APLICAÇÃO DA REGRESSÃO LOGÍSTICA E DA REDE NEURAL PROBABILÍSTICA NA CLASSIFICAÇÃO DE RESISTÊNCIA AOS ANTIRRETROVIRAIS LOPINAVIR E NELFINAVIR

Letícia Martins Raposo

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

Orientador: Flavio Fonseca Nobre

Rio de Janeiro
Fevereiro de 2014

APLICAÇÃO DA REGRESSÃO LOGÍSTICA E DA REDE NEURAL PROBABILÍSTICA NA CLASSIFICAÇÃO DE RESISTÊNCIA AOS ANTIRRETROVIRAIS LOPINAVIR E NELFINAVIR

Letícia Martins Raposo

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 BIOMÉDICA.

Examinada por:

Prof. Flavio Fonseca Nobre, Ph.D.

Prof. Marcio Nogueira de Souza, D.Sc.

Prof. Oswaldo Gonçalves Cruz, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
FEVEREIRO DE 2014

Raposo, Letícia Martins

Aplicação da regressão logística e da rede neural probabilística na classificação de resistência aos antirretrovirais Lopinavir e Nelfinavir/Letícia Martins Raposo. – Rio de Janeiro: UFRJ/COPPE, 2014.

XV, 99 p.: il.; 29, 7cm.

Orientador: Flavio Fonseca Nobre

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Biomédica, 2014.

Referências Bibliográficas: p. 85 – 90.

1. Regressão logística. 2. Redes neurais probabilísticas. 3. Classificadores. I. Nobre, Flavio Fonseca. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Biomédica. III. Título.

Dedico este trabalho aos meus pais com todo o meu amor e carinho.

Agradecimentos

Primeiramente, agradeço a Deus por tudo que tens proporcionado em minha vida, me dando forças para não desistir e proteção para me amparar.

Aos meus pais, Elizabeth e Gustavo, por todo amor, carinho e orações dedicados a mim.

Ao meu orientador Prof. Flavio Fonseca Nobre pelos seus ensinamentos, orientações, dedicação e, principalmente, pela paciência. Obrigada por estar sempre presente, me dando todo o suporte necessário para que este trabalho fosse concluído.

Ao meu tio Werley que não mediu esforços para me auxiliar na finalização deste trabalho. Obrigada por estar presente na concretização de mais uma etapa em minha vida.

Aos meus amigos de mestrado pelo companheirismo e amizade. Agradeço especialmente às minhas amigas Aline, Beatriz, Carolina, Débora, Gabriela, Liliana, Natália, Raquel e Viviane por todas as risadas e choros compartilhados. Obrigada pelas ótimas histórias vividas e por ajudar a tornar a vida de mestranda muito mais divertida. Essas lembranças serão eternamente guardadas em meu coração.

Aos meus familiares por sempre torcerem por mim e me apoiarem nas minhas decisões.

Aos meus amigos de Além Paraíba pelo carinho de sempre e presença constante em minha vida.

Aos amigos do LESS/PEB que tive o prazer de conviver nesse período, em um ótimo ambiente de crescimento intelectual e pessoal.

Aos professores do PEB que sempre contribuíram na minha formação e me apoiaram neste período de grande dedicação.

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

APLICAÇÃO DA REGRESSÃO LOGÍSTICA E DA REDE NEURAL PROBABILÍSTICA NA CLASSIFICAÇÃO DE RESISTÊNCIA AOS ANTIRRETROVIRAIS LOPINAVIR E NELFINAVIR

Letícia Martins Raposo

Fevereiro/2014

Orientador: Flavio Fonseca Nobre

Programa: Engenharia Biomédica

O HIV é o agente etiológico responsável por uma das doenças de maior impacto no mundo, a Aids. Com o advento da terapia antirretroviral, uma grande redução dos índices de morbidade e mortalidade foi registrada. Entretanto, muitos indivíduos não vêm apresentando um benefício durável, relacionado ao problema de resistência a essas terapias. Detectar a presença de resistência a determinada droga antirretroviral é uma importante ferramenta de monitoração na prática clínica. Dessa forma, observa-se que o desenvolvimento de modelos capazes de predizer a resistência aos antirretrovirais torna-se útil na tomada de decisão do melhor regime terapêutico para o indivíduo HIV positivo. O objetivo deste trabalho foi desenvolver modelos de predição de resistência aos inibidores da HIV-protease Lopinavir e Nelfinavir, fazendo uso de duas técnicas de modelagem: a regressão logística e a rede neural probabilística. Os modelos logísticos para o Lopinavir apresentaram AUCs variando de 0,78 a 0,82, acurácias de 0,66 a 0,74, sensibilidades de 0,64 a 0,91, especificidades de 0,62 a 0,76 e índices *Kappa* variando de 0,23 a 0,29. Nas redes neurais, esses valores foram: AUCs variando de 0,60 a 0,76, acurácias de 0,64 a 0,79, sensibilidades de 0,56 a 0,73, especificidades de 0,64 a 0,82 e índices *Kappa* variando de 0,16 a 0,33. Para o Nelfinavir, os modelos logísticos apresentaram AUCs variando de 0,70 a 0,82, acurácias de 0,71 a 0,82, sensibilidades de 0,53 a 0,73, especificidades de 0,75 a 0,89 e índices *Kappa* variando de 0,31 a 0,53. Nas redes neurais, esses valores foram: AUCs variando de 0,61 a 0,77, acurácias de 0,60 a 0,82, sensibilidades de 0,48 a 0,75, especificidades de 0,57 a 0,93 e índices *Kappa* variando de 0,18 a 0,45.

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

APPLICATION OF LOGISTIC REGRESSION AND PROBABILISTIC NEURAL NETWORK IN THE CLASSIFICATION OF RESISTANCE TO ANTIRETROVIRAL DRUGS LOPINAVIR AND NELFINAVIR

Letícia Martins Raposo

February/2014

Advisor: Flavio Fonseca Nobre

Department: Biomedical Engineering

HIV is the etiological agent responsible for one of the most impact disease on the world, the Aids. With the advent of antiretroviral therapy, a major reduction in morbidity and mortality was recorded. However, many individuals have not presented a lasting benefit due to the problem of resistance to these therapies. Detecting the presence of particular antiretroviral drug resistance is an important tool for monitoring in clinical practice. Thus, it is observed that the development of models capable of predicting resistance to antiretroviral drugs is useful in making decision about the best treatment regimen for HIV-positive individual. The aim of this study was to develop predictive models of resistance to HIV-protease inhibitors Lopinavir and Nelfinavir applying two modeling techniques: logistic regression and probabilistic neural network. Logistic models for Lopinavir presented the following average values: AUCs ranging from 0.78 to 0.82, accuracies from 0.66 to 0.74 , sensitivities from 0.64 to 0.91, specificities from 0.62 to 0.76 and *Kappa* indices ranging from 0.23 to 0.29. At probabilistic neural networks, these values were: AUCs ranging from 0.60 to 0.76, accuracies from 0.64 to 0.79, sensitivities from 0.56 to 0.73, specificities from 0.64 to 0.82 and *Kappa* indices ranging from 0.16 to 0.33. For Nelfinavir, the mean values for logistic models were AUCs ranging from 0.70 to 0.82 , accuracies from 0.71 to 0.82, sensitivities from 0.53 to 0.73, specificities from 0.75 to 0.89 and Kappa indices ranging from 0.31 to 0.53. At probabilistic neural networks, these values were: AUCs ranging from 0.61 to 0.77, accuracies from 0.60 to 0.82, sensitivities from 0.48 to 0.75, specificities from 0.57 to 0.93 and Kappa indices ranging from 0.18 to 0.45.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O vírus da imunodeficiência humana (*Human Immunodeficiency Virus*, HIV) é um retrovírus pertencente à família *Retroviridae* e responsável pela síndrome da imunodeficiência adquirida (*Acquired Immunodeficiency Syndrome*, Aids), documentada pela primeira vez em 1981, quando um número crescente de jovens homossexuais foi alvo de raras infecções oportunistas [1].

Existem dois tipos de HIV responsáveis pela Aids, o HIV-1 e o HIV-2. A maior parte das infecções é decorrente do HIV-1, classificado em diferentes subtipos de acordo com características relacionadas a variações genéticas [2]. A maioria das pessoas com HIV estão infectadas pelos subtipos não B do HIV-1. Entretanto, no Brasil, o subtipo B é o mais predominante, seguido pelos subtipos F, C e em menor proporção pelo D [3].

Segundo dados da UNAIDS [4], estima-se que em 2011 existiam cerca de 34 milhões de pessoas infectadas com o HIV, representando um aumento de 17% em relação ao ano de 2001, e que 2,7 milhões foram decorrentes de novas infecções ocorridas em 2010. Cerca de 1,8 milhões de pessoas infectadas por este vírus morreram em 2010, indicando uma redução do número de mortes quando comparado com meados do ano 2000 (2,2 milhões de mortes) [4].

No Brasil, os primeiros casos de Aids foram identificados no início da década de 80. Desta época até junho de 2012, foram notificados 656.701 casos da doença no país. Em 2011, foram notificados 38.776 casos e a taxa de incidência de Aids foi de 20,2 casos por 100 mil habitantes [5].

Diante da magnitude da infecção causada pelo HIV, a terapia antirretroviral tem sido oferecida de maneira gratuita e universal no Brasil desde 1996 [6]. Esse tratamento vem proporcionando uma redução dos índices de morbidade e mortalidade, assim como um aumento da qualidade de vida dos pacientes. Entretanto, para muitas pessoas, esta terapia não vem apresentando um benefício clínico durável, podendo ser resultado da baixa aderência ao tratamento e da resistência a essas terapias [7]. A resistência às drogas antirretrovirais é um dos maiores obstáculos para

a obtenção de um tratamento duradouro, levando a uma multiplicação de novas linhagens do HIV, o que gera um problema global para alcançar o sucesso a longo prazo do tratamento da Aids.

Detectar a presença de resistência à determinada droga é um importante desafio na monitoração na prática clínica. Ao se ter acesso ao resultado de um teste de resistência, o clínico pode indicar o melhor tipo de terapia a ser introduzida nos pacientes que ainda não fizeram uso de drogas, assim como alterar de forma adequada a terapia para aqueles que já apresentaram alguma falha no tratamento.

Diante deste cenário, observa-se que o desenvolvimento de modelos capazes de predizer a resistência às drogas antirretrovirais torna-se útil na escolha da melhor prática terapêutica para o indivíduo HIV positivo (HIV+). Vários estudos vêm sendo desenvolvidos com o intuito de determinar um bom modelo de predição do fenótipo de resistência do HIV às drogas, fazendo uso de diferentes variáveis preditivas e técnicas de modelagem, como métodos estatísticos, com destaque para as regressões linear e logística [8, 9], e algoritmos de aprendizagem de máquina (*Machine Learning*, ML) [10–13].

Dentre as técnicas de ML citadas na literatura, destaca-se a utilização de redes neurais artificiais (*Artificial Neural Networks*, ANNs) [10, 13], a utilização de máquinas de vetor de suporte (*Suport Vector Maquine*, SVMs) [10], aprendizagens não supervisionadas, aplicação de algoritmos com regras (*Stanford HIVdb*, VGI, ANRS, *Rega*) e análises lineares [13].

Embora o conhecimento sobre a resistência do HIV aos antirretrovirais tenha se expandido enormemente, padrões de resistência genotípica estão em constante evolução devido a mudanças nas estratégias de tratamento. Dessa forma, algumas questões precisam ser avaliadas, dentre elas a determinação das variáveis que melhor explicam esses padrões de resistência.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo deste trabalho é propor modelos baseados na regressão logística e na rede neural probabilística para o desenvolvimento de classificadores de resistência aos inibidores da HIV-protease. Através das técnicas de *bootstrap* e validação cruzada, o estudo objetiva selecionar as variáveis mais explicativas para a tarefa de classificação, otimizando o desempenho dos modelos.

1.1.2 Objetivos Específicos

- Selecionar as variáveis explicativas de cada classificador através das técnicas *Bootstrap*, *Stepwise*, Validação Cruzada e *Sequential Forward Selection*;
- Determinar o melhor método de codificação dos aminoácidos;
- Definir o melhor ponto de corte para a classificação pelos modelos logísticos e pelas redes neurais probabilísticas;
- Avaliar o desempenho dos classificadores em termos de acurácia, sensibilidade, especificidade, área sob a curva ROC e índice *Kappa*;
- Comparar o desempenho dos classificadores com os dos algoritmos *Stanford HIVdb* e *Rega*.

Capítulo 2

Fundamentos Teóricos

2.1 HIV

2.1.1 Classificação

O HIV pertence à família *Retroviridae*, sendo um membro do gênero *Lentivirus*. Dois tipos de HIV foram identificados como agente etiológico da Aids: o HIV tipo 1 (HIV-1), identificado em 1983 [14, 15] e responsável pela maior proporção das infecções, e o HIV tipo 2 (HIV-2), identificado inicialmente no oeste da África [16].

Análises filogenéticas das numerosas linhagens do HIV-1 vêm mostrando que o vírus pode ser dividido em grupos, subtipos, subsubtipos e formas recombinantes circulantes (*Circulating Recombinant Forms*, CRFs). Existem três grupos do HIV-1: M (*Major/Main*), N (*Non-M*, *Non-O/New*) e O (*Outlier*) [2]. O grupo M é o mais prevalente entre os três, sendo subdividido em nove subtipos (A-D, F-H, J e K) e em 15 CRFs [1].

Os subtipos predominantes mundialmente são o A e o C, seguidos do subtipo B [17]. No sul e leste da África, o subtipo C é o predominante, enquanto que, no oeste e no centro-oeste da África, a maioria pertence à forma recombinante CRF02-AG. Na América do Norte, Europa e Austrália, o subtipo B é o mais comum. Na América do Sul, este subtipo também é o predominante, porém os subtipos F e C também são encontrados [18]. No Brasil, a ocorrência dos subtipos segue a mesma característica, tendo adicionalmente uma pequena proporção do subtipo D [3].

2.1.2 Estrutura

O HIV, por ser um retrovírus, possui a enzima transcriptase reversa (*Reverse Transcriptase*, RT) que, em seu ciclo de replicação, catalisa a síntese de ácido desoxirribonucleico (*Deoxiribonucleic Acid*, DNA) a partir do ácido ribonucleico (*Ribonucleic Acid*, RNA). O HIV é composto de um material genético diploide, codificado por 15

proteínas virais, e constituído por duas fitas simples de RNA [19]. A partícula viral completa fora da célula hospedeira é constituída de um envelope externo, matriz, capsídeo e nucleocapsídio, como mostra a Figura 2.1.

Os principais genes encontrados no material genético do HIV são: gag, pol e env, responsáveis por codificar as principais proteínas do vírus. O gene gag codifica proteínas estruturais (no capsídeo, matriz e nucleocapsídio), como a p24; o pol é responsável pela codificação das enzimas como a RT, protease e integrase e; o env codifica proteínas do envelope do HIV, como as gp120 e gp41 [20].

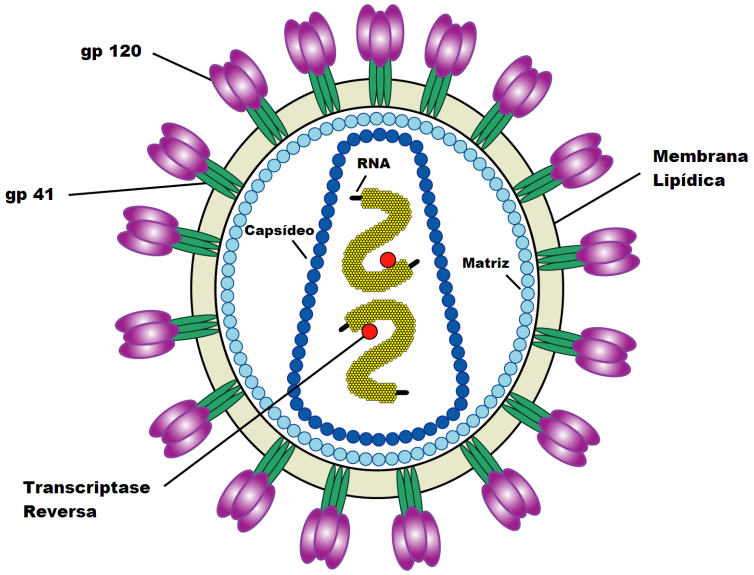

Figura 2.1: Estrutura do HIV.
Adaptado de *US National Institute of Health* (2005).

2.1.3 Patogenia e Ciclo de Replicação

O HIV é transmitido através do contato sexual, transfusão de sangue e da mãe infectada para o bebê por meio de três vias: durante o parto, perinatalmente ou através do leite materno [21].

Após a transmissão inicial do HIV, partículas virais se acumulam em altas concentrações no sangue dentro de poucas semanas, seguido de uma queda desses níveis devido à resposta imune do hospedeiro. Após esse episódio, a infecção permanece latente durante um longo período, variando de alguns anos até mesmo décadas [22].

O HIV apresenta tropismo pelas células CD4+, como linfócitos e macrófagos [21]. Ele infecta certos tipos de linfócitos denominados células T-*helper*, importantes para o sistema imune. Esse tropismo é determinado principalmente pelos receptores de

superfície das células, necessários para o vírus atacar e entrar na célula hospedeira. Normalmente, as células-alvo requerem o receptor CD4 e a quimiocina CCR5 [23]. A entrada nas células é mediada por interações sequenciais das proteínas virais gp120 e gp41 [24].

Com a redução do número desses linfócitos, o sistema imune torna-se incapaz de defender o organismo contra infecções. Durante este período, o número de linfócitos T CD4+ declina devido à morte causada pelo HIV. Este vírus, ao se replicar no interior dos linfócitos T CD4+, acaba matando-os no processo [22].

No ciclo de replicação do HIV, o vírus se funde à superfície da célula hospedeira e o material genético e as proteínas virais entram na célula. A RT catalisa a síntese de uma única molécula de DNA a partir do RNA e esta é então incorporada ao genoma da célula hospedeira através da proteína integrase. O novo RNA viral torna-se o material genético da célula e produz novas proteínas. Esse conjunto migra para a superfície da célula e um novo vírus, imaturo, é formado. A maturação do vírus é mediada pela protease, responsável por processar as proteínas virais, tornando-as funcionais [20, 25], como apresentado na Figura 2.2.

