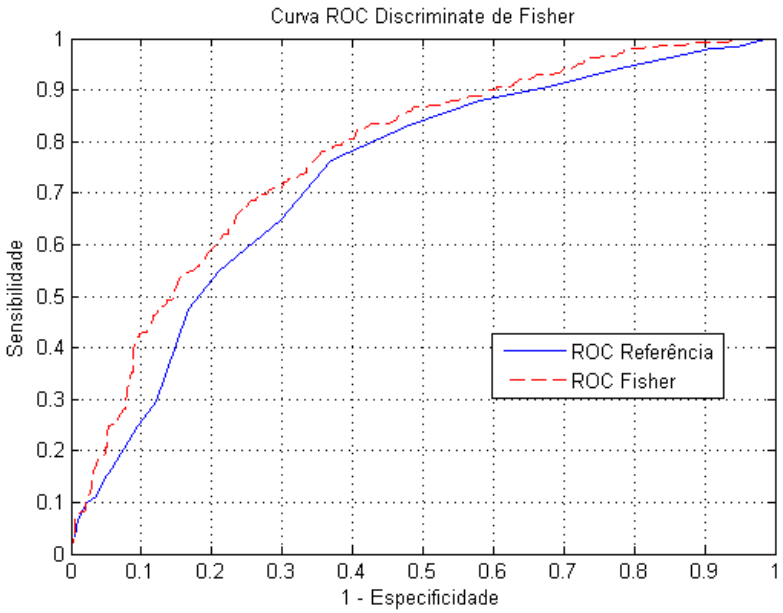


Figura 5.9: Curva ROC escore por Discriminante de Fisher

região em que a sensibilidade varia de 80% a 90%, as curvas estão muito próximas. Entretanto, o escore por Discriminante de Fisher tem um desempenho um pouco melhor, uma vez que o mesmo é mais especifico do que o escore de referência, para uma mesma sensibilidade.

Já o limiar de corte foi definido como o valor do escore que obtivesse sensibilidade de aproximadamente 80%, que no caso desse escore o limiar de corte é -7, fazendo com que o classificador tenha sensibilidade de 81,82% e especificidade de 56,76%.

5.2.2 Simulated Annealing

Foram desenvolvidos diferentes escores utilizando a técnica de Simulated Annealing, segundo os modelos descritos na seção 4.3.3. Como, para cada cenário, o valor máximo dos pesos, em módulo, de cada escore pode variar entre 1 e 15, o critério de escolha do melhor escore para cada cenário se deu da seguinte forma: foram avaliados os valores da especificidade para os valores de sensibilidade próximos de 80% e 90% e o valor máximo dos pesos. Os escores escolhidos foram aqueles que apresentaram os maiores valores de especificidade para os valores de sensibilidade desejados, priorizando aqueles para os quais os pesos sejam os menores possiveis (facilidade de cálculo).

5.2.2.1 Modelo 1 - Pontuação para Sintomas Presentes e Ausentes

Na tabela 5.22, podemos ver os resultados dos escores desenvolvidos para esse modelo em relação aos valores máximos que os pesos do escore podem atingir em módulo. Segundo o critério de escolha do escore vemos que para a sensibilidade próximo a 80% os maiores valores de especificidade foram para os escores com valores máximos de 4, 6 , 8 e 13. Analisando, para esses valores, a região de sensibilidade próximo a 90% pode-se ver que o escore que pontua os sintomas até 13 pontos teve um desempenho de especificidade melhor que os outros.

Ao analisarmos o escore escolhido, tabela 5.23, podemos ver que algumas variáveis são mais influentes na classificação, como a idade, tosse, hemoptise, ema-

Valor Máximo dos pesos	Sensibilidade ≈ 80%		Sensibilidade ≈ 90%	
	Sensibilidade	Especificidade	Sensibilidade	Especificidade
1	85,12%	49,56%	95,87%	19,62%
2	84,30%	51,88%	92,98%	28,82%
3	81,40%	56,87%	91,74%	28,71%
4	80,17%	60,31%	90,08%	35,25%
5	80,17%	59,31%	90,91%	36,70%
6	80,99%	61,53%	90,08%	35,59%
7	81,40%	58,20%	90,91%	32,71%
8	80,17%	61,86%	90,50%	33,59%
9	80,58%	57,32%	90,08%	41,46%
10	80,58%	55,54%	90,08%	35,59%
11	80,17%	58,65%	90,50%	30,93%
12	81,40%	57,98%	91,32%	37,69%
13	**80,17%**	**61,75%**	**90,08%**	**41,57%**
14	80,17%	57,76%	92,15%	31,49%
15	80,58%	59,42%	90,91%	37,03%

Tabela 5.22: Desempenho dos Escores por Simulated Annealing do modelo 1

grecimento e o sexo do paciente. Podemos ver que os pacientes jovens terão uma pontuação maior, tendendo a serem classificados como suspeitos de TB, já os pacientes acima de 35 anos pontuam muito negativamente no escore, fazendo que os mesmos tenham menos chances de serem suspeitos de TB. A pontuação atribuída à Tosse , Hemoptise e Emagrecimento vão de acordo com a importância desses sintomas no diagnóstico médico da TB, são pontuadas fortemente na presença do sintoma ou na ausência do sintoma; assim como o sexo do paciente, onde pacientes homens são pontuados positivamente e mulheres negativamente, fazendo que o escore varie muito com esses sintomas mostrando a relevância dos mesmos.

Entretanto, também podemos ver a influência dos sintomas confundidores no escore, como a Dor Torácica, Dispnéia e Tabagismo, que pontuam fortemente de forma positiva tanto na presença quanto na ausência do sintoma, assim como a Internação Hospitalar, onde foi atribuído um pesos negativo para a presença do sintoma.

Como o processo de desenvolvimento do escore foi feito para maximizar o índice SP do mesmo, o limiar de corte ideal seria 20 pontos, pois o mesmo teria sensibilidade de 70,25% e especificidade de 74,61% que é o ponto onde se tem o SP máximo do

SINTOMAS	PONTOS	
	SIM	NÃO
Idade $\leq$ 35 anos	9	X
Idade $>$ 35 anos	-12	X
Tosse	12	1
Hemoptise	9	-2
Sudorese	0	-11
Febre	2	-13
Emagrecimento	8	-12
Dispnéia	8	13
Tabagismo	13	11
Internação Hospitalar	-6	1
Homens	7	X
Mulheres	-10	X
Dor Torácica	5	5
TOTAL	Suspeito de TB $\geq$ 10 pontos	Não TB $<$ 10 pontos

Tabela 5.23: Escore por Simulated Annealing para o modelo 1 usando a Presença e a Ausência dos sintomas

classificador. Entretanto, o limiar de corte expresso na tabela 5.23 foi definido através da análise da curva ROC, figura 5.10, para que o limiar de corte do escore obtivesse sensibilidade de aproximadamente 80% . Logo, quando o somatório de pontos do escore for maior ou igual a 10 se tem uma sensibilidade de 80,17% e especificidade de 61,75%, com um $SP = 0,7066$.

5.2.2.2 Modelo 2 - Pontuação para Sintomas Presentes, Ausentes e Ignorados

O escore do Modelo 2 pode pontuar positivamente ou negativamente a resposta do paciente ao questionário de triagem, podendo o paciente declarar o sintoma presente, ausente ou ignorado. Sendo assim, o escore desenvolvido para este modelo é mais abrangente que o do Modelo 1.

Na tabela 5.24, podemos ver os resultados dos escores desenvolvidos para esse modelo em relação aos valores máximos que os pesos do escore podem atingir em módulo. Segundo o critério de escolha do escore estabelecido, vemos que para a sensibilidade próximo à 80% os maiores valores de especificidade foram para os escores com valores máximos dos pesos de 3, 7, 8, 9, 10 e 11. Para esses conjunto

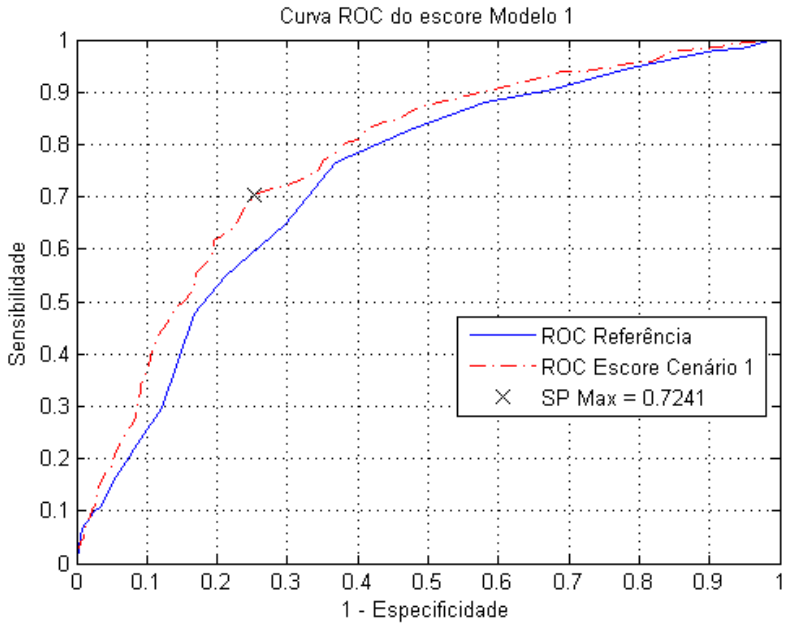


Figura 5.10: Curva ROC escore por Simulated Annealing modelo 1

Valor Máximo	Sensibilidade ≈ 80%		Sensibilidade ≈ 90%	
dos pesos	Sensibilidade	Especificidade	Sensibilidade	Especificidade
1	88,02%	39,80%	99,17%	8,76%
2	83,88%	44,35%	94,21%	22,28%
3	80,99%	60,64%	90,91%	32,15%
4	80,17%	56,76%	90,50%	38,91%
5	81,40%	55,21%	92,56%	28,60%
6	80,17%	59,53%	90,91%	30,38%
7	80,99%	58,54%	90,50%	38,03%
8	**80,99%**	**61,09%**	**90,08%**	**38,69%**
9	80,58%	60,64%	90,50%	40,24%
10	80,58%	61,20%	90,08%	39,02%
11	80,58%	60,42%	90,91%	38,03%
12	80,99%	55,99%	91,74%	32,71%
13	80,58%	55,54%	90,91%	39,91%
14	80,99%	56,43%	90,50%	35,25%
15	80,17%	59,53%	90,50%	37,25%

Tabela 5.24: Desempenho dos Escores por Simulated Annealing do modelo 2

de valores, quando analizada a região de sensibilidade próximo a 90%, pode-se ver que a especificidade varia muito pouco, com exceção do escore com pesos até 3, variando de 38,03% a 40,24%, que significa uma diferença, entre a menor e a maior especificidade, de 20 pacientes, que são TB negativas, que não serão consideradas suspeitas de TB para esse nível sensibilidade. Os resultados para os escores com valores ate 10 e 8 são muito próximos; entretanto, a escolha do escore, com pesos de valor máximo em modulo igual a 8, como o melhor classificador foi devido a ele ter tido a maior especificidade para a região de sensibilidade de aproximadamente de 80%, sensibilidade de 80,99% e especificidade de 61,09%, apesar de este escore ter especificidades menor que o escore com valores até 10 para a faixa de sensibilidade a 90%. Também foi leavdo em consideração o fato do escore, com pesos de valor até de 8 pontos, ser formado por números menores logo sendo mais fácil contabilização do total.

SINTOMAS	PONTOS		
	SIM	NÃO	IGNORADO
Idade $\leq$ 35 anos	2	X	X
35 $<$ Idade $\leq$ 65 anos	-5	X	X
Idade $>$ 65 anos	-6	X	X
Tosse	4	-7	-4
Hemoptise	7	-8	-5
Sudorese	-2	-8	-1
Febre	5	-8	0
Emagrecimento	7	-7	-7
Dispnéia	-3	-4	-3
Tabagismo	-6	-5	0
Internação Hospitalar	-5	6	7
Homens	4	X	X
Mulheres	-7	X	X
Dor Torácica	-7	3	7
TOTAL	Suspeito de TB ≥ -30 pontos	Não TB $<$ -30 pontos	

Tabela 5.25: Escore por Simulated Annealing para o modelo 2 usando a Presença,Ausência e Abstenção dos sintomas

Ao analisarmos o escore escolhido, tabela 5.25, podemos ver que as variáveis mais influentes na classificação são a idade, tosse, hemoptise, febre, emagrecimento e o sexo do paciente. Podemos ver que os pacientes jovens terão uma pontuação positiva

enquanto as outras faixas etárias são pontuadas negativamente, portanto pessoas jovens tendem a serem classificados como suspeitos de TB. A pontuação atribuída a Tosse , Febre, Hemoptise e Emagrecimento vão de acordo com a importância desses sintomas no diagnóstico médico da TB, são pontuadas fortemente na presença do sintoma ou na ausência do sintoma, assim como o sexo do paciente, onde homens são pontuados positivamente e mulheres negativamente, fazendo que o escore varie muito com esses sintomas mostrando a relevância dos mesmos.

Entretanto, vemos a Sudorese, que nos outros escores desenvolvidos fora pontuada fortemente, no escore deste modelo a sua pontuação é negativa até mesmo quando o sintoma se faz presente fazendo com que esse peso não tenha sentido clínico como as outras pontuações. Também podemos ver a influência dos sintomas confundidores no escore, como a Dispnéia e o Tabagismo, que pontuam fortemente de forma negativa tanto na presença quanto na ausência do sintoma, assim como a Internação Hospitalar que pontua negativamente a presença e positivamente a ausência do sintoma, semelhante ao escore do Modelo 1.

Como esse modelo contempla a pontuação dos sintomas caso o paciente se abstenha de responder a pergunta, podemos ver que de uma forma geral os pontos dados aos sintomas ignorados seguem a tendência da pontuação dada para os casos negativos. No entanto, na base de dados somente dois sintomas continham casos com ignorados, Tabagismo e Internação Hospitalar. Porém, a pontuação dada a esses casos não acrescentaram pontos significativos ao escore, já que no caso do Tabagismo é dada pontuação zero e na Internação Hospitalar a pontuação e quase a mesma que quando o paciente declara que não foi internado.

Ao analisarmos a curva ROC, figura 5.11, vemos que a mesma segue a curva de referência para valores de sensibilidade entre 0 e 55% e entre 90% e 100%, porém na região entre 55% e 90 %, a curva ROC do escore do Modelo 2 está bem afastada da referência, tendo o limiar de corte de -24 pontos, quando se tem o máximo do índice SP, com sensibilidade de 72.31% e especificidade de 73.84% neste ponto.

Entretanto, o limiar de corte expresso na tabela 5.25 foi escolhido de forma que

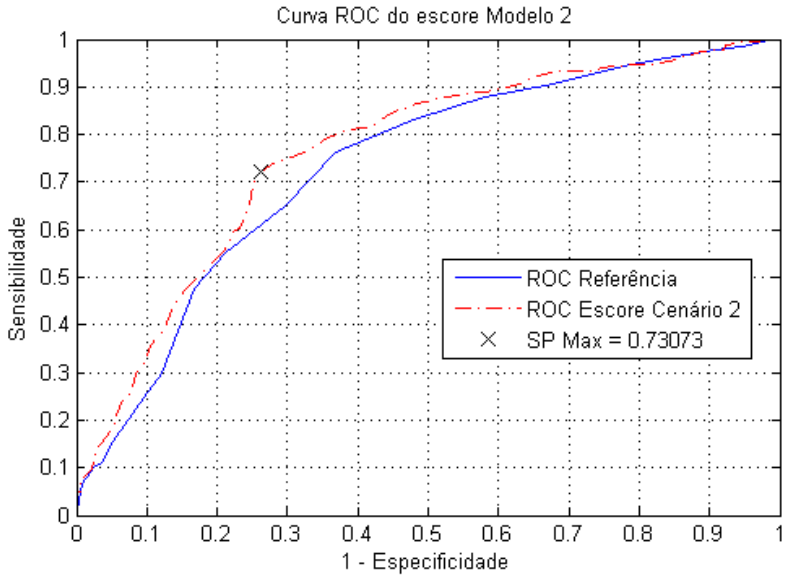


Figura 5.11: Curva ROC escore por Simulated Annealing modelo 2

escore obtivesse sensibilidade de aproximadamente 80% . Logo, quando o somatório de pontos do escore for maior ou igual a -30 tem-se um classificador com uma sensibilidade de 80,99% e especificidade de 61,09%, com um $SP = 0,7069$.

5.2.2.3 Modelo 3 - Pontuação para Sintomas Presentes

Para o Modelo 3, foram desenvolvidos dois escores distintos. Um deles pontua positivamente ou negativamente, enquanto o segundo pontua somente positivamente os sintomas que os pacientes declaram presentes. Sendo estes modelos de escore mais restritos que os outros modelos.

Na tabela 5.26, podemos ver os resultados dos escores desenvolvidos para esse modelo podem pontuar positivamente ou negativamente a presença do sintoma. Segundo o critério de escolha do escore estabelecido, vemos que para a sensibilidade próximo a 80% os maiores valores de especificidade foram para os escores com valores máximos de 9, 11 e 13. No entanto para a região de sensibilidade próxima a 90% o escore com pesos até 9 se destaca mais com uma melhor especificidade, 37,25%.

Assim como nos outros modelos, o padrão de pontuação das variáveis mais influentes e as confundidoras na classificação se mantêm no escore do Modelo 3, como pode ser visto na tabela 5.27. Podemos ver que nesse modelo ambas as faixas de

Valor Máximo dos pesos	Sensibilidade ≈ 80%		Sensibilidade ≈ 90%	
	Sensibilidade	Especificidade	Sensibilidade	Especificidade
1	87,60%	35,81%	98,35%	8,09%
2	86,36%	45,34%	97,11%	12,86%
3	81,40%	55,42%	90,50%	33,59%
4	83,88%	51,33%	92,56%	34,15%
5	80,99%	53,42%	90,08%	39,58%
6	81,82%	58,31%	90,08%	37,47%
7	81,40%	56,54%	93,80%	35,70%
8	82,23%	53,99%	90,50%	37,03%
9	**82,23%**	**56,98%**	**91,32%**	**37,25%**
10	80,99%	55,88%	90,50%	36,70%
11	80,99%	58,65%	92,15%	27,27%
12	82,71%	53,88%	92,15%	32,04%
13	80,17%	61,64%	90,91%	33,59%
14	82,64%	55,99%	90,08%	35,70%
15	80,17%	56,65%	90,50%	33,81%

Tabela 5.26: Desempenho dos Escores por Simulated Annealing do modelo 3

idade pontuam positivamente e que os pacientes mais jovens pontuam de forma a terem mais chances de serem suspeitos de TB. Também pode-se ver que nesse modelo que sintomas relevantes, do ponto de vista clínico, como Tosse e Sudorese não têm pontuação tão alta quanto os outros sintomas relevantes. No entanto, como esse modelo só prevê a pontuação da presença do sintoma, os sintomas que são confundidores ficaram com pesos negativos fazendo diminuir as chances de o paciente ser suspeito de TB quando o mesmo declara a presença de um desses sintomas.

Esse escore gera uma curva ROC sempre maior que a respectiva curva para o escore de referência, como pode ser visto na figura 5.12, principalmente na região onde se encontra o SP máximo, sensibilidade de 72,31% e especificidade de 73,71%, onde o escore tem aproximadamente oito pontos percentuais a mais de especificidade do que a referência.

Já para o escore que utiliza somente pesos positivos os escores que apresentaram as maiores especificidades para a faixa de sensibilidade de 80% foram os quem tem pesos com valor máximo até 3 e 5, conforme visto na tabela 5.28. Ambos os escores tem resultados nas faixas de interesse muito parecidos, portanto a escolha do escore com pesos ate 3 se deu pelo fato de ser uma soma muito intuitiva e muito fácil de

SINTOMAS	PONTOS	
	SIM	NÃO
Idade $\leq$ 35 anos	8	X
Idade $>$ 35 anos	2	X
Tosse	4	X
Hemoptise	7	X
Sudorese	4	X
Febre	6	X
Emagrecimento	9	X
Dispnéia	-3	X
Tabagismo	-2	X
Internação Hospitalar	-5	X
Homens	4	X
Mulheres	-1	X
Dor Torácica	-2	X
TOTAL	Suspeito de TB $\geq$ 11 pontos	Não TB $<$ 11 pontos

Tabela 5.27: Escore por Simulated Annealing para o modelo 3 usando a Presença dos sintomas

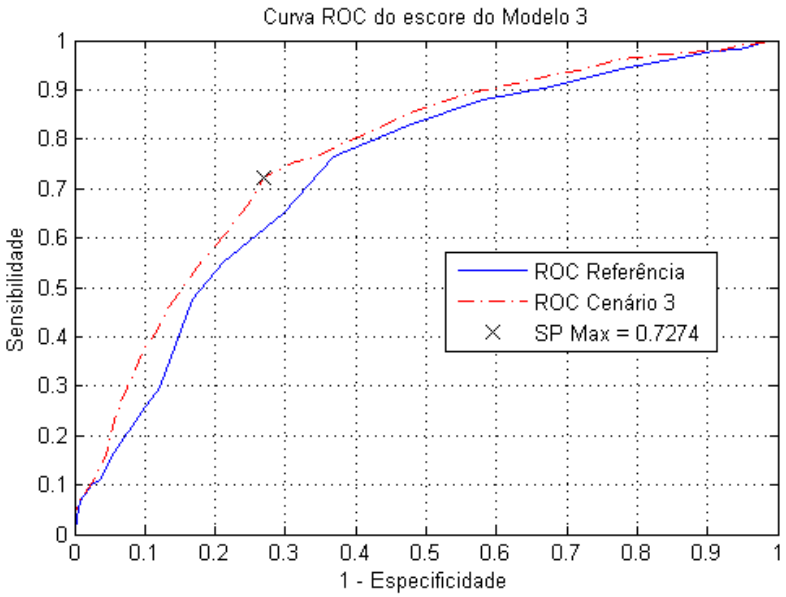


Figura 5.12: Curva ROC escore por Simulated Annealing modelo 3

ser feita.