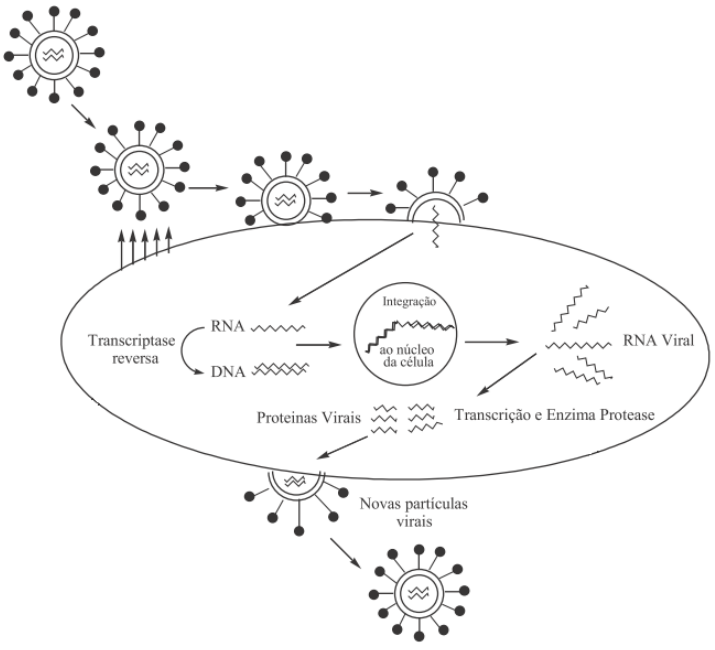

Figura 2.2: Esquema resumido do ciclo de replicação do HIV. Adaptado de Souza e Almeida (2003).

2.2 Terapia Antirretroviral

Em 1986, foi aprovado o primeiro antirretroviral (ARV), a Zidovudina, mais conhecido como AZT. Esta droga foi capaz de promover uma redução da mortalidade e prolongar a sobrevida dos indivíduos, embora por apenas mais alguns meses. Após a introdução do primeiro medicamento no mercado, diversos ARVs análogos foram produzidos com o objetivo de se obter fármacos mais eficazes.

Diversos governos e organizações internacionais têm disponibilizado tratamentos antirretrovirais para os países em desenvolvimento [18], entretanto, alguns países ainda não apresentam acesso a esta terapia e as projeções de tratamentos que visam uma vacina efetiva e a cura ainda são incertas [26]. No Brasil, os ARVs estão disponíveis ao público-alvo de maneira gratuita e universal desde 1996.

Terapias modernas combinam diferentes tipos de drogas inibidoras das enzimas virais, podendo levar a uma prolongada supressão viral e, em algumas vezes, a uma reconstrução imunológica, além de reduzir a morbidade e a mortalidade relacionada à infecção pelo HIV [27]. Em muitos países, a terapia adotada é a HAART (*Highly Active Antiretroviral Therapy*) que combina um ou dois inibidores da transcriptase reversa análogo de nucleosídeo (*Nucleoside Reverse Transcriptase Inhibitor*, NRTI), um inibidor da transcriptase reversa não análogo de nucleosídeo (*Non-nucleoside Reverse Transcriptase Inhibitor*, NNRTI) e/ou um inibidor de protease (*Protease Inhibitor*, PI) [28].

O resultado da terapia é avaliado através da carga viral do paciente. Quando os pacientes apresentam carga viral detectável e baixos níveis de linfócitos T CD4+, eles são denominados em falha terapêutica. As causas podem estar associadas a alguns fatores, tais como pouca aderência ao tratamento, devido aos efeitos colaterais desagradáveis causados por alguns medicamentos; concentrações sub-ótimas da droga; potência inadequada da droga, que consiste na quantidade necessária do fármaco para produzir efeito ou desenvolvimento de resistência viral. Uma vez que o tratamento começa a apresentar falhas, a medida a ser adotada é suspender o uso deste tipo de terapia e substituir por outro tratamento com diferentes inibidores [29].

2.2.1 Drogas Antirretrovirais

Atualmente existem mais de 20 drogas antirretrovirais que atuam em diferentes estágios no ciclo de vida do HIV: na entrada do vírus, bloqueando a fusão do HIV à membrana da célula; na transcrição reversa, compreendendo duas classes de ARVs que atuam tanto no sítio de ligação da RT, quanto alterando a sua conformação; na integração do vírus, impedindo a integração do DNA viral ao da célula infectada; e na maturação viral, inibindo competitivamente a enzima protease.

Os NRTIs atuam diretamente sobre a RT, incorporando-se à cadeia de DNA que o vírus cria, tornando-a defeituosa e impedindo a reprodução do vírus. São encontrados nesta classe sete inibidores: Zidovudina (ZDV, AZT), Estavudina (d4T), Emtricitabine (FTC), Lamivudina (3TC), Didanosina (ddl), Abacavir (ABC) e Tenofovir (TDF) [25]. Os NNRTIs são aqueles que se ligam de modo reversível e não competitivo à RT, promovendo sua inibição. Esses medicamentos apresentam a vantagem de não causar efeitos sobre os elementos formadores do sangue do hospedeiro, assim como não gerar resistência cruzada com os NRTIs. Esta classe possui 5 drogas: Nevirapina (NVP), Efavirenz (EFV), Etravirina (ETR), Delavirdina (DLV) e Rilpivirine (RPV).

Na classe de inibidores de fusão, existe apenas uma droga, a Enfuvirtida (T-20), que bloqueia a entrada do HIV nas células CD4+. Antagonistas de CCR5 também possui um único representante, o Maraviroc (MVC), que, ao bloquear a proteína CCR5, atua indiretamente na entrada do HIV. Inibidores de integrase, como o próprio nome diz, atuam sobre a integrase, responsável pela inserção do DNA do HIV ao DNA humano, inibindo a replicação do vírus. Seu representante é a droga Raltegravir (RAL) [30].

Outra classe de ARV amplamente usada no tratamento clínico da Aids é a dos PIs, introduzidos em 1995 após um intenso esforço no seu desenvolvimento [22]. Por atuarem sobre a HIV-protease, o mecanismo desses inibidores para ou reduz drasticamente a replicação do vírus, uma vez que essas drogas se ligam ao local de ativação da protease, bloqueando sua atividade. A HIV-protease é uma aspartil protease que possui apenas 99 resíduos dispostos em uma estrutura dimérica composta de dois monômeros idênticos [28]. Ela é uma enzima responsável por clivar com precisão as proteínas do HIV em diversos locais para completar a formação das partículas virais infecciosas [22].

Atualmente existem nove PIs aprovados pelo FDA (*Food and Drug Administration*): Amprenavir (APV), Atazanavir (ATV), Darunavir (DRV), Indinavir (IDV), Lopinavir (LPV), Nelfinavir (NFV), Ritonavir (RTV), Saquinavir (SQV) e Tipranavir (TPV). A efetividade dessas drogas está limitada pelo rápido processo de desenvolvimento de resistência aos ARVs na população viral [31].

2.3 Resistência aos Antirretrovirais

O desenvolvimento de resistência aos ARVs tem sido um dos maiores fatores limitantes na efetividade dessas terapias. Estudos mostram que o desenvolvimento de resistência é consequência da natureza altamente dinâmica de replicação do HIV [22].

Quando a terapia não consegue parar por completo a replicação viral devido à

alta taxa de mutação, aumentam-se as chances de que as modificações nas bases genéticas do HIV desenvolvam novas variantes resistentes aos ARVs. Com o desenvolvimento da resistência, tornam-se necessários altos níveis da droga para inibir a replicação viral. Porém, esses níveis são perigosos para os seres humanos [10].

A habilidade para predizer a resistência ao ARV pode ser útil no desenvolvimento de drogas mais efetivas e de regimes de tratamento mais duradouros. Com o aumento do número de drogas e das opções terapêuticas, os testes para avaliação de resistência a esses medicamentos vêm apresentando um importante papel no controle da infecção pelo HIV-1. Entretanto, algumas situações acabam intensificando o problema da resistência, como o limitado número de drogas aprovadas pelo FDA e a resistência cruzada, reduzindo o número de combinações efetivas de drogas [11].

2.3.1 Tipos de Resistência

Existem dois tipos de resistência às drogas: a resistência primária ou resistência transmitida, e a resistência adquirida.

A resistência primária é aquela apresentada por um indivíduo sem exposição prévia à droga, decorrente da transmissão direta de linhagens resistentes de indivíduos tratados. Elevadas taxas de resistência primária têm sido registradas em regiões em que a terapia antirretroviral está disponível por um longo tempo [32].

A resistência adquirida é resultado direto dos tratamentos com ARVs, sendo uma das causas de falha terapêutica principalmente em pacientes que experimentaram várias falhas no tratamento. Esse tipo de resistência é mais comum do que a resistência transmitida [32].

2.3.2 Testes de Resistência

O uso de testes que avaliam a presença de resistência à determinada droga é recomendado por ser uma importante ferramenta de monitoração na prática clínica. Ao ter acesso a um teste de resistência, pode-se melhorar o tipo de terapia a ser administrada aos pacientes que ainda não fizeram uso de drogas, assim como naqueles que já apresentaram alguma falha na terapia. Esses testes têm se tornado um importante passo no desenvolvimento de drogas e na otimização da combinação de terapias para o tratamento da infecção pelo HIV.

Testes de resistência genotípica (genotipagem) são capazes de determinar a presença ou ausência de específicas mutações genéticas no HIV que foram previamente associadas à resistência aos ARVs. Esses testes têm se tornado parte da rotina diagnóstica no tratamento de infecções por HIV por se tratar de um teste mais rápido, menos custoso e mais acessível [13].

Testes de resistência fenotípica (fenotipagem) fornecem uma medida quantitativa direta da suscetibilidade de linhagens do HIV a determinadas drogas. Entretanto, são testes muito caros, com uma demanda de tempo maior para gerar resultados e mais complexos, o que requer laboratórios especiais para a sua realização [13, 33].

2.4 Técnicas de Classificação

2.4.1 Regressão Logística

O modelo de regressão logística (RL) é um modelo linear generalizado, sendo um tipo de análise de regressão muito utilizado para realizar predições ou explicar a ocorrência de um evento específico quando a variável dependente (variável resposta) é de natureza binária. Quanto às variáveis independentes, estas podem ser tanto quantitativas quanto qualitativas.

Por se tratar de um modelo linear generalizado, a RL apresenta três componentes: uma componente aleatória, que consiste em uma combinação das variáveis independentes (preditoras); uma componente sistemática, que relaciona as variáveis independentes com os parâmetros do modelo, correspondendo à variável resposta que se quer modelar; e uma função de ligação, neste caso, a função *logit*, que conecta os valores esperados das observações às variáveis independentes [34].

Neste modelo de regressão, a variável resposta (y) é dicotômica, ou seja, é atribuído a ela dois valores: 1 para o acontecimento de interesse, denominado sucesso, e 0 para o acontecimento complementar, o fracasso. A probabilidade do sucesso é dada por π_i e a de fracasso é $1 - \pi_i$.

Considerando-se uma série de variáveis independentes $x_i = [x_1, x_2, ..., x_p]$, em que $x_i^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$ e um vetor $\beta = [\beta_0, \beta_1, ..., \beta_p]$ formado por parâmetros desconhecidos do modelo, a probabilidade de sucesso é dada por:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})} \tag{2.1}$$

E a probabilidade de fracasso é:

$$1 - \pi_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})} \tag{2.2}$$

O *logit* para o modelo de regressão múltipla é dado por:

$$g\left(x_i\right) = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = x_i^T\beta = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij} \tag{2.3}$$

Para estimar os valores de β, utiliza-se o método da máxima verossimilhança, que busca valores de β que maximizam essa função.

$$l(\beta) = \prod_{i=1}^{p}\pi\left(x_i\right)^{y_i}\left[1-\pi\left(x_i\right)\right]^{1-y_i} \tag{2.4}$$

$$L\left(\beta\right) = ln\left[l(\beta)\right] = \sum_{i=1}^{n}\{y_i \ln\left[\pi(x_i)\right] + (1-y_i)\ln\left[1-\pi(x_i)\right]\} \tag{2.5}$$

Para encontrar este valor, deriva-se a equação (2.5) em relação aos parâmetros do modelo e igualam-se as expressões a zero. Pelo fato dessas equações serem não lineares nos parâmetros, é necessário recorrer a métodos numéricos iterativos [34].

A significância do modelo final obtido pela RL é verificada através do desvio entre o modelo ajustado e o modelo saturado, onde todos os parâmetros se ajustam perfeitamente a todas as observações, tendo tantos parâmetros quanto dados observados. O modelo mais simples é o denominado modelo nulo, formado apenas pelo parâmetro β_0, indicando toda variação à componente aleatória. A estatística *deviance* (D) se baseia na função de verossimilhança e é utilizada para medir a discrepância de um modelo intermediário de p parâmetros em relação ao modelo saturado. Quanto menor a *deviance*, melhor o ajuste do modelo [35] .

$$D = -2\ \ln\frac{(verossimilhança\,do\,modelo\,ajustado)}{(verossimilhança\,do\,modelo\,saturado)} \tag{2.6}$$

Após obter o modelo ajustado, é necessário verificar se este apresenta uma boa descrição dos dados que foram observados através de uma análise de resíduos, que permite avaliar a qualidade do ajuste.

2.4.2 Redes Neurais

A rede neural artificial é uma técnica de aprendizagem de máquina utilizada para resolver padrões complexos, baseado no modelo biológico do cérebro humano. Esse tipo de técnica tem sido utilizado para solucionar problemas que representam uma relação não linear entre a entrada e a saída [12], sendo aplicada em diversos campos como modelos de previsão e métodos de classificação.

Uma ANN é formada por um grande número de unidades simples de processamento, usualmente denominadas de neurônios, associados a um elevado número de conexões entre eles. A informação entre os neurônios é transmitida através dos pesos sinápticos, representando a memória da rede, e cada neurônio possui uma

função de ativação, geralmente não linear [36]. As funções de ativação mais usadas são a função logística, que sempre assume valores positivos, e a função tangente hiperbólica, cuja saída pode assumir valores positivos ou negativos.

Basicamente, existem 3 tipos de redes:

- Rede *feedforward* de uma única camada, em que os neurônios de entrada simplesmente repetem o sinal de entrada na saída;
- Rede *feedforward* de multicamadas, formada pela camada de entrada, composta pelas variáveis de entrada relevantes para o modelo, a camada de saída, que fornece a solução do problema e, uma ou mais camadas intermediárias, chamadas camadas escondidas, onde os neurônios são efetivamente unidades processadoras e;
- Redes recorrentes, onde há pelo menos um laço de recorrência.

Existem diferentes algoritmos de aprendizagem para ANNs, destacando-se dois grandes grupos:

- Métodos de aprendizagem supervisionados, em que a rede neural recebe um conjunto de entradas e seus correspondentes padrões de saída e;
- Métodos de aprendizagem não supervisionados, em que a rede neural trabalha os dados de forma a determinar algumas propriedades similares entre os padrões de entrada, sem usar um conjunto de dados previamente conhecidos [37].

Ao implementarmos uma rede neural, os dados normalmente são separados em três conjuntos: treinamento, validação e teste. O conjunto de treinamento engloba uma amostra representativa do problema e é utilizado para treinar a rede e estimar seus parâmetros. O conjunto de validação tem como objetivo verificar a eficiência da rede quanto à sua capacidade de generalização durante o processo de treinamento. O conjunto de teste é utilizado para verificar a capacidade de generalização da rede ao aplicar dados reais.

As ANNs não fazem nenhuma suposição sobre a forma funcional entre as variáveis de entrada e saída, diferentemente do modelo de RL [13]. Não há necessidade de que as variáveis em estudo apresentem independência e normalidade, podendo ser utilizadas em problemas de regressão, classificação e compactação dos dados.

Um tipo de ANN adequado para problemas de classificação é a Rede Neural Probabilística (*Probabilistic Neural Network*, PNN). Esse tipo de rede foi desenvolvido por Specht (1990) [38], baseado na estratégia de decisão Bayesiana para classificar os vetores de entrada e no método de estimação de densidade de Parzen.

As estratégias Bayesianas são estratégias de decisão que minimizam o risco esperado de uma classificação. A regra de decisão ótima que minimiza os custos médios de erro de classificação é chamada de regra de decisão ótima de Bayes [38].

A rede PNN é uma rede composta tipicamente por quatro camadas:

- Camada de entrada, composta pelas variáveis explicativas e que não realiza operações sobre os vetores de entrada;
- Camada padrão, formada pelas unidades padrão, que armazenam, cada uma, um vetor do conjunto de treinamento, assim como a sua verdadeira classe;
- Camada de soma, que tem o número de elementos de processamento igual ao número de classes e;
- Camada de saída, que fornece a classificação dos dados de entrada, produzindo uma saída binária (1 em apenas uma das unidades e 0 nas demais).

A arquitetura de uma PNN está representada na Figura 2.3.

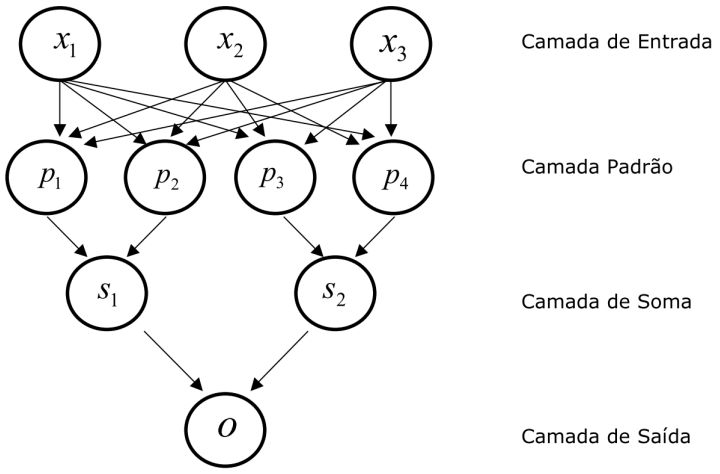

Figura 2.3: Arquitetura básica de uma rede neural probabilística.

Uma das vantagens da rede PNN é que este tipo de rede precisa de apenas uma etapa de treinamento, sendo muito rápida quando comparada a outras redes do tipo *feedforward*.