Valor Máximo dos pesos	Sensibilidade ≈ 80%		Sensibilidade ≈ 90%	
	Sensibilidade	Especificidade	Sensibilidade	Especificidade
1	89,26%	40,24%	99,17%	5,99%
2	95,87%	21,73%	95,87%	21,73%
3	**81,40%**	**61,31%**	**92,15%**	**32,59%**
4	80,58%	60,53%	93,39%	22,39%
5	80,99%	61,42%	91,32%	32,04%
6	81,82%	57,10%	92,56%	21,18%
7	81,82%	59,20%	90,91%	36,03%
8	80,58%	56,65%	92,15%	27,16%
9	80,58%	55,32%	90,91%	31,26%
10	80,99%	59,87%	90,50%	31,82%
11	80,17%	57,43%	90,50%	35,03%
12	81,40%	58,20%	90,91%	29,38%
13	80,99%	58,65%	90,50%	25,06%
14	80,58%	59,76%	90,50%	38,80%
15	82,64%	58,54%	90,50%	35,59%

Tabela 5.28: Desempenho dos Escores somente com pesos positivos por Simulated Annealing do modelo 3

Ao se analisar a tabela 5.29, que traz os pontos do escore do Modelo 3 com pesos positivos, podemos ver que a pontuação desse escore também segue a tendência dos outros escores. Dando mais pontos aos sintomas de maior importância segundo os critérios clínicos e as variáveis confundidoras não pontuam . Também nota-se a importância do sexo do paciente nesse escore, onde para os pacientes homens é atribuída a pontuação máxima e as mulheres não pontuam. Outro fato interessante de o escore pontuar até 3 pontos e que pode-se fazer uma correlação entre os três *clusters* obtidos no processo de clusterização com as pontuações atribuídas ao sintomas. Se for feita a relação de pontos com os *clusters*, somente a Sudorese não tem uma correspondência exata com a clusterização, já que esse sintoma é discriminante para o *cluster* de alto risco e no escore foi atribuída a pontuação 1 para a presença do mesmo.

Esse escore gera uma curva ROC que segue a curva de referência, sendo maior que a referencia na faixa de sensibilidade entre 60% e 90%, como pode ser visto na figura 5.13. Já o limiar de corte de 8 pontos se deu pela especificação de sensibilidade de 80% já que o limiar de corte dado pelo SP máximo, limiar de 9 pontos, nos daria

SINTOMAS	PONTOS	
	SIM	NÃO
Idade $\leq$ 35 anos	3	X
Idade $>$ 35 anos	1	X
Tosse	2	X
Hemoptise	3	X
Sudorese	1	X
Febre	3	X
Emagrecimento	3	X
Dispnéia	0	X
Tabagismo	0	X
Internação Hospitalar	1	X
Homens	3	X
Mulheres	0	X
Dor Torácica	0	X
TOTAL	Suspeito de TB $\geq$ 8 pontos	Não TB $<$ 8 pontos

Tabela 5.29: Escore por Simulated Annealing para o modelo 3 usando a Presença dos sintomas e pesos positivos

uma sensibilidade de 76,45% e especificidade de 68,18% .

5.2.3 Escolha do Escore

Nesta seção será discutida a escolha do escore que será utilizado nas análises posteriores. A escolha do escore foi baseado na comparação dos resultados dos diferentes escores para sensibilidade de aproximadamente 80% .

A tabela 5.30 traz o desempenho dos escore em função da matriz de confusão dos mesmos. Onde VP são os casos Verdadeiro Positivos, FP os Falso Positivos, VN os Verdadeiros Negativos e FN os Falsos Negativos.

Todos os escores tiveram, no limiar de corte definido, uma sensibilidade menor que o escore de referência, sendo os escores por Discriminante de Fisher e os Por Simulated Annealing dos Modelos 2 e 3 com pesos positivos os que tiveram um número de pacientes classificados corretamente com TB mais próximos da referência. Entretanto, todos os escores foram mais específicos que a referência, sendo o escore desenvolvido por Discriminante de Fisher o de menor especificidade.

Como os objetivos deste trabalho e ter um escore de fácil utilização; logo, que

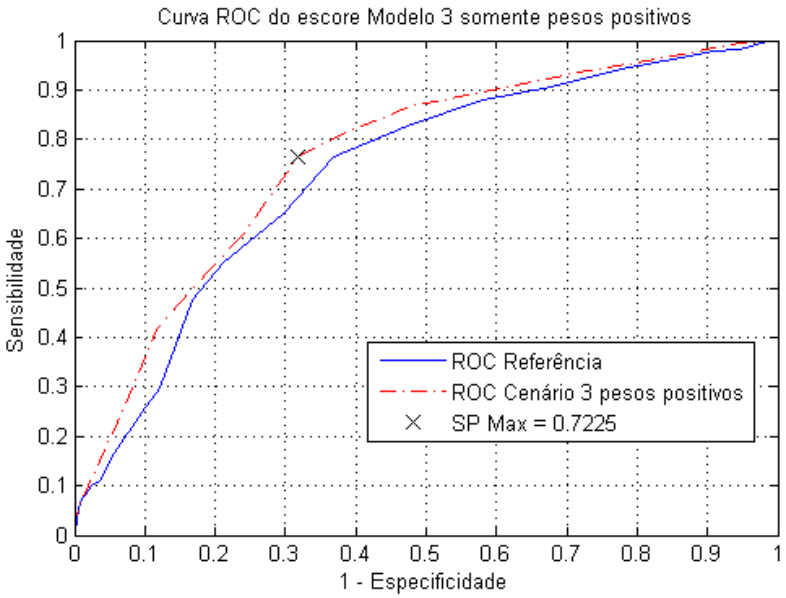


Figura 5.13: Curva ROC escore por Simulated Annealing modelo 3 com somente pesos positivos

Casos	TB negativos 902		TB positivos 242			
	Referência		Discriminate de Fisher		modelo 1	
	Casos	%	Casos	%	Casos	%
VP	201	83,06%	198	81,82%	194	80,17%
FP	433	48,00%	390	43,24%	345	38,25%
VN	469	52,00%	512	56,76%	557	61,75%
FN	41	16,94%	44	18,18%	48	19,83%
	modelo 2		modelo 3		modelo 3 pesos positivos	
	Casos	%	Casos	%	Casos	%
VP	196	80,99%	192	79,34%	197	81,40%
FP	351	38,91%	349	38,69%	349	38,69%
VN	551	61,09%	553	61,31%	553	61,31%
FN	46	19,01%	50	20,66%	45	18,60%

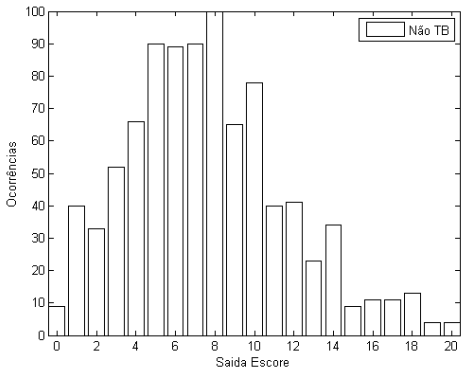
Tabela 5.30: Resultado dos Escores para Sensibilidade na faixa de 80%

não necessite de poder computacional, não se pode levar em consideração na escolha do melhor escore somente aquele que obteve o melhor desempenho numérico na classificação, mas também o impacto no serviço médico que o escore implicará.

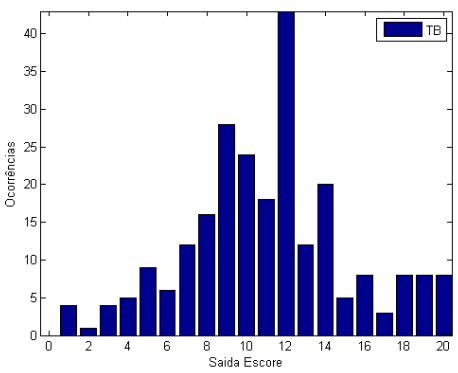
Partindo dessas premissas, o melhor escore foi o desenvolvido por Simulated Annealing, Modelo 3 pesos positivos. Do ponto de vista da sensibilidade, um teste de triagem quanto mais sensível melhor, já que mais pacientes portadores do bacilo e que irão desenvolver a TB serão atendidos, e o escore escolhido tem a segunda maior sensibilidade dos escores desenvolvidos. Já se tratando da especificidade, o escore escolhido é o que tem maior especificidade de todos os escores mostrados; portanto, o mesmo terá maior impacto na não disseminação da doença. Uma vez que se tratando de biossegurança na triagem dos pacientes, quanto maior a especificidade do escore menor são as chances de um paciente sem tuberculose ficar no mesmo ambiente de um paciente que ainda esta transmitindo o bacilo. A especificidade do escore também afeta diretamente o serviço médico já que com um menor numero de casos de Falso Alarme ocorre uma diminuição da carga de trabalho e de exames a serem realizados pelos médicos e enfermeiros, além do fato de o escore escolhido ser de simples utilização, pois somente se pontua positivamente e com pesos pequenos a presença do sintoma, semelhante ao escore de referência que hoje em dia é utilizado na unidade de saúde onde os dados foram coletados.

Analisando o histograma das saídas dos escores para os pacientes contidos na base de dados, figura 5.14, podemos ver que o escore escolhido separa melhor as classes. Logo, o escore escolhido tem uma acurácia maior que o escore de referência, 65,67% e 58,67%, respectivamente.