No treinamento, as unidades da camada padrão calculam a distância euclidiana $(x - x_{k_i})^T (x - x_{k_i})$, onde x é um vetor apresentado à rede no treinamento e x_{k_i} é o vetor da amostra que é armazenado na unidade padrão da classe i. Posteriormente, é aplicada uma função de ativação exponencial, obtendo $\exp - \left[\frac{(x-x_{k_i})^T (x-x_{k_i})}{2\sigma^2} \right]$ [38].

As unidades da camada padrão representam os vetores do conjunto de treinamento onde serão centradas as funções *kernel* utilizadas no método de Parzen.

Os resultados da camada padrão são repassados às unidades da camada de soma correspondentes a cada classe. Nesta camada ocorre apenas o somatório de suas entradas. As somas obtidas são enviadas às unidades de saída, que irão comparar os resultados recebidos e decidir por classificar x em uma das possíveis classes.

Ao assumir que as funções *kernel* utilizadas nas estimações sejam gaussianas e que as estimativas das densidades de probabilidade da população em cada classe sejam dadas respectivamente por $\hat{f}_A$ e $\hat{f}_B$, pode-se concluir que os estimadores dessas densidades são dados por:

$$\hat{f}_A(x) = \frac{1}{(2\pi)^{p/2}\,\sigma^p}\frac{1}{N_A}\sum_{i=1}^{N_A}\exp-\left[\frac{(x-x_{A_i})^T(x-x_{A_i})}{2\sigma^2}\right] \tag{2.7}$$

e

$$\hat{f}_B(x) = \frac{1}{(2\pi)^{p/2}\,\sigma^p}\frac{1}{N_B}\sum_{i=1}^{N_B}\exp-\left[\frac{(x-x_{B_i})^T(x-x_{B_i})}{2\sigma^2}\right] \tag{2.8}$$

onde p é o número de variáveis de entrada e σ é o fator de alisamento, que afeta a forma da superfície de decisão da rede.

O classificador implementado pela rede PNN irá decidir por alocar um vetor x na classe A se $\hat{f}_A(x) > \hat{f}_B(x)$. Caso o sinal da desigualdade esteja invertido, a rede alocará x na classe B.

2.5 Métodos de Seleção de Variáveis

A seleção de variáveis é um passo muito importante no desenvolvimento de modelos. Seu principal objetivo é determinar um subconjunto de variáveis independentes que melhor expliquem a variável resposta, isto é, dentre todas as variáveis explicativas disponíveis, deve-se encontrar um subconjunto daquelas mais importantes para o modelo, produzindo um erro mínimo de classificação.

Existem diferentes métodos de seleção de variáveis, dentre eles o Algoritmo Genético [39], *Sequential Search* (SS) [40] e métodos de *Stepwise* (*Forward*, *Backward* e *Both*) [40].

2.5.1 *Stepwise*

Neste tipo de seleção, as variáveis são adicionadas uma por vez ao modelo. O procedimento constrói iterativamente uma sequência de modelos de regressão pela adição ou remoção de variáveis em cada etapa. Aquelas não significativas são excluídas e

o procedimento se repete até que nenhuma outra variável possa entrar no modelo [34].

Para avaliar se as variáveis selecionadas explicam satisfatoriamente o modelo estatístico, pode-se aplicar o critério de informação de Akaike (*Akaike Information Criterion*, AIC) [41]. Os valores do AIC fornecem um meio para a seleção do modelo. Este critério penaliza modelos com muitas variáveis, sendo que valores menores de AIC são preferíveis. O AIC é dado pela fórmula:

$$AIC = -2log\left(L_p\right) + 2\left(p\right) \tag{2.9}$$

onde L_p é a função de máxima verossimilhança e p o número de parâmetros a serem estimados no modelo.

2.5.2 *Sequential Forward Selection*

O método *Sequential Forward Selection* (SFS) seleciona sequencialmente um subconjunto de variáveis explicativas que melhor prediz os resultados de uma variável resposta. Essa seleção ocorre até que não haja melhora na previsão.

Inicialmente, o método SFS avalia todos os subconjuntos formados por apenas uma variável e seleciona aquela que produz um modelo com melhor desempenho. Em seguida, as outras variáveis são adicionadas a essa que foi selecionada, formando subconjuntos bidimensionais. Aquele com melhor desempenho é selecionado e uma terceira variável é incluída a este subconjunto. Esse ciclo se repete até que nenhuma melhoria seja obtida ao aumentar o subconjunto atual. O critério de seleção das variáveis é definido pelo usuário, com destaque para o AIC e a AUC.

A Figura 2.4 mostra o método para 5 variáveis iniciais. Neste exemplo, o conjunto de variáveis foi reduzido de 5 (1-2-3-4-5) para 3 (2-3-5). Na primeira etapa, a variável 3 apresentou melhor desempenho. Subconjuntos bidimensionais foram formados com a variável previamente selecionada (Var 3) e as restantes (Var 1, Var 2, Var 4 e Var 5). Aquele que apresentou melhor desempenho foi selecionado e a partir deste, subconjuntos com 3 variáveis foram formados. Quando adicionada uma quarta variável, o subconjunto não apresentou um aumento no critério de avaliação de desempenho definido. Os quadrados cinzas indicam o melhor resultado obtido em cada passo da análise.

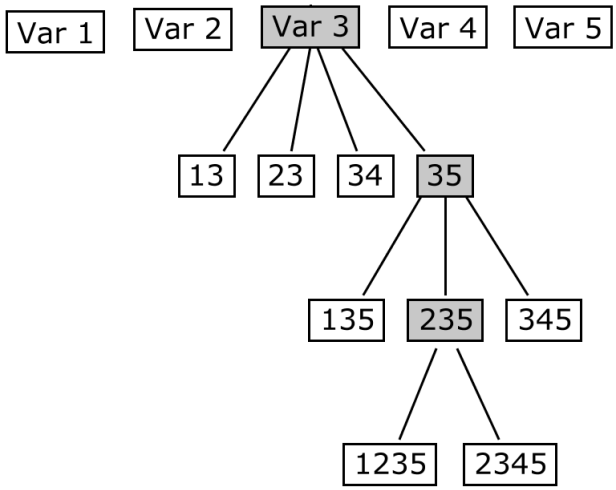

Figura 2.4: Exemplo de seleção de variáveis pelo método *Sequential Forward Selection.*

2.6 Métodos de Reamostragem

2.6.1 *Bootstrap*

O método de *Bootstrap* foi desenvolvido por Efron, em 1979 [42], sendo uma técnica de reamostragem bastante utilizada em diferentes situações estatísticas.

Baseia-se na geração de uma nova amostra de mesmo tamanho da original, a partir do sorteio aleatório com reposição de seus elementos. A substituição das observações após a amostragem permite a criação de tantas amostras quanto necessárias. Cada uma delas pode ser analisada independentemente e os resultados compilados ao longo da amostra.

2.7 Avaliação dos Classificadores

O desempenho de um classificador pode ser avaliado através de medidas calculadas a partir de uma matriz de confusão para duas classes. Essa matriz constitui-se em uma tabela de contingência 2x2 onde são representados quatro tipos de classificação segundo o resultado do modelo (Tabela 2.1).

- Classificados como positivos e pertencentes à classe positiva (verdadeiros positivos, VP);
- Classificados como negativos, mas pertencentes à classe positiva (falsos negativos, FN);

- Classificados como positivos, mas pertencentes à classe negativa (falsos positivos, FP);
- Classificados como negativos e pertencentes à classe negativa (verdadeiros negativos, VN);

A partir dessas classificações, é possível definir três medidas de desempenho mais comuns: acurácia, sensibilidade e especificidade.

Tabela 2.1: Matriz de confusão para duas classes.

	Classe positiva	*Classe negativa*
Predição positiva	VP	FP
Predição negativa	FN	VN

2.7.1 Acurácia

A acurácia é definida como a proporção de acertos do modelo. Ela é dada pela fórmula:

$$A = (VN + VP) / (VP + FN + FP + VN) \tag{2.10}$$

Para determinar o número de acertos do modelo final é necessário estabelecer uma probabilidade, denominada ponto de corte. Probabilidades estimadas pelo modelo que sejam maiores ou iguais a esse ponto de corte recebem valor 1, enquanto que as que sejam menores do que esse ponto de corte recebem valor 0.

2.7.2 Sensibilidade

Essa medida é definida como a proporção de verdadeiros positivos em relação ao total da classe positiva. A sensibilidade é calculada por:

$$S = VP / (VP + FN) \tag{2.11}$$

Um modelo, ao apresentar alta sensibilidade, raramente classificará como pertencente à classe negativa um valor da classe positiva, ou seja, fornece uma pequena taxa de falsos negativos.

2.7.3 Especificidade

Essa métrica compreende a proporção de verdadeiros negativos em relação ao total da classe negativa. A especificidade é calculada por:

$$E = VN/\left(FP + VN\right) \tag{2.12}$$

Um modelo, ao apresentar alta especificidade, dificilmente classificará como pertencente à classe positiva um valor da classe negativa, fornecendo, dessa forma, poucos resultados falsos positivos.

2.7.4 Curva ROC

Outra medida para avaliar o desempenho de um modelo de classificação foi desenvolvida no contexto de detecção de sinais eletrônicos e problemas com radares, no início dos anos 50, e é conhecida como curva ROC (*Receiver-Operating Characteristic*).

Essa curva é obtida traçando-se aos pares a sensibilidade e a especificidade a cada ponto de corte. Ela mostra a relação entre a sensibilidade e a especificidade de um teste e pode ser utilizada na decisão do melhor ponto de corte.

A área sob a curva ROC (AUC) é uma medida resumo usual de precisão, já que ela é estimada levando em consideração as sensibilidades e especificidades relativas a cada um dos valores estipulados [43].

Um modelo totalmente incapaz de discriminar valores pertencentes a uma classe ou outra apresenta uma AUC de 0,5. Quanto maior a capacidade do modelo em discriminar os valores segundo as classes, mais a curva se aproxima do canto superior esquerdo do gráfico e o valor da AUC se aproxima de 1. Na Figura 2.5 está representando um exemplo de curva ROC.

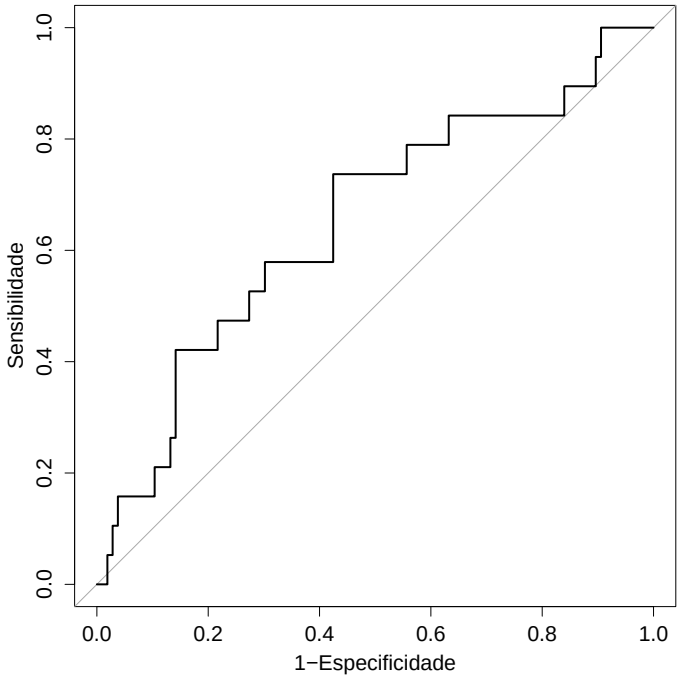

Figura 2.5: Exemplo de curva ROC.
A reta indica uma área igual a 0,5.

2.7.5 Índice *Kappa*

O índice *Kappa* mede o grau de concordância entre duas diferentes técnicas além do que seria esperado pelo acaso. Ele é calculado pela divisão da diferença entre a concordância esperada e a concordância observada e a diferença entre a concordância absoluta e a concordância esperada. Como esta última diferença representa a maior concordância possível entre a esperada e a observada, quanto maior é o índice *Kappa*, maior é a concordância entre as observações.

Landis e Koch (1977) definiram diferentes faixas para os valores de *Kappa*, segundo o grau de concordância que eles sugerem, formando a chamada escala de Landis. O índice *Kappa* pode variar de –1, indicando discordância completa, a +1, sugerindo concordância completa. Valores entre 1 e 0,80 indicam concordância quase perfeita; entre 0,79 e 0,60 representam concordância considerável; entre 0,59 e 0,40 indicam concordância moderada; entre 0,39 e 0,20 uma concordância razoável; entre 0,19 e 0 uma concordância baixa e valores menores que 0 não indicam concordância [44].

2.7.6 Validação Cruzada

A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo a partir de um conjunto de dados. É amplamente empregada em problemas

onde o objetivo da modelagem é a classificação.

2.7.6.1 Método *Holdout*

Este método consiste em dividir o conjunto total de dados em dois subconjuntos mutuamente exclusivos, um para treinamento (estimação dos parâmetros) e outro para teste, evitando a sobreposição entre os dois subconjuntos de dados.

Normalmente, $\frac{2}{3}$ dos dados são destinados para o treinamento e $\frac{1}{3}$ para o teste. Após o particionamento, a estimação do modelo é realizada e, posteriormente, os dados de teste são aplicados [45].

A desvantagem deste procedimento é que os resultados são altamente dependentes da escolha dos dados para a formação dos subconjuntos. Na Figura 2.6 está esquematizada a divisão dos dados através deste método.

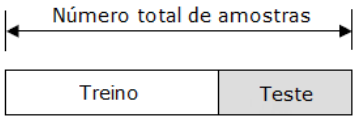

Figura 2.6: Representação da divisão dos dados na metodologia *holdout*.

2.7.6.2 Método *K-fold*

O método de validação cruzada denominado *k-fold* consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho. A partir disto, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para estimação dos parâmetros, calculando-se assim a acurácia do modelo. Este processo é realizado k vezes alternando de forma circular o subconjunto de teste [45].

Na validação cruzada 10-*fold*, comumente utilizada, o conjunto de dados é particionado aleatoriamente em 10 subconjuntos iguais. Destes, um único subconjunto é retido como dados de validação para testar o modelo e os 9 subconjuntos restantes são utilizados como dados de treino. O processo é repetido 10 vezes, com cada um dos 10 subconjuntos utilizados apenas uma vez como dados de validação. Este valor de k é muito utilizado, pois faz predições utilizando 90% dos dados, tornando-as mais suscetíveis de serem generalizadas para os dados completos. Na Figura 2.7 está esquematizado a divisão dos dados através do método *k-fold*, para k igual a 10.

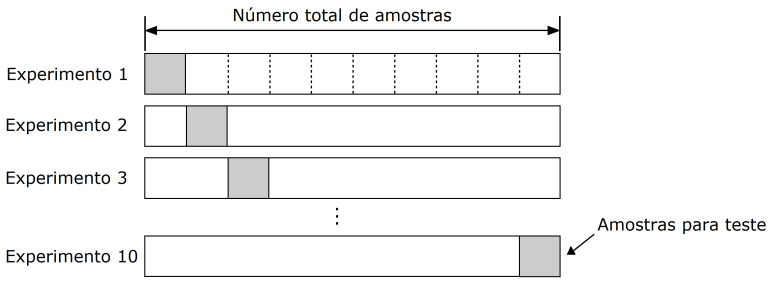

Figura 2.7: Representação da divisão dos dados na metodologia *k-fold* (k = 10).

2.7.6.3 Método *Leave-one-out*

O método *leave-one-out* é um caso específico do *k-fold*, com k igual ao número total de dados. A cada processo, uma amostra é destinada ao teste e o restante dos dados é aplicado no treinamento. Esse procedimento é largamente utilizado quando a disponibilidade dos dados é pequena.

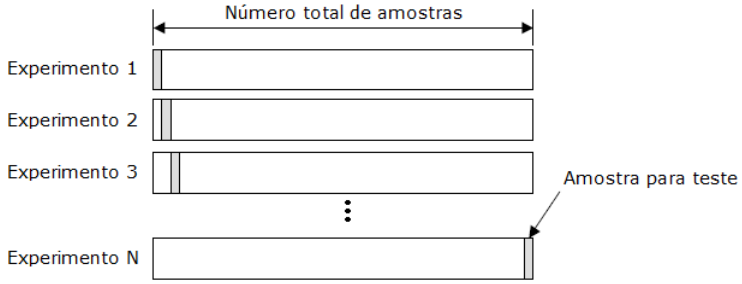

Figura 2.8: Divisão dos dados segundo o método de validação cruzada *Leave-one-out*.

Capítulo 3

Revisão de Literatura

A resistência às drogas tem sido um importante fator na falha das terapias antirretrovirais. Identificar a resistência aos medicamentos a partir do genótipo pode auxiliar na prática clínica e na escolha das drogas durante o processo terapêutico.

Existem diversos estudos com o intuito de desenvolver modelos capazes de predizer o desenvolvimento de resistência a um determinado ARV a partir de mutações. Muitos métodos supervisionados são empregados com esta finalidade, tais como árvores de decisão (*Decision Tree*, DT), SVMs e ANNs.

No estudo de Draghici e Potter (2003) [11], foi analisada a predição de resistência a dois PIs, IDV e SQV, através da construção de modelos preditores utilizando redes neurais do tipo SOM (*Self-Organizing Map*). O primeiro classificador foi construído para o IDV e baseou-se nas características estruturais do complexo inibidor-protease, uma vez que a perda de contato entre a droga e a protease acarreta em modificações no complexo e, consequentemente, em resistência à droga. O objetivo do estudo foi estabelecer uma relação entre as características estruturais das HIV-proteases com mutação e a resistência ao ARV correspondente, dado pelo IC90. Este índice consiste no resultado de uma fenotipagem e representa a quantidade de droga necessária para reduzir em 90% a replicação viral. A razão do IC90 de uma mutação específica por de um tipo selvagem é nomeado de *fold resistance*. As mutações foram classificadas segundo suas *fold resistance* em três classes: sem resistência ou resistência muito baixa (menor que *5-fold resistance*), baixa resistência (entre *5* e *10-fold resistance*) e alta resistência (maior que *10-fold resistance*). As categorias consideradas foram alta, média e baixa resistência ao PI, além de outras duas classes, uma na qual não se conhecia previamente a classificação quanto à resistência e outra mais heterogênea, na qual houve mistura de sequências resistentes e não resistentes ao PI. O preditor baseado nas informações estruturais apresentou uma acurácia entre 60% e 70% ao utilizar validação cruzada do tipo *leave-one-out*. O segundo classificador utilizou sequências de aminoácidos da HIV-protease com mutações para predizer a resistência ao SQV. As categorias consideradas foram baixa (menor que *5-fold re-*

sistance), média (entre *5* e *10-fold resistance*) e alta (maior que *10-fold resistance*). Os aminoácidos da HIV-protease que não sofreram mutação receberam valor zero e aqueles que diferiram da sequência original foram ordenados por frequência de ocorrência. A segunda técnica apresentou um classificador com 68% de acurácia e uma cobertura (razão entre os padrões de teste que foram classificados e o total de padrões de teste) de 69%. Ao combinar múltiplas redes, a acurácia foi de 85% e a cobertura foi de 84%, demonstrando que a resistência pode ser predita tanto por informações estruturais quanto pela sequência da proteína. O artigo apenas incorpora a acurácia para determinar o poder de discriminação do modelo, o que limita a sua interpretação. A adoção de outras métricas, como sensibilidade e especificidade, forneceriam informações adicionais para avaliar mais detalhadamente o modelo proposto.