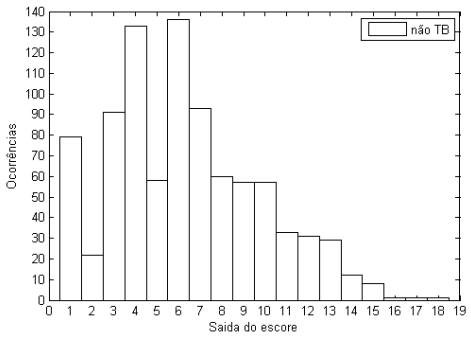
Ao comparamos as saídas dos dois escores, para os pacientes que não foram diagnosticados com TB, figura 5.15 (c), podemos ver que as variáveis utilizadas neste trabalho torna o melhor escore desenvolvido mais específico que o modelo de variável utilizado no escore de referência. Já que existe uma grande quantidade de casos no quadrante superior esquerdo, onde o escore de referência pontua acima do limiar de corte, representado pela linha preta, para suspeito de TB e no escore



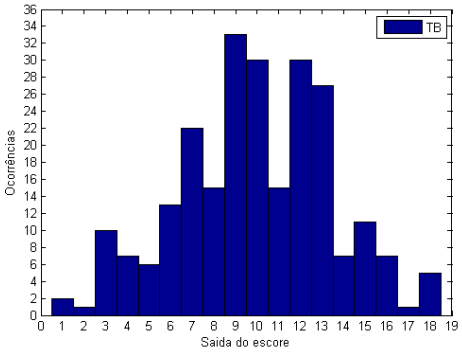
(a) Pacientes sem TB para a referência



(b) Pacientes com TB para a referência

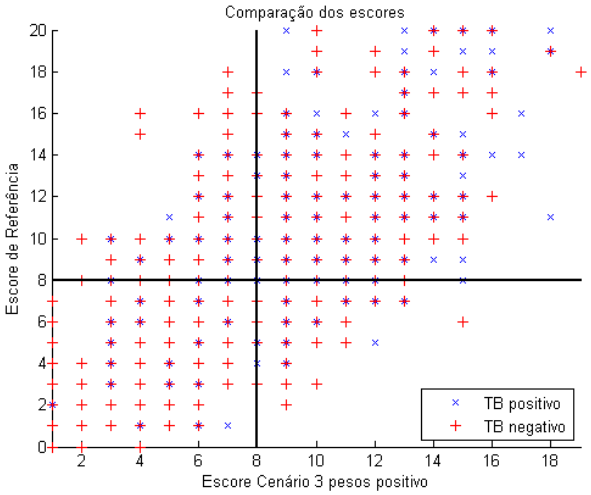


(c) Pacientes sem TB para o escore escolhido

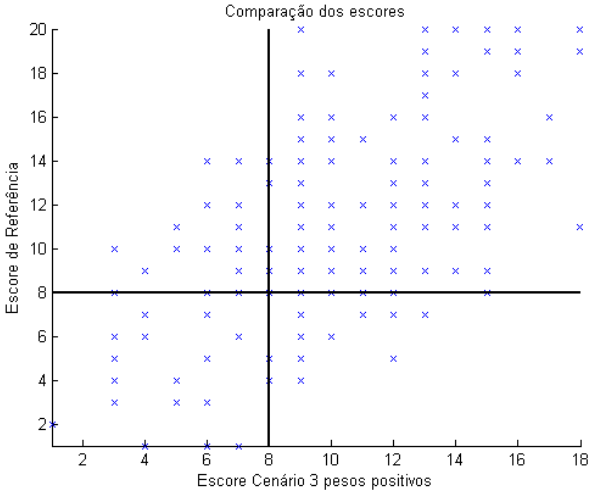


(d) Pacientes com TB para o escore escolhido

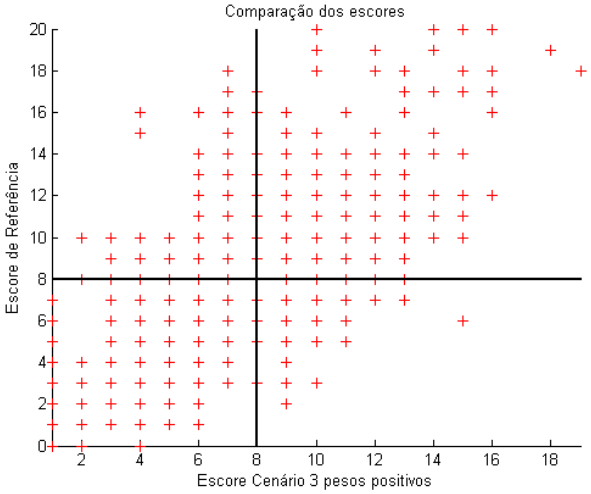
Figura 5.14: Comparativo dos histogramas das saídas do escore de referência e do melhor escore desenvolvido



(a) Todos os casos



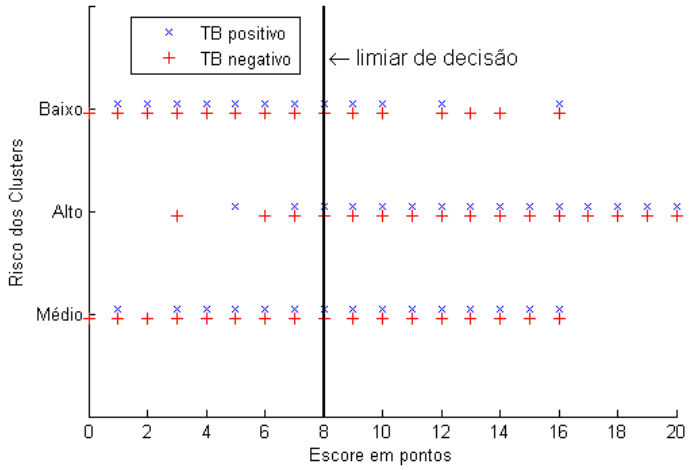
(b) Somente casos TB positivo



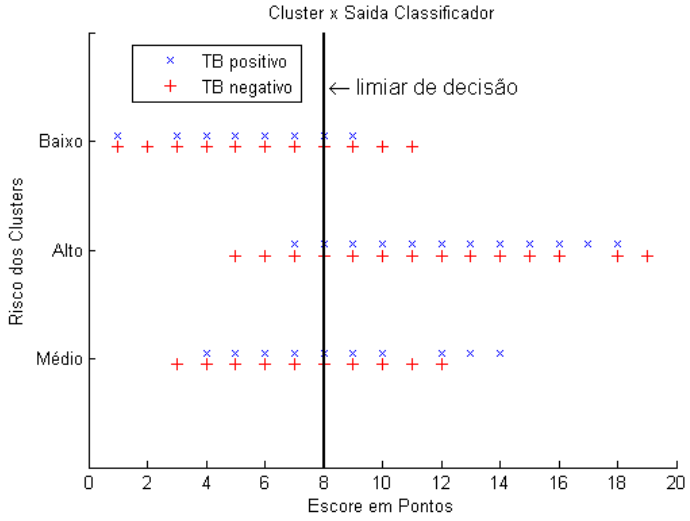
(c) Somente casos TB negativo

Figura 5.15: Comparação entre as saídas do escore de referência e do melhor escore desenvolvido

desenvolvido a pontuação do mesmo caso esta abaixo do respectivo limiar de corte. Tendo muitos desses casos obtido uma pontuação muito alta no escore de referência, acima de 12 pontos, enquanto no escore desenvolvido obtiveram uma pontuação entre 4 e 7 pontos.



(a) Clusters x Escore de Referência



(b) Clusters x Escore Desenvolvido

Figura 5.16: Comparação entre as saídas do escore de referência e do melhor escore desenvolvido com os grupos de risco

Outro ponto que podemos destacar é o relacionamento da saída dos escores com o resultado da clusterização utilizada para designar os grupos de risco. Quando comparado a saída do escore para um paciente com o *cluster* que o mesmo está associado, vê-se que o escore proposto gera grupos mais coesos e melhor delimitados do que na mesma comparação para o escore de referência, como pode ser visto na

figura 5.16. Indicando que a pontuação dada pelo escore tem relacionamento direto com o risco de o paciente ter TB.

Escore de Referência			
Risco	Baixo	Médio	Alto
Sensibilidade	46,88%	62,22%	95,76%
Especificidade	53,13%	37,78%	4,24%
Melhor Escore Desenvolvido			
Risco	Baixo	Médio	Alto
Sensibilidade	31,25%	68,89%	94,55%
Especificidade	68,75%	31,11%	5,45%

Tabela 5.31: Comparativo de desempenho por cluster dos escores

Considerando os limiares de decisão dos respectivos escores foi feita a análise de desempenho dos *clusters* em função da sensibilidade e especificidade, tabela 5.31. Mostrando que os pacientes que tiveram uma pontuação menor, no escore desenvolvido neste trabalho, têm de fato um menor risco de ter TB já que o seu desempenho para a especificidade é aproximadamente 15 pontos percentuais melhor que o escore de referência.

Também foi avaliado o desempenho dos escores para diferentes recomendações do tempo de tosse no diagnóstico da TB. O padrão utilizado no escore de referência e neste trabalho de tempo de tosse segue as recomendações da Organização Pan-Americana de Saúde (OPAS) onde pessoas com tosse por tempo igual ou superior a duas semanas são consideradas suspeitas de TB pulmonar. Entretanto, o Ministério da Saúde (MS) recomenda que pessoas com três semanas ou mais de tosse procurem uma unidade de saúde [67]. Logo, em cima desses critérios foi avaliado a distribuição dos pacientes atendidos segundo as duas recomendações, tabela 5.32, e a performance dos escores, tabela5.33.

Distribuição da Tosse				
	OPAS		MS	
	Não	Sim	Não	Sim
TB negativo	157	745	452	450
TB positivo	13	229	75	167

Tabela 5.32: Distribuição da Tosse segundo as diretrizes de diagnóstico da OPAS e do MS

Escore de Referência			
	OPAS	MS	
pontos	8	8	7
Sensibilidade	83,06%	79,75%	86,78%
Especificidade	58,00 %	57,87%	47,12%
Melhor Escore Desenvolvido			
	OPAS	MS	
pontos	8	8	7
Sensibilidade	81,40%	74,79%	83,88%
Especificidade	61,31%	67,85%	57,54%

Tabela 5.33: Performance dos escores para diferentes padrões de tempo de Tosse

Podemos ver pela distribuição dos casos que ao se utilizar o padrão do MS pode-se perder a chance de detectar pacientes em estado prematuro da doença e não se tem mais uma definição bem clara nos casos de não TB. Já que para os pacientes sem TB o padrão de tosse se distribui igualitariamente.

Ao se analisar o desempenho dos escores, caso se utilize o padrão de Tosse do MS e mantendo o mesmo limiar de decisão do escore original, desenvolvido com a recomendação da OPAS, ambos os escores perdem desempenho, sendo que o escore de referência é mais imune a essa mudança pois perdeu 3,31 pontos percentuais (p.p.) de sensibilidade e 0,13 p.p. de especificidade enquanto o escore desenvolvido neste trabalho perdeu 6,61 p.p. em sensibilidade e ganhou 6,54 p.p. em especificidade. Portanto, o escore de referência poderia ser utilizado normalmente para ambos os padrões com o mesmo limiar de corte. Já o escore desenvolvido, quando usado a recomendação do MS, poderia ser utilizado com o limiar de decisão de 8 pontos quando fosse desejado um teste mais especifico e com um limiar de decisão de 7 pontos a performance do mesmo se equipara a performance do escore de referência para o padrão OPAS.