Wang e Larder (2003)[13] também fizeram uso das ANNs para a criação de modelos preditivos. Neste estudo, a droga analisada foi a LPV, um inibidor de protease coadministrado com RTV para melhorar os níveis de droga no plasma dos pacientes. Dois classificadores foram desenvolvidos, um apresentando como entrada 11 mutações previamente relatadas na literatura e o outro com 28 posições selecionadas por uma análise de prevalência. A prevalência foi calculada para cada grupo (resistentes e suscetíveis) e a sua variação entre os grupos foi analisada. Mutações que apresentaram diferenças significantes entre os grupos (p-valor $< 0{,}05$) foram selecionadas como mutações significantes. As posições que não apresentavam mutação foram codificadas com 0 e as posições com mutação foram codificadas com 1. A variável resposta foi categorizada em suscetível ($<$ *2,5-fold resistance*) e resistente ($\geq$ *2,5-fold resistance*). Os modelos foram treinados, validados e testados. Estes foram construídos utilizando o algoritmo de retropropagação (*back-propagation*), no qual os pesos das conexões entre os neurônios são atualizados até a obtenção de um erro mínimo. A técnica de validação cruzada do tipo *10-fold* foi aplicada na divisão dos conjuntos de treino e validação. Quando comparado os resultados do modelo de 28 posições com o modelo de 11 posições, o coeficiente de correlação ($\mathbf{R^2}$) aumentou de 0,84 para 0,88 para o conjunto de teste e de 0,85 para 0,88 para o conjunto completo. O intervalo de confiança de 95% da AUC para o modelo de 11 posições foi igual a (0,88 - 0,92) e (0,84 - 0,90) para o conjunto completo e de teste, respectivamente. Para o modelo de 28 posições, o intervalo de confiança de 95% da AUC foi (0,92 - 0,96) e (0,93 - 0,99) no conjunto completo e de teste, respectivamente. Esses resultados indicaram que o classificador com mais posições foi significantemente mais acurado em termos de AUC para os dois conjuntos analisados (p-valor $< 0{,}05$). Este estudo contribuiu na comprovação da eficiência das ANNs no desenvolvimento de classificadores, assim como identificou novas mutações associadas à resistência da droga. Entretanto, valores de sensibilidade e especificidade não foram apresentados,

sendo apenas utilizados na construção da curva ROC (também não apresentada).

Wang *et al.* [46], em 2004, utilizaram uma regressão linear para a construção de modelos de resistência para 7 PIs e 10 inibidores da RT. A variável dependente foi o logaritmo natural da IC50 *fold change* (quantidade de droga necessária para reduzir em 50% a replicação viral) categorizada em suscetível e resistente de acordo com o ponto de corte de cada ARV. As variáveis independentes foram representadas pelas posições cujas mutações são consideradas importantes no desenvolvimento de resistência, segundo o banco de dados de *Stanford*. Cada variável independente recebeu valor 1 nos casos em que a sequência continha alguma mutação e valor 0 nos casos contrários. O método de *backward* foi utilizado para selecionar as variáveis independentes do modelo final. Os modelos foram avaliados através da validação cruzada *leave-one-out* e testados por um conjunto de teste independente. Os modelos foram comparados em termos de desempenho aos algoritmos de interpretação genotípica publicamente disponíveis (*Stanford HIVdb program (HIVdb)*, *Visible Genetics/Bayer Diagnostics Guidelines* (VGI), *Agence Nationale de Recherches sur le SIDA* (ANRS) e *Rega*), incluindo também os métodos de árvore de decisão e máquinas de vetores de suporte. Os coeficientes de correlação entre o logaritmo natural da IC50 *fold change* obtido pelo experimento e o predito variaram de 0,761 a 0,887. A sensibilidade dos modelos variou de 67,7% a 97,8% e a especificidade de 1,8% a 98,0%. O modelo linear superou todos os 7 algoritmos, apresentando uma acurácia média de 88,7%.

No estudo de Rhee *et al.* (2006) [47], cinco métodos de aprendizagem (árvores de decisão, redes neurais do tipo *feed-forward*, regressão por vetores-suporte, regressão por mínimos quadrados e regressão de ângulo mínimo) foram aplicados para relacionar as mutações da protease e da transcriptase reversa do HIV-1 à susceptibilidade de 16 ARVs. A suscetibilidade aos ARVs foi expressa em função da IC50 e seus resultados foram classificados em suscetível, resistência baixa/intermediária e resistência elevada. Os métodos de aprendizagem foram treinados e testados utilizando a validação cruzada do tipo *k-fold* (k = 5). Cada validação cruzada foi realizada 10 vezes para estimar a variabilidade da acurácia média. Para cada método de aprendizagem, quatro conjuntos de mutação foram utilizados como dados de entrada: um conjunto completo com todas as mutações presentes em mais de 2 sequências do banco de dados, um conjunto com as 30 mutações mais comuns, um conjunto de mutações selecionadas por um painel de especialistas, e um conjunto de mutações não polimórficas (que não podem ser separadas em classes distintas e bem definidas). A acurácia média dos métodos variou de 76,1% para as redes neurais a 79,7% para a regressão de ângulo mínimo. Os métodos de regressão (79,9%) apresentaram maior acurácia para os PIs do que as árvores de decisão e as redes neurais (75,5%). O conjunto de mutações com melhor desempenho foi o de mutações não polimórficas

(80,1%), seguido pelo conjunto de mutações selecionadas por um painel de especialistas (77,5%) e o conjunto completo (76,1%). Em relação aos PIs, o coeficiente de correlação (R^2) médio entre os coeficientes da regressão por vetores-suporte e da regressão por mínimos quadrados foi igual a 0,98 e entre os coeficientes da regressão por mínimos quadrados e da regressão de ângulo mínimo foi de 0,96. Para os NRTIs, o coeficiente de correlação (R^2) médio entre os coeficientes da regressão por vetores-suporte e da regressão por mínimos quadrados foi igual a 0,94 e entre os coeficientes da regressão por mínimos quadrados e da regressão de ângulo mínimo foi de 0,91.

Em 2007, Larder *et al.* [48] desenvolveram modelos de predição baseados em redes neurais com o objetivo de predizer a resposta virológica. Foram desenvolvidos 3 tipos de modelos classificados de acordo com as variáveis de entrada. O primeiro tipo de modelo recebeu 71 variáveis explicativas: 55 mutações selecionadas da literatura (HIV-protease e HIV-transcriptase reversa), 14 ARVs, carga viral *baseline* e tempo de acompanhamento da carga viral, sendo classificados pelo artigo como "modelos padrões". O segundo modelo, além de apresentar as 71 variáveis descritas anteriormente, incluíram também a contagem de células T CD4+ *baseline* e 4 variáveis relacionadas à história de tratamento: uso anterior de Zidovudina, Lamivudina, qualquer NNRTI e qualquer PI. Estes foram nomeados de "modelos aumentados". O terceiro tipo de modelo foi classificado como *"HAART-only"*, apresentando as mesmas variáveis de entrada do modelo aumentado, mas envolvendo apenas as amostras com uso de três ou mais drogas. Para cada categoria, dez redes neurais foram treinadas e validadas usando o método *leave-n-out*, em que 10% dos dados de treinamento foram selecionados aleatoriamente como conjunto de validação e o restante usado para o treinamento das redes. O desempenho dos modelos foi avaliado através do coeficiente de correlação (R^2) entre a resposta virológica predita e a real do conjunto de teste. Os modelos padrões apresentaram um R^2 médio igual a 0,53, enquanto que os modelos aumentados e *HAART-only* apresentaram valores superiores, iguais a 0,69 e 0,61, respectivamente. Neste estudo, valores de acurácia, sensibilidade e especificidade não foram apresentados.

Em 2007, Bonet *et al.* [10] demonstraram em seu estudo que a energia dos aminoácidos da proteína pode ser utilizada no desenvolvimento de modelos preditivos, uma vez que a energia envolvida em um aminoácido ou entre aminoácidos pode auxiliar na representação do genótipo do HIV por se aproximar da configuração tridimensional. Neste estudo, a energia associada a cada aminoácido da protease e a variação de energia entre os aminoácidos da sequência original e da analisada foram utilizadas como característica para predizer a resistência aos ARVs. Os métodos SVM, redes neurais multicamadas (*Multilayers Perceptrons*, MLPs) e redes neurais recorrentes bidimensionais (*Bidirectional Recurrent Neural Networks*, BRNN) fo-

ram aplicados na construção de classificadores para sete inibidores da HIV-protease: APV, ATV, NFV, RTV, SQV, LPV e IDV. A variável resposta foi codificada em relação à quantidade necessária do ARV para inibir a HIV-protease: valores menores que 3,5 foram codificados como suscetíveis e maiores que o ponto de corte foram classificados como resistentes. Para a avaliação dos resultados, a validação cruzada do tipo *10-fold* foi aplicada ao banco de dados. A rede MLP com entrada igual à energia dos aminoácidos apresentou uma acurácia de 80,00% a 92,33% para os ARVs enquanto que a mesma rede, mas com entrada igual à diferença de energia, apresentou valores de acurácia entre 74,16% a 92,55%. Em relação à SVM, utilizando a energia como representação dos aminoácidos, sua acurácia variou desde 68,59% a 92,57%. Para a rede BRNN, a acurácia variou de 81,33% a 94,73% e a codificação foi a variação de energia. Quando avaliado a sensibilidade e a especificidade das redes MLP e BRNN, estas apresentaram resultados próximos, sugerindo que os dois tipos de rede podem ser utilizados neste objetivo. A rede MLP apresentou sensibilidade entre 80,95% e 95,72% e especificidade entre 73,33% e 94,52%. A rede BRNN, por sua vez, apresentou sensibilidade entre 84,81% e 96,25% e especificidade entre 77,78% e 100%. Foi demonstrado que a energia dos aminoácidos assim como a diferença de energia entre a mutação e o aminoácido original são boas características para representar o genótipo do HIV e a BRNN pode ser usada como um classificador.

Outro estudo que também fez uso das ANNs foi o de Pasomsub *et al.* [12], em 2010, utilizando a técnica para predizer a resistência fenotípica a partir do genótipo de 14 ARVs (PIs e inibidores da RT). A rede utilizada foi do tipo *feedforward* e as sequências do HIV-1 foram treinadas, validadas e testadas. O desempenho do modelo foi obtido através da validação cruzada do tipo *10-fold*. A acurácia dos modelos e o coeficiente de correlação de postos de Spearman foram calculados. A acurácia dos modelos também foi comparada entre os subtipos virais. Os resultados encontrados foram satisfatórios, indicando que a resistência aos ARVs pode ser predita por meio da sequência das proteínas-alvo das drogas. Os maiores coeficientes de correlação foram observados para os PIs (0,87 - 0,91) e NRTIs (0,78 - 0,90), seguidos pelos NNRTIs (0,69 - 0,79). As acurácias dos modelos foram altas para todas as classes de drogas: 91-96% para PIs, 84-93% para NRTIs e 92-96% para NNRTIs. Quando avaliados os modelos preditivos de resistência segundo o subtipo viral, a acurácia foi maior para o subtipo B em relação aos dos subtipos não B. Porém, a diferença não foi estatisticamente significante. O estudo também comparou a concordância dos fenótipos obtidos de maneira experimental com a classificação fornecida pelas ANNs e pelos sistemas de interpretação *Geno2pheno* e *Stanford HIVdb*. A análise da curva ROC mostrou uma pequena diferença entre os três classificadores para os NNRTIs e PIs. Porém, para as classes dos NRTIs, a diferença entre o modelo da ANN e dos sistemas de interpretação foi significantemente maior.

No estudo de Silva [49], foi desenvolvido um modelo computacional híbrido composto pelo algoritmo genético para a seleção de atributos e pelo classificador de *kernel* de Fisher para a seleção de variáveis. A hidrofobicidade foi aplicada na codificação das sequências de aminoácidos da HIV-protease. O modelo mostrou-se promissor na previsão de resistência em sequências da HIV-protease de pacientes portadores do HIV-1 dos subtipos B e C em falha terapêutica no Brasil para os inibidores SQV, NFV e LPV. Além disso, o estudo foi capaz de identificar novas mutações de resistência no gene da protease do genoma do HIV-1, responsáveis por aumentar o desempenho dos modelos preditivos dos três ARVs em questão. A incorporação das posições de mutações mais frequentes selecionadas pelo modelo no classificador de *kernel* de Fisher possibilitou um desempenho superior em todas as métricas utilizadas, principalmente a sensibilidade, onde o valor médio obtido foi maior que 99% em pacientes portadores do HIV-1 de subtipo B e maior que 95% para o subtipo C.

Diversos são os métodos propostos com o intuito de desenvolver um modelo capaz de predizer a resistência aos ARVs com bom desempenho. Como pode ser observado, os trabalhos têm apresentado metodologias e resultados diferentes, variando desde modelos com baixo poder preditivo até classificadores com altos desempenhos. Essa variação de resultados, além de depender do método aplicado, está relacionada tanto aos dados utilizados no desenvolvimento dos modelos, bem como ao próprio tratamento aplicado a esses dados.

Cabe ainda a realização de novos estudos que sejam capazes de estabelecer um modelo de aplicação mais ampla, cujo desempenho não esteja limitado ao banco de dados aplicados no desenvolvimento do modelo.

Capítulo 4

Materiais e Métodos

4.1 Conjunto de Dados

O banco de dados utilizado neste estudo é constituído por sequências de aminoácidos da enzima protease do gene pol (polimerase) do HIV-1 de 625 pacientes infectados por este vírus, subtipo B.

Além das sequências da HIV-protease, o conjunto de dados possui informações referentes à contagem de linfócitos T CD4+ e carga viral no último período de tratamento, indicação se o paciente apresentou ou não falha terapêutica ao medicamento, ano de diagnóstico da infecção pelo HIV, data da genotipagem, estado de origem do paciente, sexo e idade. A identificação de falha terapêutica foi realizada a partir do Algoritmo Brasileiro, um programa baseado em regras que classifica o nível de resistência com base na sequência genotípica [http://algoritmo.aids.gov.br/atualizacao_algoritmo/site/]. As informações básicas referentes aos pacientes incluídos no estudo estão descritas na Tabela 4.1.

Os dados foram cedidos pelo Laboratório de Virologia Molecular do Centro de Ciências da Saúde da Universidade Federal do Rio de Janeiro (CCS - UFRJ/Brasil), integrante da rede de laboratórios de genotipagem do Ministério da Saúde (RENAGENO).

Tabela 4.1: Resumo das características clínicas dos pacientes (n=625).

Características	*Valores*
Idade média (anos), ±dp	38,68 (13,72)
Sexo masculino, %	340 (54,40)
Carga viral média (cópias/ml), IQR	4,60 (4,11 - 5,02)
Contagem de T CD4+ (céls./mm^3), IQR	299,4 (124,5 - 407,0)

dp: desvio-padrão
IQR: intervalo interquartílico

Neste estudo, os ARVs selecionados para o desenvolvimento dos modelos preditivos foram o LPV e NFV. A utilização desses inibidores se deu em função primeiramente do número de pacientes experimentados no último regime terapêutico,

4.2 Pré-processamento dos Dados

4.2.1 Normalização

Normalizar significa pré-processar os dados de entrada para que todos tenham o mesmo peso e medida, evitando que dados com valores muito elevados ou muito baixos produzam resultados com distribuições enviesadas degradando o modelo. Neste estudo, as variáveis com valores contínuos foram normalizadas com média 0 (zero) e desvio-padrão 1.

4.2.2 Separação em Conjunto de Treino e Teste

Os dados foram selecionados de maneira aleatória e divididos em dois conjuntos: conjunto de treino, para selecionar as variáveis dos modelos e seus parâmetros e, conjunto de teste, utilizado para determinar o desempenho dos classificadores com dados que não foram previamente utilizados. Na Figura 4.1 está representado o fluxograma da divisão dos dados para cada ARV segundo o número de pacientes com resistência e sem resistência.

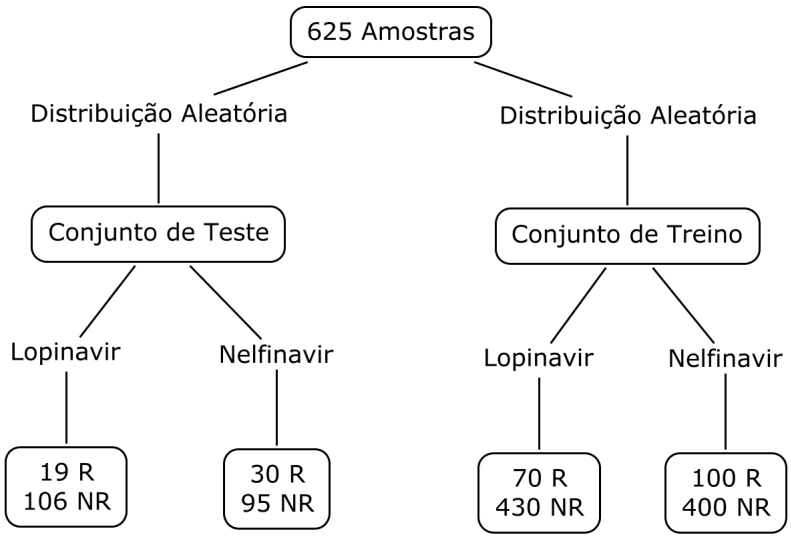

Figura 4.1: Divisão dos dados em conjunto de teste e conjunto de treino.
R: Resistentes e NR: Não Resistentes.