5.3 Uso do Escore para Triagem e Diagnóstico

Após a definição do melhor escore desenvolvido se faz necessária a escolha dos limiares de decisão que atenda às condições impostas pelos modelos de triagem de pacientes e auxílio ao diagnostico médico da tuberculose, conforme descrito na seção

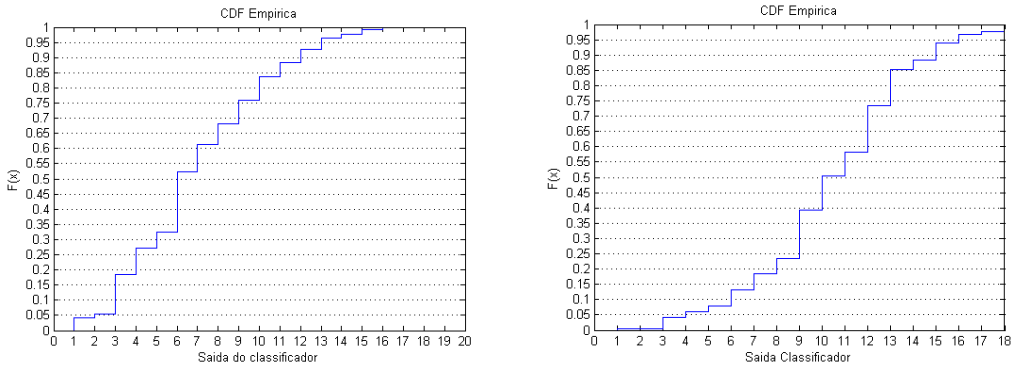
4.4.

Apesar de o limiar de decisão que melhor balanceia a sensibilidade e a especificidade ser de 8 pontos, pode-se definir dois limiares de corte distintos no mesmo escore. Um responsável pela triagem dos pacientes, excluindo do atendimento os pacientes que não tem risco de estarem contaminados com TB e ao mesmo tempo em que tenha a maior sensibilidade possível, e o outro responsável em auxiliar ao diagnóstico médico, onde é interessante uma alta especificidade para que o médico tenha confiança de que o paciente tem grandes chances de ter TB.

Portanto a escolha desses dois limiares de decisão foi baseada na análise dos quartis da saída do escore dados aos pacientes utilizados neste trabalho e nos pontos que compõem a curva ROC do escore. Através de uma inspeção visual da função de distribuição acumulada para as saídas do escore para os casos de TB e não TB, figura 5.17, podemos estipular os respectivos quartis das distribuições,que estão representados na tabela 5.34.

	TB Negativo	TB Positivo
1º Quartil	4	9
2º Quartil	6	10
3º Quartil	9	13

Tabela 5.34: Quartis dos casos de TB negativa e positiva do escore



(a) Pacientes TB Negativos (b) Pacientes TB Positivos

Figura 5.17: Função de distribuição acumulada da saída do escore

Vale notar, que o segundo quartil da saída do escore para os casos de não TB é um valor baixo de pontuação, e que o valor referente ao terceiro quartil é o mesmo

que o primeiro quartil da pontuação atribuída aos casos de TB, dando indício de que um corte próximo ao primeiro quartil dos casos de TB atenderia o modelo de triagem. Já para o modelo de auxilio ao diagnóstico, um limiar de decisão próximo do terceiro quartil dos casos de TB ajudaria ao médico na tarefa do diagnostico da TB já que o paciente com esse escore tem pelo menos 75% de chance de estar doente.

Entretanto, esses limiares de corte podem ser melhores definidos através da análise da curva ROC do escore, figura 5.13, tendo seus pontos mais detalhados na tabela 5.35. Como no modelo de triagem é importante que o máximo possível de pacientes que tenha TB receba atendimento, e ao mesmo tempo deve-se ter a preocupação de manter os pacientes que não tenham risco de terem TB fora do mesmo ambiente que os portadores do bacilo, evitando a exposição dos pacientes à doença. Portanto, o limiar de decisão de 7 pontos é o que melhor atende aos requisitos da triagem já que mais de 80% dos pacientes portadores da TB passarão no teste e um pouco mais da metade dos pacientes que não tem suspeita de TB não irão passar no teste.

Pontos da curva ROC						
Pontos	1	2	3	4	5	6
Sensibilidade	100,00%	99,59%	99,59%	95,87%	93,80%	92,15%
Especificidade	0,00%	4,10%	5,32%	18,63%	27,16%	32,59%
Pontos	7	8	9	10	11	12
Sensibilidade	86,78%	81,40%	76,45%	60,74%	49,59%	41,74%
Especificidade	52,33%	61,31%	68,18%	76,05%	83,59%	88,25%
Pontos	13	14	15	16	17	18
Sensibilidade	26,45%	14,88%	11,57%	6,20%	3,31%	2,48%
Especificidade	92,68%	96,34%	97,56%	99,22%	99,78%	99,78%

Tabela 5.35: Desempenho do escore por limiar de decisão

A partir da população que ficou acima do limiar de decisão da triagem, podemos estimar qual é o desempenho do escore para essa nova população, e decidir qual seria o melhor limiar de decisão para o modelo de diagnóstico médico. Portanto, verificamos na tabela 5.36 que o limiar de decisão com 14 pontos atende às necessidades do modelo do escore para diagnóstico por apresentar uma sensibilidade de 92,33%, logo poucos pacientes seriam diagnosticados erroneamente com tuberculose.

Pontos	8	9	10	11	12	13
Sensibilidade	93,81%	88,10%	70,00%	57,14%	48,10%	30,48%
Especificidade	18,84%	33,26%	49,77%	65,58%	75,35%	84,65%
Pontos	14	15	16	17	18	19
Sensibilidade	17,14%	13,33%	7,14%	3,81%	2,86%	0,00%
Especificidade	92,33%	94,88%	98,37%	99,53%	99,53%	100,00%

Tabela 5.36: Desempenho do escore com os pacientes que passaram na triagem

Outro fator importante é a análise do desempenho do escore em função do Valor Preditivo Positivo (VPP), que diz respeito à quantidade de pacientes que de fato foram diagnosticas com TB e que o modelo previu corretamente como suspeito de TB, e do Valor Preditivo Negativo (VPN), que é uma situação análoga a do VPP; porém, para os casos sem TB que que ficaram abaixo do limiar de corte. Com isso, podemos ver o comportamento do escore quando utilizado em outras populações com diferentes prevalências da TB na população. A tabela 5.37 traz os valores de VPP e VPN obtidos para populações com prevalência de 5%, 10%, 15% e 21,5%; na qual, 21,5% é a prevalência de TB, dos casos atendidos na unidade de saúde em que os dados foram coletados.

Podemos ver que os limiares escolhidos, para a população em estudo, foram satisfatórios, tendo em vista que no modelo de triagem um VPP de 32,81% é satisfatório, pois que em cada 3 pacientes que passaram na triagem 1 foi diagnosticado com TB pulmonar. Já para o modelo de diagnóstico, de cada 2 pacientes que tem a pontuação acima ou igual ao limiar, 1 tem TB, dando uma certa garantia ao corpo médico de começar alguns procedimentos de tratamento da doença sem a necessidade de esperar os resultados de alguns exames mais demorados como o baciloscopia.

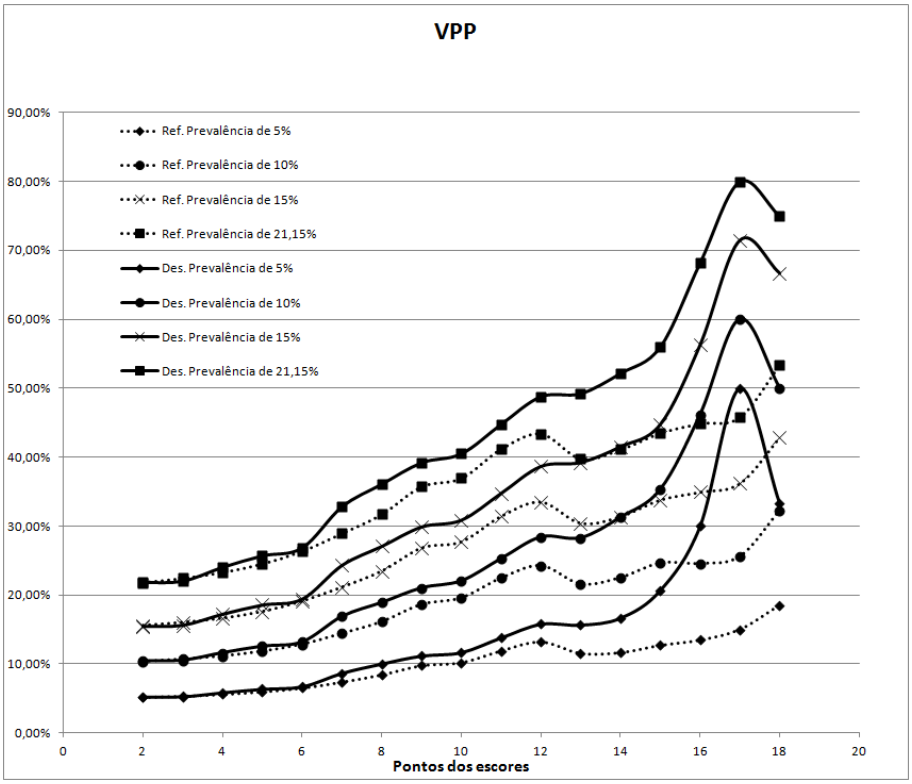
Porém, para localidades com prevalência baixa da TB, pode-se ver que o escore não tem um desempenho tão bom quanto o obtido para a população em estudo. Portanto para essas populações não se pode segmentar a pontuação do escore em dois limiares de decisão. E para populações com prevalência muito baixa como a de 5% o escore perde muito poder de presunção dos casos de TB positiva, inviabilizando o seu uso nessa prevalência.