4.2.3 Codificação dos Aminoácidos

Os aminoácidos foram codificados de 2 maneiras: codificação binária e codificação segundo a escala de Eisenberg.

Na codificação binária, as posições que apresentam alguma mutação (aminoácido diferente do original) são codificadas com valor 1 enquanto que as posições que possuem o aminoácido na forma consenso recebem valor 0. Nesta codificação, as mutações não são diferenciadas entre si, indicando apenas a presença ou não de mutação.

Na codificação segundo a escala de Eisenberg, cada aminoácido recebe um valor de acordo com o seu nível de hidrofobicidade, possibilitando a diferenciação entre eles. Na Tabela 4.2 está representada a escala de Eisenberg.

Tabela 4.2: Escala de hidrofobicidade de Eisenberg.

Aminoácido	*Símbolo*	*Valor*	*Categoria*
Isoleucina	I	1,38	Hidrofóbico
Valina	V	1,08	Hidrofóbico
Leucina	L	1,06	Hidrofóbico
Fenilalanina	F	1,19	Hidrofóbico
Cisteína	C	0,29	Hidrofóbico
Metionina	M	0,64	Hidrofóbico
Alanina	A	0,62	Hidrofóbico
Glicina	G	0,48	Neutro
Treonina	T	-0,05	Neutro
Serina	S	-0,36	Neutro
Triptofano	W	0,81	Neutro
Tirosina	Y	0,26	Neutro
Prolina	P	0,12	Neutro
Histidina	H	-0,40	Hidrofílico
Glutamina	Q	-0.85	Hidrofílico
Asparagina	N	-0,78	Hidrofílico
Ácido Glutâmico	E	-0,74	Hidrofílico
Ácido Aspártico	D	-0,90	Hidrofílico
Lisina	K	-1,50	Hidrofílico
Arginina	R	-2,53	Hidrofílico

4.3 Seleção das Variáveis

Foi considerada como variável resposta a resistência aos inibidores LPV e NFV. Para os pacientes que, no último regime terapêutico, não utilizaram ou não apresentaram resistência ao inibidor da HIV-protease, a variável resposta foi codificada com o valor 0, enquanto que aqueles que apresentaram resistência ao ARV, a variável foi codificada com o valor 1. A análise foi realizada separadamente para cada ARV.

Dois conjuntos diferentes de variáveis explicativas foram utilizados neste estudo:

- (1) as posições com mutações mais frequentes encontradas no gene da HIV-protease associadas à resistência a cada ARV, a carga viral e a contagem de linfócitos TCD4+, ambos no último regime terapêutico. A seleção das posições mais frequentes baseou-se nos dados do artigo de Johnson [50], uma atualização da Sociedade Internacional Antiviral (*International Antiviral Society, IAS-USA*) que lista as mutações associadas à resistência aos ARVs. As posições são representadas por uma anotação que consiste na letra correspondente ao nome do aminoácido da sequência consenso, seguido pelo número referente à posição do aminoácido. As posições selecionadas foram: L10, K20, L24, V32, L33, M46, I47, I50, F53, I54, L63, A71, G73, L76, V82, I84 e L90 para o LPV e L10, D30, M36, M46, A71, V77, V82, I84, N88 e L90 para o NFV. Este conjunto foi denominado "Lista da IAS".

- (2) as posições da HIV-protease com mais de 1% de variabilidade entre as amostras (62 de 99 posições), a carga viral e a contagem de linfócitos TCD4+, ambos no último regime terapêutico. Esta variabilidade no resíduo de aminoácido depende fundamentalmente de dois fatores: ocorrência de mutações de resistência à terapia antirretroviral e polimorfismos característicos do subtipo viral. Inicialmente, cada exemplo é representado pelos 99 aminoácidos da sequência da HIV-protease. Visando à exclusão das posições com pequena variação, foram consideradas somente as posições com variação mínima de 1% dentre os exemplos, que é a taxa estimada de mutação no HIV-1 ao ano. Este conjunto foi denominado "Todas as Posições".

Para selecionar as variáveis dos modelos preditivos, foram aplicados métodos distintos para cada classificador.

4.3.1 Regressão Logística

A seleção das variáveis foi realizada através das técnicas *bootstrap* e *stepwise*. Para o LPV, dos 70 pacientes resistentes do conjunto de treino, 70 amostras foram selecionadas por *bootstrap*, enquanto que dos 430 pacientes não resistentes do conjunto

de treino, 86 foram selecionados aleatoriamente com reposição, a fim de se obter um conjunto de dados próximo do balanceamento. Para o NFV, a partir das amostras resistentes do conjunto de treino, foram obtidas 100 amostras por *bootstrap* combinadas com 100 amostras não resistentes do conjunto de treino selecionadas aleatoriamente com reposição, resultando em um conjunto equilibrado de dados.

Esse procedimento foi repetido 1000 vezes, resultando em 1000 subconjuntos para cada ARV. Para cada um desses subconjuntos, um modelo logístico foi desenvolvido e as variáveis de cada um desses modelos foram selecionadas pelo método *stepwise*. Para a escolha das variáveis finais, dois critérios de seleção foram definidos: (1) "Critério $\geq 50\%$", em que as variáveis presentes em 50% ou mais dos modelos são eleitas para compor os modelos finais e (2) "Critério $\geq 60\%$" , em que as variáveis presentes em 60% ou mais dos modelos são selecionadas.

4.3.2 Rede Neural Probabilística

O conjunto de treino de cada ARV foi utilizado para a seleção das variáveis explicativas e do fator de alisamento de cada PNN. Nesta metodologia, foram aplicadas as técnicas *bootstrap*, validação cruzada e SFS.

A partir do conjunto de treino, ao invés de 1000, 100 subconjuntos foram obtidos da mesma forma descrita acima na seleção das variáveis da RL. Para cada um desses subconjuntos, foi implementada uma rede PNN. O melhor conjunto de variáveis foi obtido utilizando o método de seleção SFS e a AUC.

Inicialmente, foram desenvolvidas redes PNN para cada variável de entrada. A AUC média foi calculada através da técnica de validação cruzada do tipo *k-fold* com k igual a 10. A variável com melhor valor de AUC médio foi selecionada. A partir deste passo, todos os possíveis vetores bidimensionais contendo o vencedor do passo anterior foram formados.

Novas redes PNN, em cada caso, foram treinadas e validadas, a AUC média foi calculada e, como anteriormente, o vetor que resultou em maior AUC média foi selecionado. A cada passo uma nova variável foi adicionada, gerando a finalização do algoritmo quando o n-ésimo vetor calculado a partir da n-ésima etapa não melhorou a AUC. Este processo é repetido para cada valor do fator de alisamento e o modelo com maior AUC é escolhido, armazenando as variáveis que foram selecionadas e o valor do fator de alisamento correspondente. Este procedimento é repetido para cada uma das 100 amostras *bootstrap*, sendo computado o número de vezes que cada variável de entrada é selecionada. As variáveis de entrada finais são aquelas selecionadas pelos dois critérios definidos previamente ($\geq 50\%$ e $\geq 60\%$).

A escolha do melhor fator de alisamento ocorreu juntamente com a seleção das variáveis. Cada um dos 100 subconjuntos balanceados gerou redes PNN com fatores

de alisamento variando de 0,1 a 1, com passos de 0,1. Aqueles que apresentaram melhor AUC média tiveram o seu fator de alisamento selecionado. O fator de alisamento final foi definido como a média dos 100 melhores fatores provenientes de cada um dos modelos. Na Figura 4.2 está representado o esquema de seleção das variáveis e dos parâmetros dos modelos.

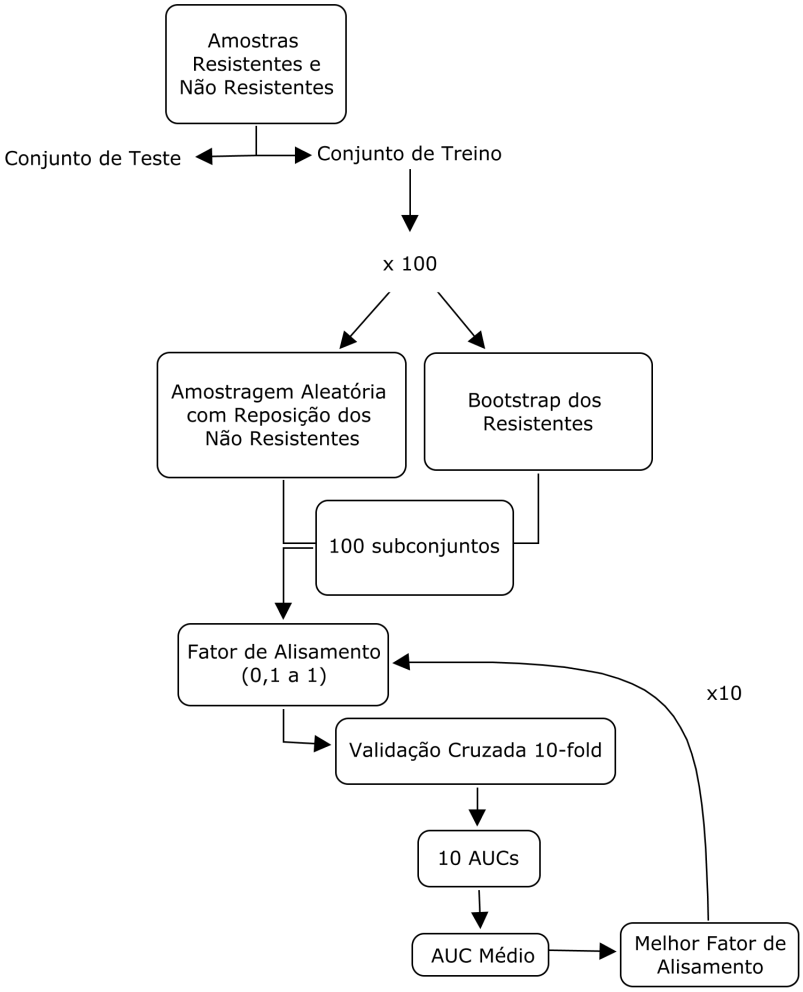

Figura 4.2: Esquema resumido para a seleção de variáveis das redes neurais probabilísticas.

4.4 Ponto de Corte

Para ajudar a decidir se o resultado de um teste é positivo ou negativo, um ponto de corte deve ser escolhido. Os resultados que estão acima do limiar são considerados anormais enquanto que os resultados que estão abaixo do ponto de corte são considerados normais. No entanto, nem todos aqueles que estão acima do ponto de

corte terão necessariamente o diagnóstico positivo e nem todos aqueles abaixo do ponto de corte apresentarão um resultado negativo.

Uma vez que o aumento na sensibilidade de um teste implica, na maioria das vezes, em uma redução da especificidade e vice-versa, a escolha do ponto de corte deve estar relacionada com o objetivo principal a que se destina o teste.

Em geral, o ponto de corte utilizado nos estudos é o de 0,5, em que valores acima são considerados positivos e valores abaixo são classificados como negativos.

Neste presente trabalho, além da utilização do ponto de corte de 0,5, para cada grupo de modelos foi escolhido um melhor ponto de corte. Este foi definido como a melhor combinação entre os valores de sensibilidade e especificidade, correspondendo à maior soma entre os pares.

4.5 Algoritmos *Stanford HIVdb* e *Rega*

Os modelos obtidos pela RL e pela PNN foram comparados com os algoritmos de interpretação *Stanford HIVdb* (versão 6.2.0) [51] e *Rega* (versão 8.0.2) [52]. As posições utilizadas nos dois algoritmos foram todas aquelas consideradas por eles relevantes para o desenvolvimento de resistência. Ambos tiveram seus desempenhos obtidos através do mesmo conjunto de teste utilizado na avaliação dos modelos desenvolvidos.

No algoritmo *Stanford HIVdb*, as posições mutadas são fornecidas ao programa e este disponibiliza o resultado em cinco níveis de resistência: suscetível, resistência potencialmente baixa, baixa, intermediária e elevada. O algoritmo *Rega*, por outro lado, fornece o ponto de corte para definir o nível de resistência. Cada posição mutada possui um valor e a soma final fornece o nível de resistência. Valores abaixo de 1,25 são considerados suscetíveis, valores maiores ou iguais a 1,25 e menores que 2,0 representam uma resistência intermediária enquanto que valores maiores ou iguais a 2,0 indicam resistência elevada.

4.6 Softwares

Os dados foram compilados em planilha eletrônica do programa Microsoft Excel® e para as análises estatísticas foram utilizados os *softwares* R versão 3.0.1 na construção dos modelos logísticos e MATLAB versão R2009b, que possui a função *newpnn*, própria para a criação de redes PNN.

Capítulo 5

Resultados

Foram desenvolvidos 8 tipos de modelos, que variaram segundo as variáveis explicativas iniciais, a codificação dos aminoácidos e o ARV.

As variáveis explicativas iniciais foram: as posições com mutações mais frequentes, segundo a lista da IAS, encontradas no gene da HIV-protease associadas à resistência a cada ARV, a carga viral e a contagem de linfócitos TCD4+, cujo conjunto foi intitulado de "Lista da IAS" e; as posições da HIV-protease com mais de 1% de variabilidade entre as amostras (62 de 99 posições), a carga viral e a contagem de linfócitos T CD4+, sendo este conjunto intitulado de "Todas as Posições". As codificações utilizadas para os aminoácidos foram a binária e a de Eisenberg, e os ARVs foram o LPV e NFV.

Dessa forma, o estudo se divide em quatro grupos para cada ARV: (1) "Lista da IAS e Codificação Binária"; (2) "Lista da IAS e Codificação de Eisenberg"; (3) "Todas as Posições e Codificação Binária" e (4) "Todas as Posições e Codificação de Eisenberg".

5.1 Lopinavir

5.1.1 Modelos de Regressão Logística

5.1.1.1 Lista da IAS e Codificação Binária

As variáveis selecionadas na RL para o LPV foram apenas as posições I47, I54 e L90, e também a carga viral (CV) e a contagem de linfócitos TCD4+ (CD4), para o critério$\geq$ 50%. O melhor ponto de corte encontrado para esses modelos foi igual a 0,345. Para o critério$\geq$ 60%, as mesmas variáveis foram selecionadas, com exceção da CD4. Neste caso, o melhor ponto de corte foi igual a 0,383. Na Figura 5.1 está representado o gráfico das frequências de seleção de cada variável inicialmente incorporada aos modelos.

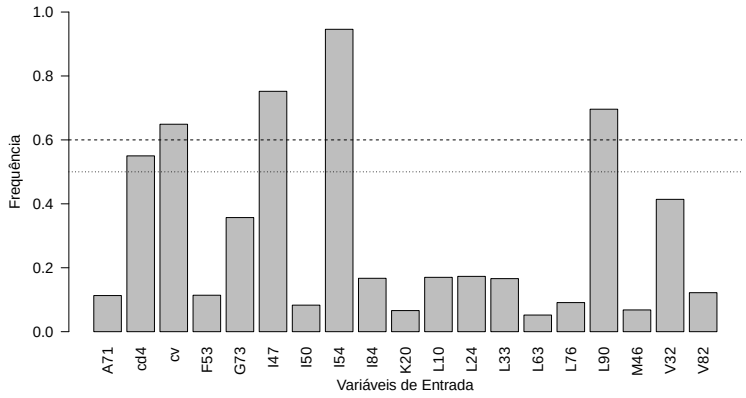

Figura 5.1: Frequência das variáveis nos modelos logísticos para o Lopinavir utilizando a codificação binária e o conjunto "Lista da IAS".

Os cinco modelos obtidos pela RL para o critérios de seleção de variáveis $\geq 50\%$ apresentaram as seguintes equações e AICs:

Modelo 1: $Resistência = -0,74 + 1,01I47 + 2,27I54 - 1,00L90 - 0,15CD4 + 0,37CV$, com um AIC igual a 189,8.

Modelo 2: $Resistência = -0,70 + 17,33I47 + 1,72I54 - 0,86L90 - 0,17CD4 + 0,62CV$, com um AIC de 181,3.

Modelo 3: $Resistência = -084, +2,66I47 + 3,37I54 - 1,69L90 - 0,74CD4 + 0,30CV$, e AIC igual a 165,2.

Modelo 4: $Resistência = -0,64 + 1,11I47 + 2,42I54 - 1,21L90 + 0,04CD4 + 0,33CV$, e AIC igual a 185,9.

Modelo 5: $Resistência = -0,83 + 1,63I47 + 2,24I54 - 0,65L90 - 0,29CD4 + 0,47CV$, e AIC igual a 182,4.

Os cinco modelos para o critério$\geq$ 60% apresentaram as seguintes equações e AICs:

Modelo 1: $Resistência = -0,74 + 1,00I47 + 2,26I54 - 0,98L90 + 0,41CV$, com um AIC igual a 188,3.

Modelo 2: $Resistência = -0,70 + 17,31I47 + 1,68I54 - 0,79L90 + 0,66CV$, com um AIC de 180,1.

Modelo 3: $Resistência = -0,77 + 2,03I47 + 3,08I54 - 1,42L90 + 0,41CV$, e AIC igual a 173,2.

Modelo 4: $Resistência = -0,64 + 1,11I47 + 2,42I54 - 1,21L90 + 0,32CV$, e AIC igual a 183,9.

Modelo 5: $Resistência = -0,82 + 1,60I47 + 2,19I54 - 0,58L90 + 0,53CV$, e AIC igual a 182,6.

Na Tabela 5.1 está representada a média do desempenho dos modelos logísticos em termos de AUC, acurácia, sensibilidade, especificidade e índice *Kappa* de acordo

com os pontos de corte e com os critérios de seleção ($\geq 50\%$ e $\geq 60\%$). Em média, os modelos logísticos para o LPV apresentaram AUC igual a 0,82 e 0,79 para os critérios $\geq 50\%$ e $\geq 60\%$, respectivamente. Para o ponto de corte de 0,345, em média, a acurácia foi igual a 0,66, a sensibilidade foi de 0,91, a especificidade igual a 0,62 e o índice *Kappa* foi igual a 0,29, indicando uma concordância razoável. Para o ponto de corte de 0,383, em média, a acurácia foi igual a 0,70, a sensibilidade foi de 0,71, a especificidade igual a 0,70 e o índice *Kappa* foi igual a 0,26, indicando uma concordância razoável.