Para efeito de comparação, podemos ver na figura 5.18 a, que a partir de 7

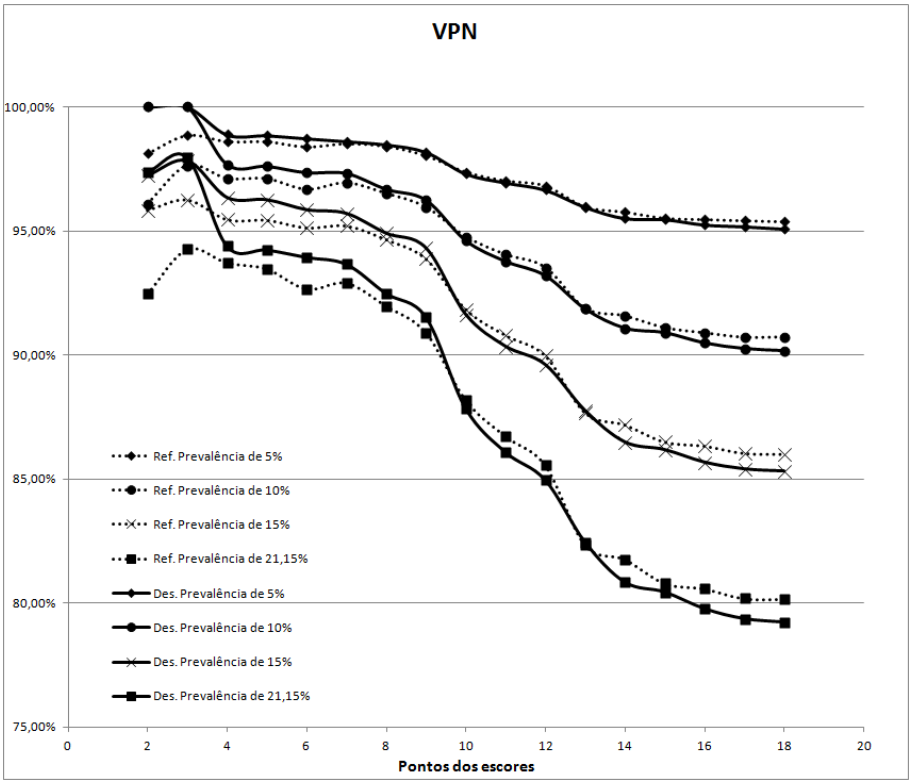
	Prevalência de 5%		Prevalência de 10%	
limiar de corte	VPP	VPN	VPP	VPN
2	5,20%	100,00%	10,38%	100,00%
3	5,27%	100,00%	10,50%	100,00%
4	5,85%	98,88%	11,59%	97,67%
5	6,36%	98,85%	12,53%	97,60%
6	6,71%	98,73%	13,16%	97,34%
7	8,67%	98,61%	16,86%	97,31%
8	10,02%	98,48%	18,88%	96,67%
9	11,18%	98,18%	20,99%	96,24%
10	11,67%	97,31%	22,02%	94,61%
11	13,81%	96,95%	25,25%	93,77%
12	15,79%	96,66%	28,38%	93,19%
13	15,66%	95,97%	28,26%	91,85%
14	16,67%	95,51%	31,25%	91,07%
15	20,69%	95,47%	35,29%	90,89%
16	30,00%	95,25%	46,15%	90,48%
17	50,00%	95,18%	60,00%	90,25%
18	33,33%	95,09%	50,00%	90,16%
	Prevalência de 15%		Prevalência de 21,15%	
limiar de corte	VPP	VPN	VPP	VPN
2	15,46%	97,22%	21,79%	97,37%
3	15,62%	97,83%	22,01%	97,96%
4	17,22%	96,34%	24,02%	94,38%
5	18,55%	96,25%	25,68%	94,23%
6	19,41%	95,85%	26,84%	93,93%
7	24,30%	95,70%	32,81%	93,65%
8	27,05%	94,90%	36,08%	92,47%
9	29,87%	94,31%	39,19%	91,52%
10	30,85%	91,63%	40,50%	87,84%
11	34,74%	90,34%	44,78%	86,07%
12	38,65%	89,61%	48,79%	84,95%
13	39,22%	87,75%	49,23%	82,45%
14	41,51%	86,48%	52,17%	80,84%
15	44,74%	86,17%	56,00%	80,44%
16	56,25%	85,67%	68,18%	79,77%
17	71,43%	85,40%	80,00%	79,37%
18	66,67%	85,31%	75,00%	79,23%

Tabela 5.37: Desempenho do escore para populações com diferentes prevalências de TB

pontos no escore, o valor preditivo positivo do escore desenvolvido neste trabalho é sempre maior que os valores obtidos pelo escore de referência, e essa diferença aumenta conforma a prevalência dos casos de TB na população. Entretanto, ambos os escores tem comportamento semelhante para os valores preditivos negativos, como visto na figura 5.18 b.



(a) Comparativo do VPP para diferentes prevalências



(b) Comparativo do VPN para diferentes prevalências

Figura 5.18: Função de distribuição acumulada da saída do escore

Capítulo 6

Conclusões

Segundo a estratégia DOTS, o exame prioritário para os casos suspeitos de TB pulmonar é a baciloscopia do escarro, devido ao seu baixo custo e pela fácil execução do exame, também são recomendados os exames clínicos de cultura, prova tuberculínica e exames radiológicos do tórax. Entretanto, os métodos tradicionais para o diagnóstico da tuberculose pulmonar apresentam limitações. A baciloscopia tem baixa sensibilidade, entre 40% a 60%, a cultura, em meio sólido, demanda de 4 a 8 semanas para se ter um resultado, a prova tuberculínica indica se o paciente teve contato com o *Mycobacterium tuberculosis* e nem todas as unidades básicas de saúde contam com equipamentos de radiografia. Além disso, o paciente portador do bacilo é a principal fonte de manutenção de transmissão da TB na população, portanto a detecção de novos casos é primordial para o controle da endemia.

Apesar do comprometimento do governo para com as ações de controle de TB, ainda nos deparamos com o grande desafio da expansão de cobertura das ações para o atendimento básico dado a população além de um aumento efetivo na detecção precoce de casos de TB. Portanto o uso de escores clínicos pode ser uma ferramenta que auxilie o atendimento básico e na detecção precoce de casos de TB por meio de profissionais da equipe de enfermagem.

Na prática, o profissional de enfermagem responsável por esse primeiro contato muitas vezes, além de identificar alguns sintomas clínicos, tem de tomar decisões sobre encaminhamento ou adiantamento de exames. Por utilizar uma avaliação to-

talmente pessoal na abordagem do suspeito de TB, não havendo uma padronização no atendimento dificultando a comparação e avaliação da qualidade dos procedimentos de atendimento.

Diante do exposto, o objetivo principal deste trabalho é fornecer uma ferramenta de apoio a tomada de decisão relativa a triagem de pacientes suspeitos de tuberculose pulmonar , usando somente informações clínicas e que fosse de fácil obtenção através de anamnese. Portanto, para um melhor entendimento do relacionamento entre os sintomas e a TB pulmonar foi utilizado redes neurais não supervisionadas do tipo SOM para identificação dos agrupamentos existentes nos pacientes atendidos na Policlínica Augusto do Amaral Peixoto e o risco associado aos pacientes associados a estes agrupamentos de terem TB pulmonar. Ao todo cinco escores diferentes foram propostos, utilizando-se de diferentes técnicas de classificação como o uso de Discriminante de Fisher e de Simulated Annealing.

Do ponto de vista da clusterização , a análise da distribuição espacial dos sintomas a partir dos Planos de Componentes do SOM mostrou-se bastante útil e de fácil entendimento. Tanto o SOM, quanto as análises estatísticas das variáveis, mostrou que existe um sentido para a distribuição espacial dos sintomas e o relacionamento entre eles.

Já a análise de dependência espacial entre os sintomas e os clusters formados pela segmentação do mapa gerado no treinamento do SOM mostrou que existe uma relação entre alguns sintomas e o risco de se ter TB. Mostrando que pessoas jovens, do sexo masculino, com presença de febre, emagrecimento, sudorese noturna estão mais sujeitas a contraírem TB pulmonar, enquanto a ausência de tosse, sexo feminino, são sinais e sintomas que caracterizam o cluster de baixo risco. Também pode ser visto que alguns sintomas são confundidores , ou não relevantes, para a clusterização como tabagismo, internação hospitalar e dispnéia.

Os modelos propostos se utilizaram somente de sinais e sintomas dos pacientes, visando uma fácil execução do teste, e os pesos dos escores eram números inteiros tornando o escore de fácil utilização em postos de saúde que não disponham de recur-

sos computacionais onde não possam ser utilizados sistemas de apoio ao diagnóstico mais refinados.

O melhor modelo foi o escore que pontua com pesos positivos entre 1 e 3 quando os sinais e sintomas estão presentes no paciente, tendo uma sensibilidade de 81,4% e especificidade de 61,3%. Tendo este escore apresentado uma curva ROC sempre maior que o atual escore utilizado na triagem de pacientes na PAAP, enquanto que o escore de referência tem 83,06% e 52,00% de sensibilidade e especificidade, respectivamente.

O escore quando utilizado com limiar de corte de 7 pontos, apresenta sensibilidade de 86,78% e sensibilidade de 52,33%, pode auxiliar o corpo de enfermagem na triagem dos pacientes pois, 1 em cada 3 pacientes triados como suspeitos de TB serão diagnosticado como TB positivo. E no ponto de vista da não proliferação do bacilo na população atendida no posto de saúde, apenas 7 em cada 100 pacientes que não passaram na triagem portam o bacilo. Portanto, o paciente que tiver uma pontuação no escore maior ou igual a 7 pontos passará na triagem e pode-se adotar como procedimento de atendimento o encaminhamento do paciente pelo corpo de enfermagem para realização de exames clínicos e laboratoriais para confirmar o diagnóstico da doença.

Uma vez que o paciente foi considerado como suspeito de TB pulmonar, o escore pode ser utilizado como ferramenta para seleção de pacientes que irão se submeter a testes mais complexos e de custo maior, otimizando tempo e recursos, como cultura para micobactéria, exames radiológicos do tórax e outros. Uma vez que 1 em cada 2 pacientes que obteve pontuação no escore maior que 14 pontos foi diagnosticado como TB pulmonar. Esse limiar de decisão também pode ser utilizado, eventualmente, na orientação do tratamento para não proliferação do bacilo em situações e locais onde os recursos humanos são escassos.

Os modelos propostos neste trabalho podem ser utilizados de forma inovadora como ferramenta de triagem e apoio ao diagnóstico da TB pulmonar. Já que agiliza o serviço médico e alem disso, devido a sua simplicidade, os mesmos podem ser

utilizados requerendo um esforço mínimo de treinamento para sua operação.

6.1 Trabalhos Futuros

Como trabalhos futuros sugerem-se o teste dos escores desenvolvidos em outras populações com diferentes prevalências da TB pulmonar na população para uma melhor avaliação dos sinais e sintomas utilizados neste escore.

Também o estudo do impacto do uso do escore no serviço médico, em relação aos custos de exames feitos desnecessariamente e agilidade no atendimento de pacientes. Assim como seu impacto em sítios com poucos recursos humanos, onde o inicio das ações profiláticas contra a proliferação do bacilo podem ser iniciada baseada no escore do paciente.

Indo mais além, se pode investigar a criação de modelos de escore específicos para situações como populações com índices elevado de co-infecção TB/HIV, ou onde há uma grande prevalência de casos de TB e diabetes, assim como para TB infantil, pleural e até bovina, já que a prevalência de TB no rebanho brasileiro chega até 32% [4].

Referências Bibliográficas

[1] WHO. *Global tuberculosis control 2010*. World Health Organization, 2010.

[2] SECRETÁRIA MUNICIPAL DE SAÚDE E DEFESA CIVIL DO RIO DE JANEIRO. `http://www.saude.rio.rj.gov.br/media/tuberculose.htm`. acessado em 21 de Junho de 2011.

[3] RUFFINO-NETTO, A. "Programa de controle da tuberculose no Brasil: Situação atual e novas perspectivas". In: *Informe Epidemiológico do SUS*, v. 10, pp. 129–138, 2001.

[4] ARAUJO, F. R., OSÓRIO, A. L. A. R., JORGE, K., et al. "Atualização em tuberculose bovina", *Embrapa Gado de Corte. Comunicado técnico, 121*, 2009.

[5] FUNDAÇÃO NACIONAL DE SAÚDE. *Tuberculose - guia de vigilância epidemiológica*. Ministério da Saúde, 2002.

[6] MINISTÉRIO DA SAÚDE DO BRASIL. `http://portal.saude.gov.br/portal/arquivos/pdf/apresentacao_incidencia_05_04_11.pdf`. acessado em 21 de Junho de 2011.

[7] KRITSKI, A. L., RUFFINO-NETTO, A. "Health sector reform in brazil: impact on tuberculosis control". In: *International Journal Tuberculosis Lung Disease*, v. 4, pp. 622–626, 2000.

[8] MELLO, F. C. Q. *Modelos preditivos para tuberculose pulmonar paucibacilar*. Tese de D.Sc., Faculdade de Medicina / UFRJ, Rio de Janeiro, RJ, Brasil, 2001.

[9] WHO. *THE GLOBAL PLAN TO STOP TB 2011 - 2015*. World Health Organization, 2011.

[10] HIJJAR, M., PROCÓPIO, M., FREITAS, L., et al. "Epidemiologia da tuberculose: importância no mundo, no Brasil e no Rio de Janeiro". In: *Pulmão RJ*, pp. 310–314, 2005.

[11] KRITSKI, A. L., CONDE, M. B., MUSY, G. R. *Tuberculose: do ambulatório a enfermaria*. Atheneu, 2006.

[12] FRIEDMAN, H. H. *Manual de Diagnóstico Clínico*. 3 ed. Rio de Janeiro, 1985.

[13] STONE, B., BURMAN, W., M.V., H., et al. "The diagnostic yeld of acid-fast-bacillus smear-positive sputum specimens". In: *Journal Clinical Microbiology*, pp. 1030–1031, 1997.

[14] SCHIRM, J., OOSTENDORP, L. A., MULDR, J. G. "Comparasion of amplicor, in house PCR and conventional culture for detection of mycobacterium in clinical samples". In: *Journal Clinical Microbiology*, pp. 3321–3324, 1995.

[15] SREERAMAREDDY, C. T., KISHORE, P. V., MENTEN, J., et al. "Time delays in diagnosis of pulmonary tuberculosis: a systematic review of literature". In: *BMC Infectious Diseases*, v. 9, p. 91, 2009. doi: 10.1186/1471-2334-9-91.

[16] AGRESTI, A. *An Introduction to Categorical Data Analysis*. Wiley, 2007.

[17] NEMES, S., JONASSON, J. M., GENELL, A., et al. "Bias in odds ratios by logistic regression modelling and sample size". In: *BMC Medical Research Methodology*, v. 9, p. 56, 2009. doi: 10.1186/1471-2288-9-56.

[18] KORB, K., NICHOLSON, A. E. *Bayesian Artificial Intelligence*. Chapman & Hall /CRC, 2003.

[19] DÍEZ, F. J., MIRA, J., ITURRALDE, E., et al. "DIAVAL, a Bayesian expert system for echocardiography". In: *Artificial Intelligence in Medicine*, v. 10, pp. 59–73, 1997.

[20] ANTAL, P., FANNES, G., TIMMERMAN, D., et al. "Using literature and data to learn Bayesian networks as clinical models of ovarian tumors". In: *Artificial Intelligence in Medicine*, v. 30, pp. 257–281, 2004.

[21] VISSCHER, S., LUCAS, P. J., SCHURINK, C. A., et al. "Modelling treatment effects in a clinical Bayesian network using Boolean threshold functions". In: *Artificial Intelligence in Medicine*, v. 46, pp. 251–256, 2009.

[22] STASIS, A., LOUKIS, E., PAVLOPOULOS, S., et al. "Using decision tree algorithms as a basis for a heart sound diagnosis decision support system". In: *4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine*, pp. 354–357, 2003.

[23] TU, M. C., SHIN, D., SHIN, D. "Effective Diagnosis of Heart Disease through Bagging Approach". In: *2nd International Conference on Biomedical Engineering and Informatics*, pp. 1–4, 2009.

[24] ZHAO, H., GUO, S., CHEN, J., et al. "Characteristic Pattern Study of Coronary Heart Disease with Blood Stasis Syndrome Based on Decision Tree". In: *4th International Conference on Bioinformatics and Biomedical Engineering*, pp. 1–3, 2010. doi: 10.1109/ICBBE.2010.5515979.

[25] MELLO, F., BASTOS, L., SOARES, S., et al. "Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: a cross-sectional study", *BMC Public Health*, v. 6, pp. 1–8, 2006.

[26] HAYKIN, S. *Neural Networks and Learning Machines*. Prentice-Hall, Inc., 2008.

[27] ARIF, M., MALAGORE, I., AFSAR, F. "Automatic Detection and Localization of Myocardial Infarction Using Back Propagation Neural Networks". In: *4th International Conference on Bioinformatics and Biomedical Engineering*, pp. 1–4, 2010. doi: 10.1109/ICBBE.2010.5514664.

[28] SMOLAR, P., SINCAK, P., JAKSA, R. "Application of AI in Cardiology". In: *IEEE 8th International Symposium on Applied Machine Intelligence and Informatics*, pp. 267–270, 2010.

[29] RAFIEE, A., MASOUMI, H., ROOSTA, A. "Using neural network for liver detection in abdominal MRI images". In: *IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 21–26, 2009. doi: 10.1109/ICSIPA.2009.5478613.

[30] BOCK, N. N., MCGOWAN JR, J. E., AHN, J., et al. "Clinical predictors of tuberculosis as a guide for a respiratory isolation policie." In: *Am J Respir Crit Care Med*, pp. 1468–1472, 1996.

[31] SAMB, B., HENZEL, D., DALEY, C. L. "Methods for diagnosing tuberculosis among in-patients in easten africa whose sputum smears are negative". In: *International Journal Tuberculosis Lung Disease*, pp. 25–30, 1997.

[32] EL-SOLH, A. A., HSIAO, C., GOODNOUGH, S., et al. "Predicting active pulmonary tuberculosis using an artificial neural network". In: *Chest*, n. 4, pp. 968–973, 1999.

[33] KANAYA, A. M., GLIDDEN, D. V., CHAMBERS, H. F. "Identifying pulmonary tuberculosis in patients with negative sputum smear results". In: *Chest*, n. 2, pp. 349–355, 2001.

[34] ARIS, E. A., BAKARI, M., CHONDE, T. M., et al. "Diagnosis of tuberculosis in sputum negative patients in dar es salaam". In: *East Afri Med J*, pp. 630–634, 1999.

[35] SANTOS, A. M. *Redes Neurais e Árvores de Classificação Aplicadas ao Diagnóstico da Tuberculose Pulmonar Paucibacilar*. Tese de D.Sc., COPPE / UFRJ, Rio de Janeiro, RJ, Brasil, 2003.

[36] SANTOS, A. M., PEREIRA, B., SEIXAS, J. M., et al. "Neural networks: An Application for Predicting Smear Negative Pulmonary Tuberculosis", *Advances in Statistical Methods for the Health Sciences*, pp. 279–289, 2007.

[37] BENF, Y., HONGMEI, S., YE, S., et al. "Study on the Artificial Neural Network in the Diagnosis of Smear Negative Pulmonary Tuberculosis". In: *WRI World Congress on Computer Science and Information*, v. 5, pp. 584 – 588, 2009.

[38] UCAR, T., KARAHOCA, D., KARAHOCA, A. "Predicting the existence of mycobacterium tuberculosis infection by Bayesian Networks and Rough Sets". In: *Biomedical Engineering Meeting BIYOMUT 2010*, pp. 1–4, 2010. doi: 10.1109/BIYOMUT.2010.5479850.

[39] ASHA, T., NATARAJAN, S., MURTHY, K. "Diagnosis of tuberculosis using ensemble methods". In: *3rd IEEE International Conference on Computer Science and Information Technology*, v. 8, pp. 409–412, 2010.

[40] ROKACH, L. "Ensemble-based classifiers", *Artificial Intelligence Review*, v. 33, pp. 1–39, 2010. doi: 10.1007/s10462-009-9124-7.

[41] POLIKAR, R. "Ensemble based systems in decision making". In: *IEEE Circuits and Systems Magazine*, v. 6, pp. 21–45, 2006.

[42] BOEHME, C. C., NABETA, P., HILLEMANN, D., et al. "Rapid Molecular Detection of Tuberculosis and Rifampin Resistance". In: *New England Journal of Medicine*, v. 363, pp. 1005–1015, 2010.

[43] CASTRO, C. *Avaliação de um escore clínico para identificação de suspeitos de tuberculose pulmonar em cenário de atenção básica*. Dissertação de M.Sc., Faculdade de Medicina / UFRJ, Rio de Janeiro, RJ, Brasil, 2010.

[44] SANTOS, A., PEREIRA, B., SEIXAS, J., et al. "Neural Networks: An Application for Predicting Smear Negative Pulmonary Tuberculosis". In: *Advances in Statistical Methods for the Health Sciences*, pp. 275–287, 2007. doi: 10.1007/978-0-8176-4542-7_18.

[45] SOUZA FILHO, J., SEIXAS, J., ANTUNES, P., et al. "Redes Neurais Aplicadas ao Diagnóstico da Tuberculose Pulmonar Paucibacilar". In: *VIII Congresso Brasileiro de Rede Neurais*, Florianopolis, 2007.

[46] EDWARDS, A. "The measure of association in a 2x2 table", *Journal of the Royal Statistical Society*, v. 126, pp. 109–114, 2009.

[47] DUDA, R. O., HART, P. E., STORK, D. G. *Pattern Classification*. Wiley, 2001.

[48] WITTEN, I. H., FRANK, E., HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.

[49] THEODORIDIS, S., KOUTROUMBAS, K. *Pattern Recognition*. Elsevier, 2009.

[50] JAIN, A. K., MURTY, M. N., FLYNN, P. J. "Data Clustering: a Review". In: *ACM Computing Surveys*, v. 31, pp. 264–323, 1995.

[51] HARTIGAN, J. *Clustering Algorithms*. Wiley, 1975.

[52] KOHONEN, T. *Self-Organizing Maps*. Springer, 2000.