Na Figura 5.2 estão representadas as curvas ROC dos cinco modelos logísticos deste grupo.

Tabela 5.1: Desempenho médio dos modelos logísticos para o Lopinavir utilizando a codificação binária e o conjunto "Lista da IAS" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,345	*PC: 0,500*	*PC: 0,383*	*PC: 0,500*
AUC	0,82±0,03	0,82±0,03	0,79±0,01	0,79±0,01
Acurácia	0,66±0,03	0,79±0,02	0,70±0,02	0,79±0,03
Sensibilidade	0,91±0,06	0,58±0,07	0,71±0,06	0,54±0,07
Especificidade	0,62±0,03	0,82±0,02	0,70±0,03	0,83±0,03
Índice Kappa	0,29±0,05	0,33±0,05	0,26±0,03	0,31±0,07

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

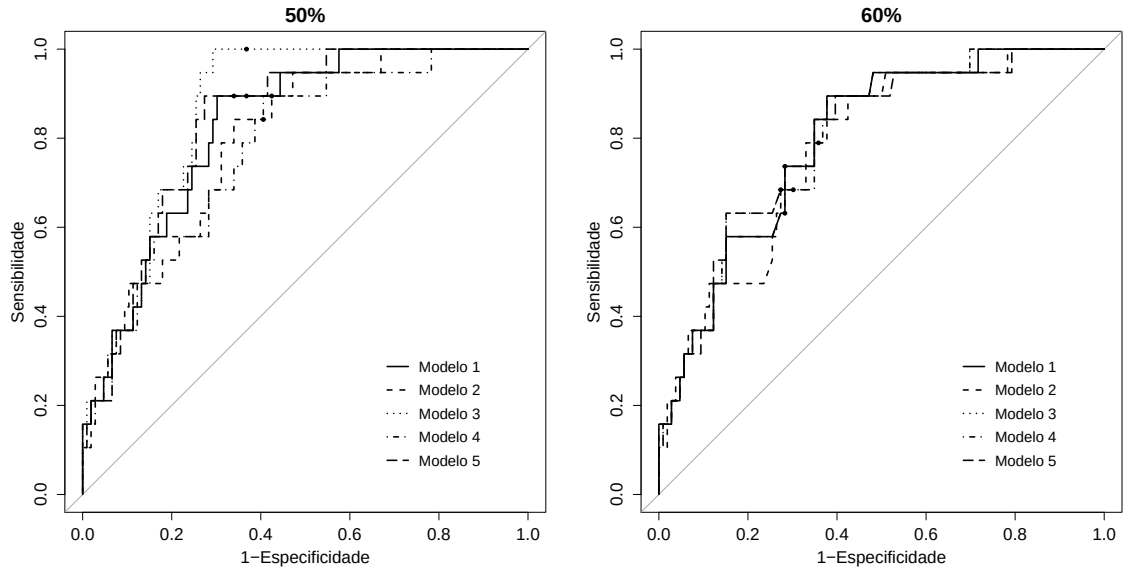

Figura 5.2: Curvas ROC dos modelos logísticos para o Lopinavir utilizando a codificação binária e o conjunto "Lista da IAS" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.1.1.2 Lista IAS e Codificação de Eisenberg

Neste grupo, as variáveis selecionadas na RL para o LPV em ambos os critérios de seleção de variáveis foram as posições V32, I47, I54 e L90, e CV. O melhor ponto de corte encontrado foi igual a 0,367. Na Figura 5.3 estão representadas as frequências de seleção de cada variável.

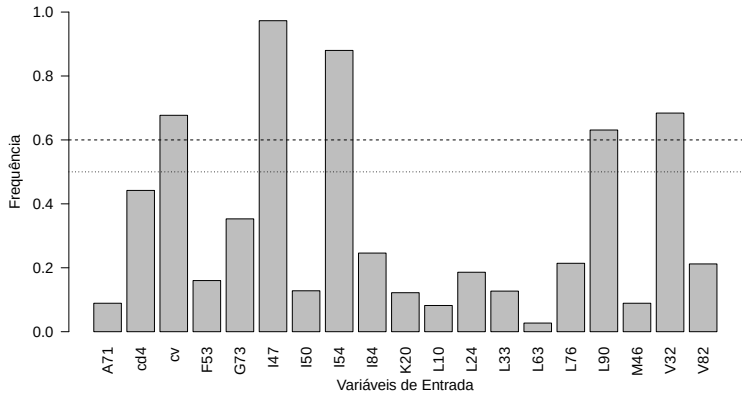

Figura 5.3: Frequência das variáveis nos modelos logísticos para o Lopinavir utilizando a codificação de Eisenberg e o conjunto "Lista da IAS".

Os cinco modelos obtidos neste grupo apresentaram as seguintes equações e AICs:

Modelo 1: $Resistência = 0,49 - 3,65V32 - 5,50I47 - 1,11I54 + 0,21L90 + 0,39CV$, com um AIC igual a 191,3.

Modelo 2: $Resistência = 1,45 - 3,05V32 - 9,95I47 - 1,03I54 + 0,21L90 + 0,70CV$, com um AIC de 178,6.

Modelo 3: $Resistência = 0,66 - 2,59V32 - 5,13I47 - 1,46I54 + 0,27L90 + 0,35CV$, e AIC igual a 188,1.

Modelo 4: $Resistência = 0,43 - 4,21V32 - 6,62I47 - 1,78I54 + 0,51L90 + 0,36C$V, e AIC igual a 176,4.

Modelo 5: $Resistência = 0,51 - 2,97V32 - 4,96I47 - 1,60I54 + 0,18L90 + 0,52CV$, e AIC igual a 177,9.

Na Tabela 5.2, pode-se observar o desempenho médio dos modelos logísticos para os pontos de corte iguais a 0,367 e 0,500. Em média, os modelos apresentaram AUC igual a 0,82. Para ambos os critérios de seleção de variáveis e para o ponto de corte de 0,367, os valores médios foram: acurácia igual a 0,68, sensibilidade de 0,80, especificidade de 0,66 e índice *Kappa* igual a 0,27 (concordância razoável). Para o ponto de corte igual a 0,500, em média, a acurácia foi de 0,80, a sensibilidade e especificidade iguais a 0,56 e 0,85, respectivamente e o índice *Kappa* igual a 0,35, caracterizando uma concordância razoável.

Na Figura 5.4 estão representadas as curvas ROC dos cinco modelos logísticos.

Tabela 5.2: Desempenho médio dos modelos logísticos para o Lopinavir utilizando a codificação de Eisenberg e o conjunto "Lista da IAS" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,367	*PC: 0,500*	*PC: 0,367*	*PC: 0,500*
AUC	0,82±0,02	0,82±0,02	0,82±0,02	0,82±0,02
Acurácia	0,68±0,03	0,80±0,04	0,68±0,03	0,80±0,04
Sensibilidade	0,80±0,06	0,56±0,05	0,80±0,06	0,56±0,05
Especificidade	0,66±0,05	0,85±0,05	0,66±0,05	0,85±0,05
Índice Kappa	0,27±0,03	0,35±0,06	0,27±0,03	0,35±0,06

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

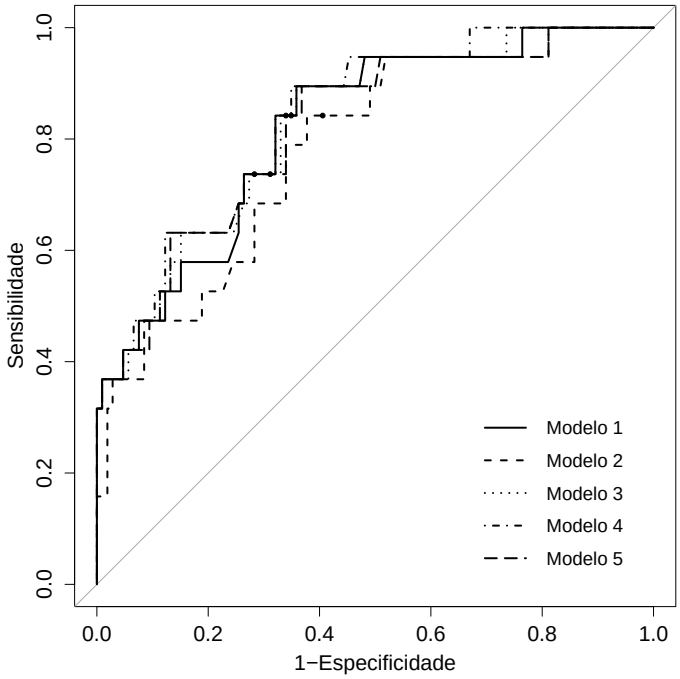

Figura 5.4: Curvas ROC dos modelos logísticos para o Lopinavir utilizando a codificação binária e o conjunto "Lista da IAS" para os dois critérios de seleção de variáveis.
Os círculos representam os pontos de corte encontrados para os critérios de seleção de variáveis.

5.1.1.3 Todas as Posições e Codificação Binária

As variáveis selecionadas na RL para o LPV quando assumido o critério de seleção $\geq 50\%$ foram as posições I13, E35, R41, K45, I47, I54, I62, G73 e L90, e também as variáveis CD4 e CV. O melhor ponto de corte encontrado foi igual a 0,389. Quando o critério de seleção foi $\geq 60\%$, as variáveis selecionadas foram apenas as posições I47, I54 e L90, e as variáveis clínicas CD4 e CV. O melhor ponto de corte, neste caso, foi igual a 0,345. Na Figura 5.5 está representado o gráfico das frequências de seleção de cada variável.

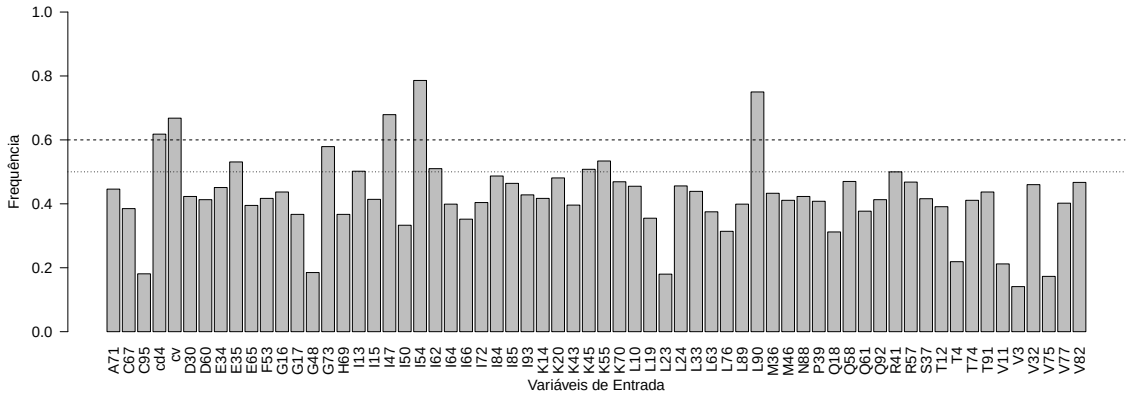

Figura 5.5: Frequência das variáveis nos modelos logísticos para o Lopinavir utilizando a codificação binária e o conjunto "Todas as Posições".

Os cinco modelos obtidos pela RL para o critério $\geq 50\%$ apresentaram as seguintes equações e AICs:

Modelo 1: $Resistência = -0,84 + 0,02I13 - 0,04E35 + 0,08R41 + 0,26K45 + 0,83I47 + 2,22I54 + 0,09I62 + 1,99G73 - 1,94L90 - 0,12CD4 + 0,36CV$, com um AIC igual a 195,7.

Modelo 2: $Resistência = -1,26 + 0,42I13 - 0,34E35 + 0,52R41 + 0,79K45 + 17,43I47 + 1,55I54 + 1,24I62 + 2,41G73 - 2,36L90 - 0,22CD4 + 0,62CV$, com um AIC de 175,6.

Modelo 3: $Resistência = -1,05 - 0,07I13 - 1,10E35 + 0,44R41 + 2,24K45 + 3,54I47 + 3,64I54 + 0,78I62 + 1,23G73 - 2,25L90 - 0,89CD4 + 0,22CV$, e AIC igual a 163,9.

Modelo 4: $Resistência = -1,19 + 0,35I13 + 0,10E35 + 0,60R41 + 0,88K45 + 0,83I47 + 2,18I54 + 0,57I62 + 0,64G73 - 1,54L90 - 0,04CD4 + 0,30CV$, e AIC igual a 192,9.

Modelo 5: $Resistência = -1,07 - 0,24I13 + 0,06E35 + 0,83R41 - 0,56K45 + 1,68I47 + 1,94I54 + 0,20I62 + 2,23G73 - 1,65L90 - 0,26CD4 + 0,43CV$, e AIC igual a 180,3.

Para o critério $\geq 60\%$, os modelos obtidos foram:

Modelo 1: $Resistência = -0,74 + 1,01I47 + 2,27I54 - 1,00L90 - 0,15CD4 + 0,37CV$, com um AIC igual a 189,8.

Modelo 2: $Resistência = -0,70 + 17,33I47 + 1,72I54 - 0,86L90 - 0,17CD4 + 0,62CV$, com um AIC de 181,3.

Modelo 3: $Resistência = -0,84 + 2,66I47 + 3,37I54 - 1,69L90 - 0,74CD4 + 0,29CV$, e AIC igual a 165,2.

Modelo 4: $Resistência = -0,64 + 1,11I47 + 2,42I54 - 1,21L90 - 0,04CD4 + 0,33CV$, e AIC igual a 185,9.

Modelo 5: $Resistência = -0,83 + 1,63I47 + 2,24I54 - 0,65L90 - 0,29CD4 + 0,47CV$, e AIC igual a 182,4.

Na Tabela 5.3 está representado o desempenho médio dos modelos logísticos em termos de AUC, acurácia, sensibilidade e especificidade para os dois critérios de seleção e seus respectivos pontos de corte. Em média, os modelos logísticos para os critérios $\geq 50\%$ e $\geq 60\%$ apresentaram AUC igual a 0,80 e 0,82, respectivamente. Para o ponto de corte de 0,389, a acurácia média foi igual a 0,70, sensibilidade e especificidade médias de 0,75 e 0,69, respectivamente, e índice *Kappa* médio igual a 0,28 (concordância razoável). Para o ponto de corte de 0,345, a acurácia média foi igual a 0,66, sensibilidade e especificidade médias de 0,91 e 0,62, respectivamente, e índice *Kappa* médio igual a 0,29 (concordância razoável).

Na Figura 5.6 estão representadas as curvas ROC dos cinco modelos logísticos para cada critério de seleção de variáveis.

Tabela 5.3: Desempenho médio dos modelos logísticos para o Lopinavir utilizando a codificação binária e o conjunto "Todas as Posições" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,389	*PC: 0,500*	*PC: 0,345*	*PC: 0,500*
AUC	0,80±0,04	0,80±0,04	0,82±0,03	0,82±0,03
Acurácia	0,70±0,04	0,79±0,02	0,66±0,03	0,79±0,02
Sensibilidade	0,75±0,40	0,63±0,06	0,91±0,06	0,58±0,07
Especificidade	0,69±0,05	0,82±0,05	0,62±0,03	0,82±0,02
Índice Kappa	0,28±0,07	0,35±0,06	0,29±0,05	0,33±0,05

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

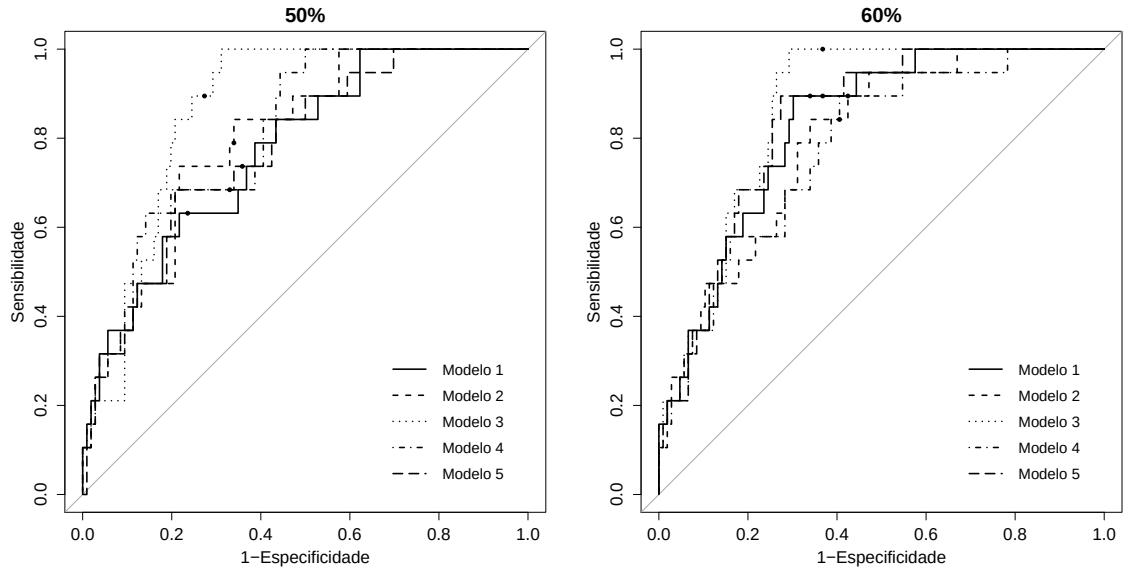

Figura 5.6: Curvas ROC dos modelos logísticos para o Lopinavir utilizando a codificação binária e o conjunto "Todas as Posições" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.1.1.4 Todas as Posições e Codificação de Eisenberg

Neste grupo, quando aplicado o critério de seleção $\geq 50\%$, as variáveis selecionadas foram as posições I13, K14, D30, V32, E35, R41, K45, I47, I54, R57, I62, I64, K70, G73, I84, I85, N88, L90, T91 e as variáveis CD4 e CV. O melhor ponto de corte definido foi igual a 0,447. Quando o critério foi $\geq 60\%$, apenas as posições K45, I47, I54, G73, L90 e T91, e a CV foram selecionadas. O melhor ponto de corte, neste caso, foi igual a 0,355. As frequências de seleção de cada variável na fase de treinamento estão representadas na Figura 5.7.