[53] VESANTO, J., HIMBERG, J., ALHONIEME, E., et al. *SOM Toolbox for Matlab 5 Documentation*. In: Report A57, Helsinki University of Technology, Finland, 2000.

[54] JOLLIFFE, I. *Principal Component Analysis*. Springer, 2002.

[55] ULTSCH, A. "Self-Organizing Neural Networks for Visualization and Classification". In: *Information and Classification*. Springer.

[56] COSTA, J. A. F. "Uma Nova Abordagem para Visualização e Detecção de Agrupamentos em Mapas de Kohonen Baseado em Gradientes das Componentes". In: *Learning and Nonlinear Models,Journal of the Brazilian Neural Network Society*, v. 9, pp. 20–31, 2011.

[57] VESANTO, J., ALHONIEMI, E. "Clustering of the Self-Organizing Map". In: *IEEE Transactions on Neural Networks*, v. 11, pp. 586–600, 2000.

[58] DAVIES, D., BOULDIN, D. "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. PAMI-1, pp. 224–227, 1979.

[59] KIRKPATRICK, S., JR, C. D. G., VECCHI, M. P. "Optimization by Simulated Annealing". In: *Science*, v. 220, pp. 671–680, 1983.

[60] FISHER, R. A. "The Use of Multiple Measurements in Taxonimic Problems". In: *Annuals of Eugenics*, v. 7, pp. 179–188, 1936.

[61] BISHOP, M. *Pattern Recognition and Machine Learning*. Springer, 2007.

[62] DOS ANJOS, A., TORRES, R., SEIXAS, J. "Neural triggering system operating on high resolution calorimetry information". In: *Nuclear Instruments and Methods in Physics Research*, v. 559, pp. 134–138, 2006.

[63] KOOPMANS, T., BECKMANN, M. "Assignment problems and the location of economic activities". In: *Econometrica*, pp. 53–76, 1957.

[64] LOIOLA, E. M., DE ABREU, N. M. M., NETTO, P. O. B. "Uma revisão comentada das abordagens do problema quadrático de alocação". In: *Pesquisa Operacional*, v. 24, pp. 73–109, 2004.

[65] PENG, T., HUANCHEN, W., DONGME, Z. "Simulated annealing for the quadratic assignment problem: A further study". In: *18th International Conference on Computers and Industrial Engineering*, v. 31, pp. 925–928, 1996.

[66] LAURSEN, P. S. "Simulated annealing for the QAP – Optimal tradeoff between simulation time and solution quality", *European Journal of Operational Research*, v. 69, n. 2, pp. 238–243, 1993.

[67] SECRETARIA DE VIGILÂNCIA EM SAÚDE. *Manual de Recomendações para o Controle da Tuberculose no Brasil*. Ministério da Saúde, 2010.

Apêndice A

Termo de Consentimento Livre e Esclarecido

XXII. **Termo de Consentimento**

Número do Prontuário: _____ No. Ficha_____

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Projeto: Implantação de um sistema de informações integrado e de modelos matemáticos para o diagnostico de TB pulmonar nas Unidades de Saúde do Estado do Rio de Janeiro

PROPOSTA PARA PROJETO DE PESQUISA

Este é um estudo que pretende avaliar a utilidade de um sistema computadorizado integrado para agilizar o diagnostico de tuberculose (TB), em indivíduos moradores nas cidades do Rio de Janeiro que tem elevada ocorrência de casos de tuberculose no pulmão em atividade e, elevada a proporção de pacientes que recebem o tratamento para tuberculose sem a confirmação bacteriológica. As propostas deste estudo são: a) aumentar a proporção de pacientes com diagnostico de certeza TB no pulmão (com confirmação bacteriologica através de exame direto no escarro, chamado baciloscopia e também de exame de cultura do escarro, que raramente e realizado nas Unidades de Saúde; b) verificar a utilidade de testes matemáticos com os dados de exame medico, analise do RX de tórax e teste cutâneo para confirmar o diagnostico de TB no pulmão. Pois, pode-se indicar tratamento com medicamentos para prevenir a doença quando se identifica precocemente a tuberculose infecção, principalmente entre aqueles que moram com pacientes portadores de tuberculose no pulmão. Como você está sendo investigado para a TB no pulmão, você esta sendo convidado a participar deste estudo. Aproximadamente, 1200 pessoas irão participar deste estudo.

PROCEDIMENTOS

Se você concordar em participar, ou permitir que o menor sob sua responsabilidade legal participe, as seguintes coisas acontecerão:

1. Você irá responder algumas questões sobre a sua história clínica. Isto levará cerca de 20 minutos.
2. Você será submetido a um exame físico de rotina. Isto levará cerca de 20 minutos.
3. Caso você tenha procurado a Unidade de Saúde sob suspeita de TB no pulmão, você será submetido aos exames de rotina: exame de escarro, radiografia do tórax e teste cutâneo chamado prova tuberculínica.
4. Caso apresente sintomas respiratórios (tosse por mais de duas semanas), informe ter tido tratamento para tuberculose no pulmão no passado, informe contato com alguém que tratou de TB no pulmão nos últimos 5 anos, ou seu teste de pele mostrar enduração maior de 15 mm, você fará uma radiografia do tórax. Isto levará cerca de 15 minutos. Este exame faz parte da investigação de rotina de situações similares caso você procurasse um Posto de Saúde ou Hospital, portanto não será realizada por causa deste estudo.
5. Você terá sangue coletado do seu braço com uma agulha para realizar teste anti-HIV. Uma parte deste sangue será guardada para testes confirmatórios, se necessário. Isto levará cerca de 5 minutos.
6. Você será submetido ao teste da pele chamado prova tuberculínica. A agulha poderá causar algum desconforto por alguns segundos e uma pequena reação inflamatória poderá surgir em poucos dias, mas irá desaparecer. Importante lembrar que, em 48 a 72 horas, um profissional de saúde deve medir a reação inflamatória da pele no local da aplicação para definir se o teste foi válido ou não.
7. Caso tenha tosse por mais de duas semanas, você terá uma amostra do seu escarro enviada para o laboratório para ser testada para a tuberculose com os testes rotineiros e com o teste sob investigação.
8. Você terá uma amostra de fezes para o laboratório para ser testada a presença de verminose. Caso seja detectada verminose você será orientado para receber tratamento adequado.

Sua participação neste estudo deverá ser de 1-15 dias nas próximas semanas, até a confirmação ou não do diagnostico de TB, dependendo do número de testes que seu médico achar necessário no seu caso.

17

Número do Prontuário: _____ No. Ficha_____

Em caso de injúria
Se você sentir que tem um problema médico relativo a sua participação neste estudo, ou tiver questões sobre o tratamento, por favor contate o Dr Afrânio Kritski ou Dra Fernanda Carvalho Queiroz Mello nos telefones 2562-2426 ou 2550-6903.

Se você sofrer qualquer problema (injúria) como resultado da sua participação neste estudo, ou se acreditar que não tenha sido tratado razoavelmente, por favor, contate a Dra Fernanda Carvalho Queiroz Mello nos telefones 2562-2426 ou 2550-6903 ou os coordenadores locais na Unidade de Saúde em que foi atendido. As Unidades de Saúde da Secretaria Municipal do Rio de Janeiro, e o Programa Acadêmico de Tuberculose da Universidade Federal do Rio de Janeiro, na Ilha do Fundão, não possuem um programa para reembolsá-lo caso seja ferido ou tiver outro resultado desfavorável que não seja falha dos médicos deste estudo.
Os custos de qualquer tratamento ou hospitalização que você venha a necessitar, como resultado da sua participação neste estudo, serão financiados pelo Ministério da Saúde através de clínicas públicas e hospitais.

Questões a respeito dos seus direitos como participante de estudo de pesquisa
Se você tiver alguma pergunta a respeito dos seus direitos como participante deste estudo de pesquisa, você deverá contatar o Comitê de Ética para Pesquisa em Seres Humanos do Hospital Universitário Clementino Fraga Filho no telefone 2562-2480.

Participação por livre opção (opção voluntária pelo estudo)
Você não precisa participar deste ou de qualquer outro estudo. Se você decidir pela sua participação, e depois mudar de idéia, você poderá deixá-lo em qualquer momento. Se você recusar a sua participação ou desistir no decorrer deste estudo, você não será penalizado ou perderá qualquer benefício para o qual seja candidato.

O que a sua assinatura significa:
A sua assinatura abaixo significa que você compreendeu a informação fornecida sobre este estudo e sobre este termo de consentimento. Se você assinar este termo, isto significará que você concorda em participar deste estudo.
Nós daremos a você uma cópia deste termo de consentimento.

_____ ____/____/____
Assinatura do Paciente Data

Para crianças ou adultos sem capacidade de dar o consentimento

_____ ____/____/____
Assinatura do Parente ou guardião legal Data

_____ ____/____/____
Assinatura do Parente 2 (quando aplicável) Data

Relação do responsável com o indivíduo: _____

Assinatura do Investigador

_____ ____/____/____
Assinatura do Investigador ou indivíduo Data
\com designação aprovada pelo CEP/UFRJ
Assinatura da Testemunha

_____ ____/____/____
Testemunha se o indivíduo for analfabeto ou Data
incapaz de assinar o seu nome

Apêndice B

Carta de aprovação do Comitê de ética

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
Hospital Universitário Clementino Fraga Filho
Faculdade de Medicina
Comitê de Ética em Pesquisa - CEP

CEP - MEMO - nº 318/06 Rio de Janeiro, 02 de maio de 2006.

Do: Coordenador do CEP

A (o): Sr. (a) Pesquisador (a): Prof. Afrânio Lineu Kritski

Assunto: Parecer sobre projeto de pesquisa.

Sr. (a) Pesquisador (a),

Informo a V. Sa que o CEP constituído nos Termos da Resolução n. ° 196/96 do Conselho Nacional de Saúde e, devidamente registrado na Comissão Nacional de Ética em Pesquisa, recebeu, analisou e emitiu parecer sobre a documentação referente ao Protocolo e seu respectivo Termo de Consentimento Livre e Esclarecido, conforme abaixo *discriminado*:

Protocolo de Pesquisa: 067/06 - CEP

Título: "Implantação de um sistema de informações integrado e de modelos matemáticos para o diagnóstico de TB pulmonar nas unidades de saúde do estado do Rio de Janeiro."

Pesquisador (a) responsável: Prof. Afrânio Lineu Kritski

Data de apreciação do parecer: 24/04/2006

Parecer: "APROVADO"

Informo ainda, que V. Sa. deverá apresentar relatório semestral, previsto para 24/10/2006, anual e/ou relatório final para este Comitê acompanhar o desenvolvimento do projeto. (item VII. 13.d., da Resolução n. ° 196/96 – CNS/MS).

Atenciosamente,

Prof. Luiz Carlos Duarte de Miranda
Coordenador do CEP