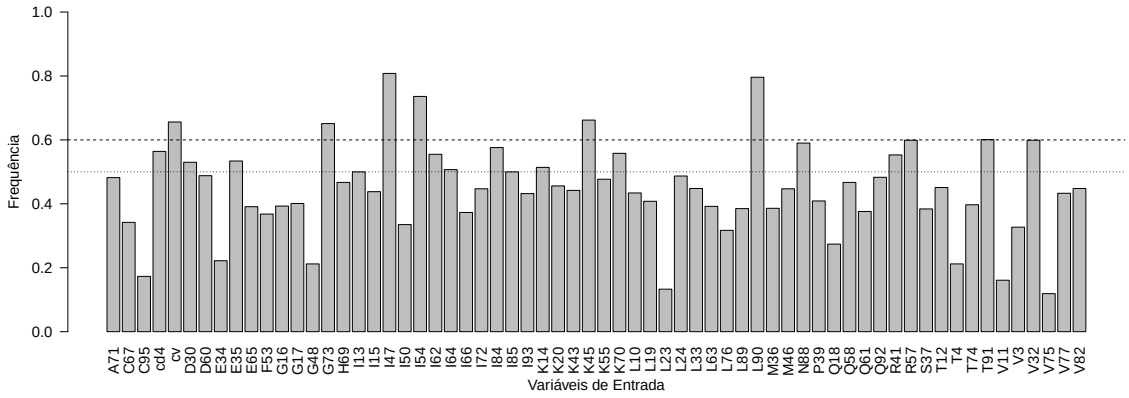

Figura 5.7: Frequência das variáveis nos modelos logísticos para o Lopinavir utilizando a codificação de Eisenberg e o conjunto "Todas as Posições".

Os cinco modelos obtidos pela RL quando utilizado o critério$\geq 50\%$ apresentaram as seguintes equações e AICs:

Modelo 1: $Resistência = 0,26 + 0,11I13 - 0,34K14 - 4,45D30 - 3,65V32 + 0,07E35 + 0,14R41 + 0,04K45 - 5,49I47 - 1,00I54 + 0,43R57 - 0,13I62 - 0,27I64 - 0,20K70 - 0,69G73 - 0,32I84 + 0,04I85 - 7,81N88 + 0,67L90 - 2,94T91 - 0,05CD4 + 0,38CV$, com um AIC igual a 201,6.

Modelo 2: $Resistência = -0,71 - 0,43I13 - 0,39K14 - 19,64D30 - 3,21V32 + 0,28E35 - 0,21R41 - 0,97K45 - 10,76I47 - 2,40I54 + 0,59R57 - 0,75I62 - 0,38I64 - 0,32K70 - 1,05G73 + 0,08I84 + 0,20I85 - 8,16N88 + 1,61L90 - 5,93T91 - 0,40CD4 + 0,78CV$, com um AIC igual a 156,4.

Modelo 3: $Resistência = 0,95 - 0,01I13 - 0,05K14 - 0,47D30 - 2,25V32 + 0,14E35 + 0,18R41 - 0,53K45 - 4,87I47 - 1,20I54 + 0,32R57 - 0,38I62 - 0,15I64 - 0,37K70 - 0,52G73 - 0,16I84 + 0,16I85 - 0,29N88 + 0,60L90 - 2,96T91 - 0,70CD4 + 0,21CV$, com um AIC igual a 187,5.

Modelo 4: $Resistência = 1,16 - 0,19I13 - 0,24K14 + 0,08D30 - 4,59V32 + 0,26E35 + 0,22R41 - 0,36K45 - 5,91I47 - 1,95I54 + 0,45R57 - 0,19I62 - 0,36I64 - 0,32K70 - 0,46G73 - 0,85I84 + 0,95I85 - 0,25N88 + 0,96L90 - 5,44T91 - 0,04CD4 + 0,60CV$, com um AIC igual a 177,7.

Modelo 5: $Resistência = 0,75 - 0,07I13 - 0,24K14 - 0,33D30 - 2,73V32 + 0,19E35 + 0,51R41 - 0,07K45 - 4,55I47 - 1,81I54 + 0,56R57 - 0,28I62 - 0,24I64 + 0,06K70 - 0,85G73 - 0,43I84 + 0,47I85 - 1,81N88 + 0,86L90 - 2,94T91 - 0,44CD4 + 0,41CV$, com um AIC igual a 174,6.

Quando utilizado o critério $\geq 60\%$, as equações e seus AICs correspondentes foram:

Modelo 1: $Resistência = 0,35 - 0,03K45 - 0,31I47 - 1,10I54 - 0,57G73 + 0,64L90 - 3,03T91 + 0,38CV$, com um AIC igual a 192,4.

Modelo 2: $Resistência = 0,82 - 0,52K45 - 5,28I47 - 1,60I54 - 0,78G73 + 0,98L90 - 6,05T91 + 0,76CV$, com um AIC igual a 160,3.

Modelo 3: $Resistência = 0,46 - 0,38K45 - 0,60I47 - 1,46I54 - 0,53G73 + 0,59L90 - 2,96T91 + 0,36CV$, com um AIC igual a 182.

Modelo 4: $Resistência = 0,63 - 0,33K45 - 0,39I47 - 1,21I54 - 0,22G73 + 0,49L90 - 5,03T91 + 0,33CV$, com um AIC igual a 186,8.

Modelo 5: $Resistência = 0,48 - 0,05K45 - 0,44I47 - 1,58I54 - 0,89G73 + 0,69L90 - 2,95T91 + 0,52CV$, com um AIC igual a 172,7.

Na Tabela 5.4 está representado o desempenho dos modelos logísticos em termos de AUC, acurácia, sensibilidade e especificidade para cada critério de seleção de variáveis e seus respectivos pontos de corte. Em média, para o critério $\geq 50\%$ e $\geq 60\%$, os modelos logísticos apresentaram AUCs iguais a 0,82 e 0,78, respectivamente. Para o ponto de corte de 0,447, o desempenho médio apresentou uma acurácia igual a 0,74, sensibilidade de 0,64, especificidade de 0,76 e índice *Kappa* igual a 0,28, indicando uma concordância razoável. Para o ponto de corte de 0,355, o desempenho médio apresentou uma acurácia igual a 0,67, sensibilidade de 0,73, especificidade igual a 0,66 e índice *Kappa* igual a 0,23, uma concordância também razoável segundo a escala de Landis.

Na Figura 5.8 estão representadas as curvas ROC dos cinco modelos logísticos para os dois critérios de seleção.

Tabela 5.4: Desempenho médio dos modelos logísticos para o Lopinavir utilizando a codificação de Eisenberg e o conjunto "Todas as Posições" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,447	*PC: 0,500*	*PC: 0,355*	*PC: 0,500*
AUC	0,82±0,05	0,82±0,05	0,78±0,02	0,78±0,02
Acurácia	0,74±0,02	0,76±0,04	0,67±0,02	0,81±0,04
Sensibilidade	0,64±0,11	0,59±0,08	0,73±0,04	0,53±0,04
Especificidade	0,76±0,04	0,79±0,05	0,66±0,03	0,86±0,05
Índice Kappa	0,28±0,05	0,23±0,04	0,23±0,02	0,34±0,05

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

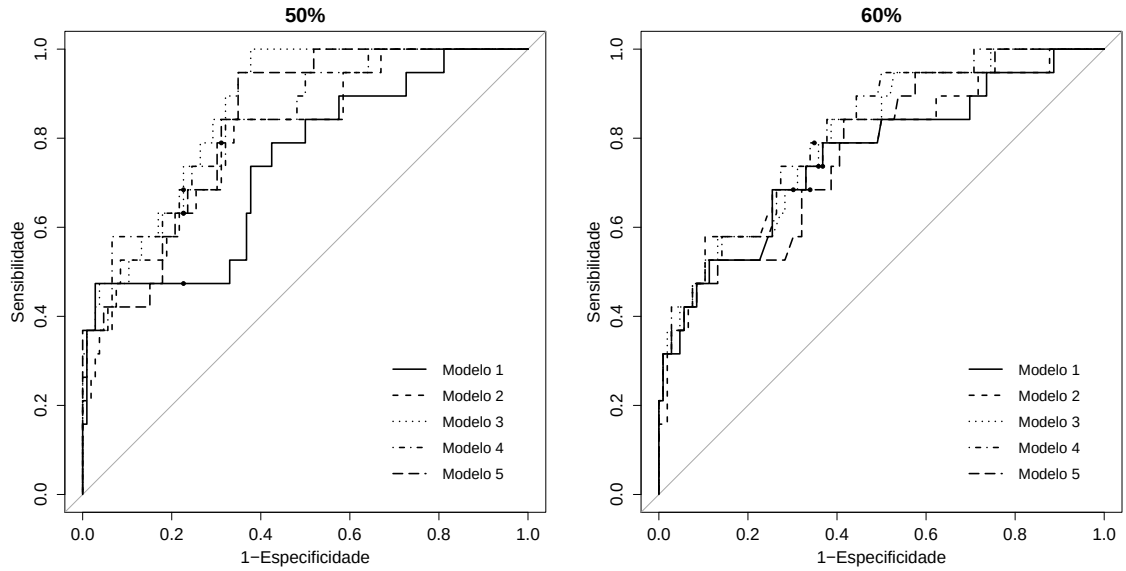

Figura 5.8: Curvas ROC dos modelos logísticos para o Lopinavir utilizando a codificação de Eisenberg e o conjunto "Todas as Posições" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.1.2 Modelos de Redes Neurais Probabilísticas

5.1.2.1 Lista da IAS e Codificação Binária

As variáveis selecionadas nas redes PNN para o LPV foram as posições I54 e L63, e as variáveis CD4 e CV para o critério de seleção $\geq 50\%$. Para o critério $\geq 60\%$, as mesmas variáveis foram selecionadas com exceção da posição L63. O melhor ponto de corte encontrado para esses modelos foi igual a 0,501 e 0,502 para o critério $\geq 50\%$ e $\geq 60\%$, respectivamente. Na Figura 5.9 está representado o gráfico das frequências de seleção de cada variável inicialmente incorporada aos modelos.

O fator de alisamento encontrado na fase de treinamento e aplicado nas redes PNN foi igual a 0,180.

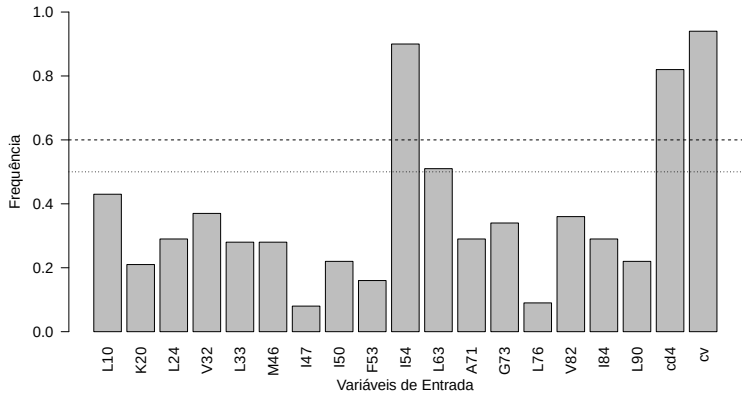

Figura 5.9: Frequência das variáveis nas redes neurais probabilísticas para o Lopinavir utilizando a codificação binária e o conjunto "Lista da IAS".

Na Tabela 5.5 está representado o desempenho dos modelos logísticos em termos de AUC, acurácia, sensibilidade e especificidade para cada critério de seleção de variáveis e seus respectivos pontos de corte. Em média, para os critérios $\geq 50\%$ e $\geq 60\%$, os modelos de PNN apresentaram AUC igual a 0,64. Para o ponto de corte de 0,501, o desempenho médio apresentou uma acurácia igual a 0,71, sensibilidade de 0,56, especificidade de 0,74 e índice *Kappa* igual a 0,21, indicando uma concordância razoável. Para o ponto de corte de 0,502, o desempenho médio apresentou uma acurácia igual a 0,69, sensibilidade de 0,61, especificidade de 0,71 e índice *Kappa* igual a 0,21, indicando uma concordância razoável. Para o ponto de corte de 0,500, os modelos obtidos apresentaram desempenhos muitos próximos.

Na Figura 5.10 estão representadas as curvas ROC dos cinco modelos obtidos pela PNN para os dois critérios de seleção de variáveis.

Tabela 5.5: Desempenho médio das redes neurais probabilísticas para o Lopinavir utilizando a codificação binária e o conjunto "Lista da IAS" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,501	*PC: 0,500*	*PC: 0,502*	*PC: 0,500*
AUC	0,64±0,08	0,64±0,08	0,63±0,06	0,63±0,06
Acurácia	0,71±0,02	0,66±0,04	0,69±0,03	0,67±0,03
Sensibilidade	0,56±0,08	0,61±0,10	0,61±0,05	0,61±0,05
Especificidade	0,74±0,01	0,67±0,04	0,71±0,03	0,68±0,03
Índice Kappa	0,21±0,06	0,18±0,08	0,21±0,06	0,19±0,06

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

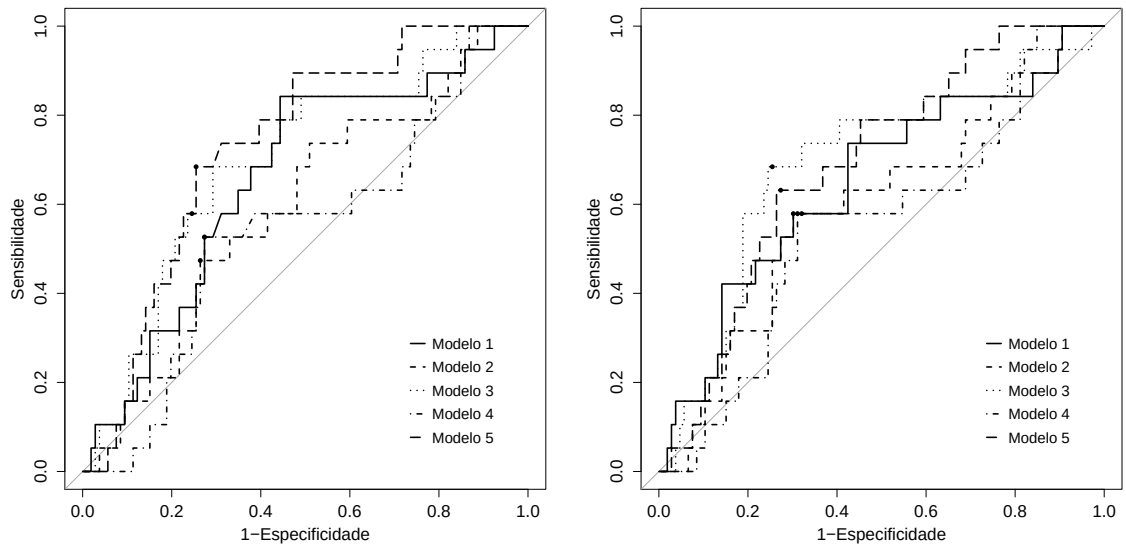

Figura 5.10: Curvas ROC das redes neurais probabilísticas para o Lopinavir utilizando a codificação binária e o conjunto "Lista da IAS" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.1.2.2 Lista da IAS e Codificação de Eisenberg

Neste grupo, as variáveis selecionadas nas redes PNN para ambos os critérios de seleção de variáveis foram a posição I54 e as variáveis clínicas CD4 e CV. O melhor ponto de corte encontrado foi igual a 0,500. Na Figura 5.11 estão representadas as frequências de seleção de cada variável.

O fator de alisamento encontrado na fase de treinamento e aplicado nas redes PNN foi igual a 0,222.

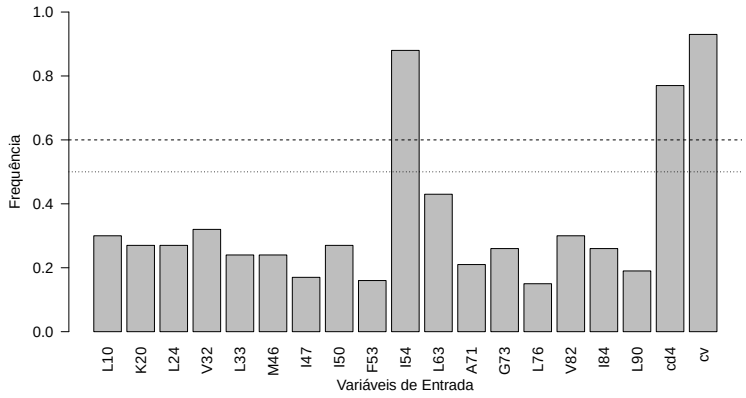

Figura 5.11: Frequência das variáveis nas redes neurais probabilísticas para o Lopinavir utilizando a codificação de Eisenberg e o conjunto "Lista da IAS".

Na Tabela 5.6, pode-se observar o desempenho médio dos modelos para o pontos de corte iguais a 0,500. Em média, os modelos apresentaram AUC igual a 0,60, acurácia igual a 0,64, sensibilidade de 0,63, especificidade de 0,64 e índice *Kappa* igual a 0,16, indicando uma concordância baixa.

Na Figura 5.12 estão representadas as curvas ROC dos modelos de redes PNN.

Tabela 5.6: Desempenho médio das redes neurais probabilísticas para o Lopinavir utilizando a codificação de Eisenberg e o conjunto "Lista da IAS" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,500	*PC: 0,500*	*PC: 0,500*	*PC: 0,500*
AUC	0,60±0,02	0,60±0,02	0,60±0,02	0,60±0,02
Acurácia	0,64±0,04	0,64±0,04	0,64±0,04	0,64±0,04
Sensibilidade	0,63±0,06	0,63±0,06	0,63±0,06	0,63±0,06
Especificidade	0,64±0,04	0,64±0,04	0,64±0,04	0,64±0,04
Índice Kappa	0,16±0,05	0,16±0,05	0,16±0,05	0,16±0,05

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

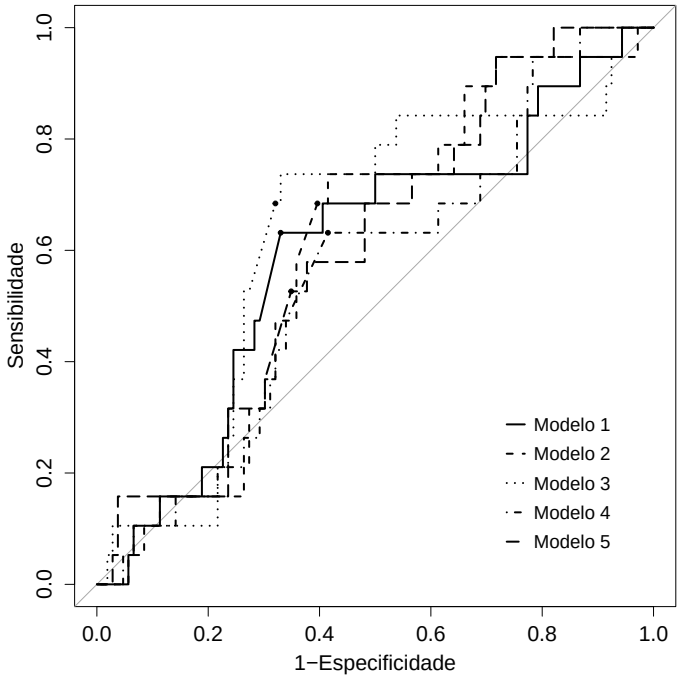

Figura 5.12: Curvas ROC das redes neurais probabilísticas para o Lopinavir utilizando a codificação de Eisenberg e o conjunto "Lista da IAS" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.1.2.3 Todas as Posições e Codificação Binária

As variáveis selecionadas neste grupo foram as posições I54, K70 e CV para o critério de seleção $\geq 50\%$. Para o critério $\geq 60\%$, apenas a posição I54 e a variável CV foram eleitas. Os melhores pontos de corte encontrados para esses modelos foram iguais a 0,480 e 0,001 para os critérios $\geq 50\%$ e $\geq 60\%$, respectivamente. Na Figura 5.13 está representado o gráfico das frequências de seleção de cada variável inicialmente incorporada às redes.

O fator de alisamento encontrado na fase de treinamento e aplicado nas redes PNN foi igual a 0,518.

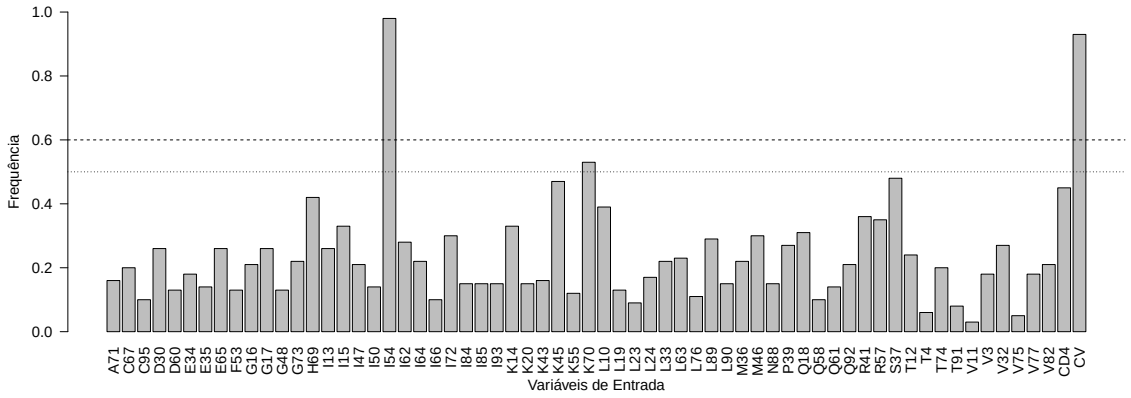

Figura 5.13: Frequência das variáveis nas redes neurais probabilísticas para o Lopinavir utilizando a codificação binária e o conjunto "Todas as Posições".

Em média, os modelos obtidos para os critérios $\geq 50\%$ e $\geq 60\%$ apresentaram AUCs iguais a 0,76 e 0,75, respectivamente. Para o ponto de corte de 0,480, a acurácia média foi igual a 0,79, sensibilidade e especificidade médias de 0,58 e 0,82, respectivamente, e índice *Kappa* médio igual a 0,33 (concordância razoável). Para o ponto de corte de 0,001, a acurácia média foi igual a 0,66, sensibilidade e especificidade médias de 0,73 e 0,65, respectivamente, e índice *Kappa* médio igual a 0,22 (concordância razoável). Na Tabela 5.7 está representado o desempenho médio das redes PNN em termos de AUC, acurácia, sensibilidade e especificidade para os dois critérios de seleção e seus respectivos pontos de corte. Na Figura 5.14 estão representadas as curvas ROC dos modelos obtidos pelas redes PNN para os dois critérios de seleção de variáveis.

Tabela 5.7: Desempenho médio dos modelos de PNN para o Lopinavir utilizando a codificação binária e o conjunto "Todas as Posições" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,480	*PC: 0,500*	*PC: 0,001*	*PC: 0,500*
AUC	0,76±0,01	0,76±0,01	0,75±0,01	0,75±0,01
Acurácia	0,79±0,04	0,79±0,04	0,66±0,03	0,80±0,04
Sensibilidade	0,58±0,04	0,57±0,04	0,73±0,09	0,54±0,02
Especificidade	0,82±0,05	0,82±0,05	0,65±0,06	0,85±0,05
Índice Kappa	0,33±0,05	0,32±0,04	0,22±0,02	0,33±0,05

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

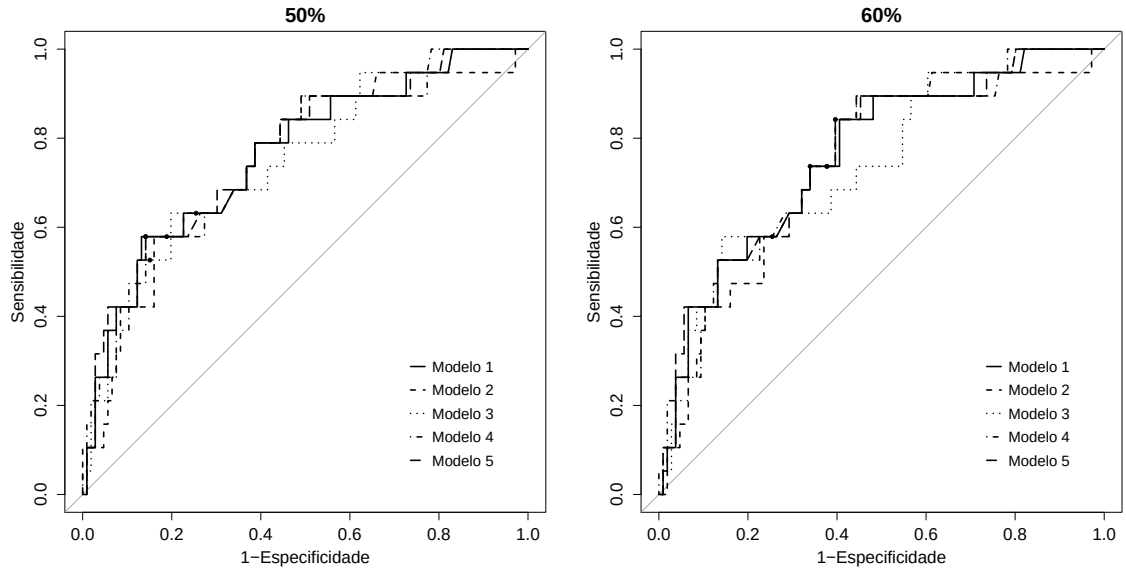

Figura 5.14: Curvas ROC das redes neurais probabilísticas para o Lopinavir utilizando a codificação binária e o conjunto "Todas as Posições" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.1.2.4 Todas as Posições e Codificação de Eisenberg

Neste grupo, as variáveis selecionadas quando assumido o critério de seleção $\geq 50\%$ foram as posições I54, K14, K70, S37 e a variável clínica CV. O melhor ponto de corte encontrado foi igual a 0,429. Quando o critério foi $\geq 60\%$, as variáveis selecionadas foram apenas as posições I54, K14 e K70 e a variável clínica CV. O melhor ponto de corte encontrado foi igual a 0,462. As frequências de seleção de cada variável na fase de treinamento estão representadas na Figura 5.15.

O fator de alisamento encontrado na fase de treinamento e aplicado nas redes PNN foi igual a 0,555.

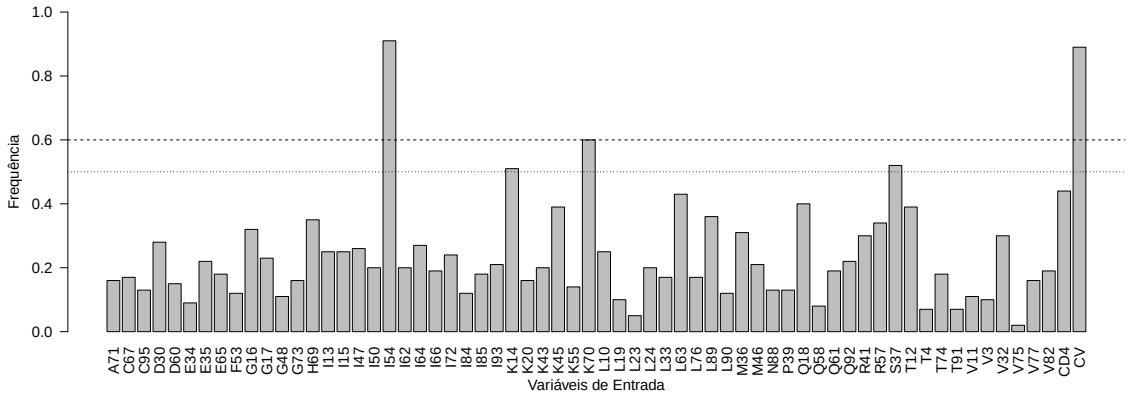

Figura 5.15: Frequência das variáveis nas redes neurais probabilísticas para o Lopinavir utilizando a codificação de Eisenberg e o conjunto "Todas as Posições".

Na Tabela 5.8 está representado os valores médios do desempenho das redes PNN em termos de AUC, acurácia, sensibilidade, especificidade e índice *Kappa* de acordo com os pontos de corte de 0,429 e 0,462 e com os critérios de seleção ($\geq$ 50% e $\geq$ 60%). Em média, os modelos apresentaram AUCs iguais a 0,66 e 0,74 para os critérios $\geq$ 50% e $\geq$ 60% . Para o ponto de corte de 0,429, em média, a acurácia foi igual a 0,67, a sensibilidade igual 0,62, a especificidade igual a 0,68 e o índice *Kappa* foi igual a 0,19 indicando uma concordância baixa. Para o ponto de corte de 0,462, os valores médios foram: acurácia igual a 0,78, sensibilidade igual a 0,59, especificidade de 0,82 e índice *Kappa* igual a 0,33, indicando uma concordância razoável.

Tabela 5.8: Desempenho médio das redes neurais probabilísticas para o Lopinavir utilizando a codificação de Eisenberg e o conjunto "Todas as Posições" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq$ 50%		*Critério* $\geq$ 60%	
	PC: 0,429	*PC: 0,500*	*PC: 0,462*	*PC: 0,500*
AUC	0,66±0,01	0,66±0,01	0,74±0,01	0,74±0,01
Acurácia	0,67±0,05	0,70±0,05	0,78±0,02	0,78±0,02
Sensibilidade	0,62±0,04	0,54±0,06	0,59±0,02	0,57±0,04
Especificidade	0,68±0,06	0,73±0,06	0,82±0,03	0,82±0,03
Índice Kappa	0,19±0,04	0,19±0,06	0,33±0,03	0,32±0,03

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

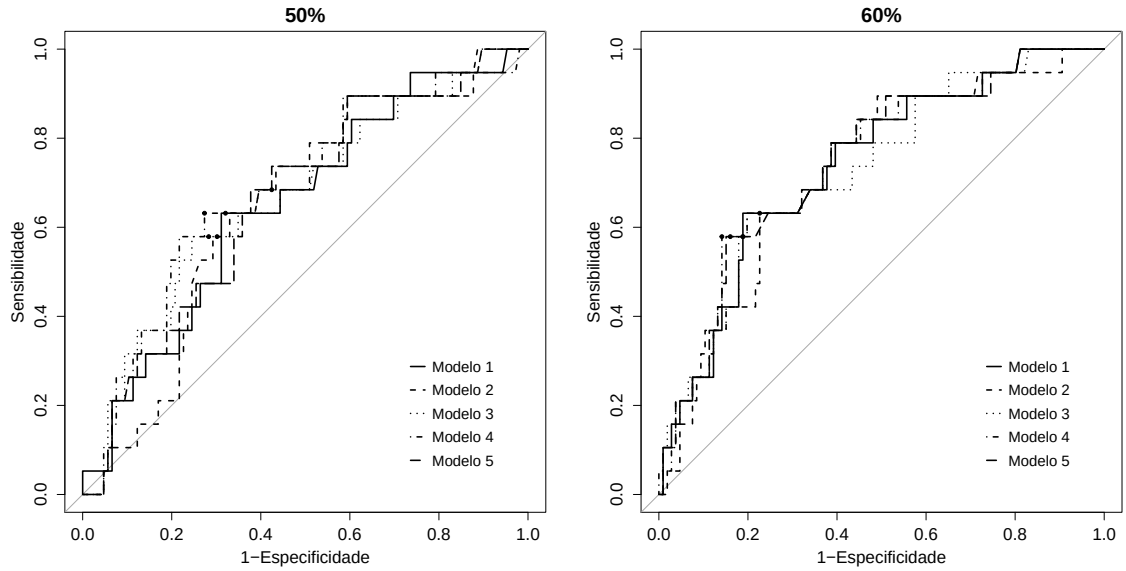

Figura 5.16: Curvas ROC das redes neurais probabilísticas para o Lopinavir utilizando a codificação de Eisenberg e o conjunto "Todas as Posições" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.2 Nelfinavir

5.2.1 Modelos de Regressão Logística

5.2.1.1 Lista da IAS e Codificação Binária

As variáveis selecionadas na RL para o NFV quando assumido o critério de seleção $\geq 50\%$ foram as posições D30, V82, I84, N88 e L90. O melhor ponto de corte encontrado foi igual a 0,428. Quando o critério foi $\geq 60\%$, as variáveis selecionadas foram as mesmas, com exceção da posição N88, e o melhor ponto de corte foi igual a 0,558. Na Figura 5.17 está representado o gráfico das frequências de seleção de cada variável inicialmente incorporada aos modelos.

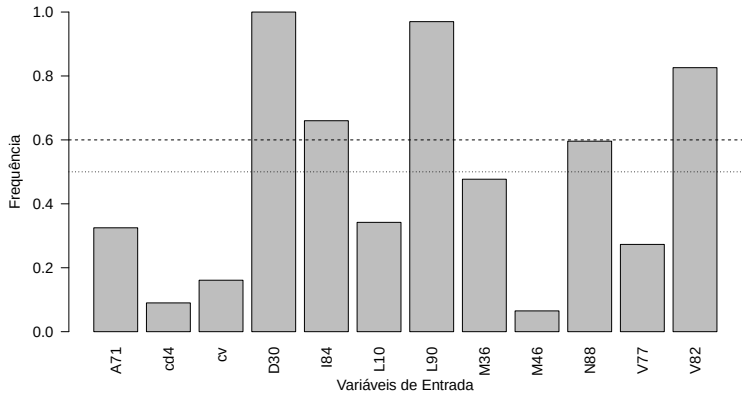

Figura 5.17: Frequência das variáveis nos modelos logísticos para o Nelfinavir utilizando a codificação binária e o conjunto "Lista da IAS".

Os modelos obtidos pela RL quando utilizado o critério $\geq 50\%$ apresentaram as seguintes equações e AICs:

Modelo 1: $Resistência = -0,73 + 2,50D30 - 0,72V82 - 0,35I84 + 0,59N88 + 1,00L90$, com um AIC igual a 232,8.

Modelo 2: $Resistência = -0,96 + 2,34D30 - 1,21V82 - 0,93I84 + 1,61N88 + 2,07L90$, com um AIC de 209,6.

Modelo 3: $Resistência = -0,79 + 2,29D30 - 0,71V82 - 1,16I84 + 1,04N88 + 1,37L90$, e AIC igual a 226,3.

Modelo 4: $Resistência = -1,00 + 2,43D30 - 0,83V82 - 0,89I84 + 1,24N88 + 1,86L90$, e AIC igual a 218,4.

Quando assumido o critério de seleção de variáveis $\geq 60\%$, os modelos obtidos foram:

Modelo 1: $Resistência = -0,69 + 2,87D30 - 0,72V82 - 0,34I84 + 1,00L90$, com um AIC igual a 231,7.

Modelo 2: $Resistência = -0,85 + 3,07D30 - 1,16V82 - 1,02I84 + 2,03L90$, com um AIC de 214,1.

Modelo 3: $Resistência = -0,73 + 2,92D30 - 0,66V82 - 1,11I84 + 1,31L90$, e AIC igual a 227.

Modelo 4: $Resistência = -0,89 + 3,06D30 - 0,79V82 - 0,85I84 + 1,76L90$, e AIC igual a 220,6.

Na Tabela 5.9 pode ser observado o desempenho dos modelos segundo seus critérios de seleção de variáveis e seus respectivos pontos de corte. Em média, os modelos logísticos apresentaram AUCs iguais a 0,76 e 0,74 para os critérios $\geq 50\%$ e $\geq 60\%$, respectivamente. Quando assumido um ponto de corte de 0,428, a acurácia média foi igual a 0,75, sensibilidade e especificidade médias foram iguais a 0,67 e 0,78, respectivamente, e índice *Kappa* médio igual a 0,40, indicando uma concordância

moderada. Para o ponto de corte de 0,558, a acurácia média foi igual a 0,82, sensibilidade e especificidade médias foram iguais a 0,60 e 0,89, respectivamente, e índice *Kappa* médio igual a 0,51, também indicando uma concordância moderada.

Na Figura 5.18 estão representadas as curvas ROC dos modelos logísticos para o NFV para cada critério de seleção de variáveis.

Tabela 5.9: Desempenho médio dos modelos logísticos para o Nelfinavir utilizando a codificação binária e o conjunto "Lista da IAS" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,428	*PC: 0,500*	*PC: 0,558*	*PC: 0,500*
AUC	0,76±0,00	0,76±0,00	0,74±0,00	0,74±0,00
Acurácia	0,75±0,03	0,80±0,03	0,82±0,00	0,77±0,06
Sensibilidade	0,67±0,00	0,65±0,03	0,60±0,00	0,60±0,00
Especificidade	0,78±0,04	0,84±0,04	0,89±0,00	0,83±0,08
Índice Kappa	0,40±0,05	0,47±0,05	0,51±0,00	0,41±0,11

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

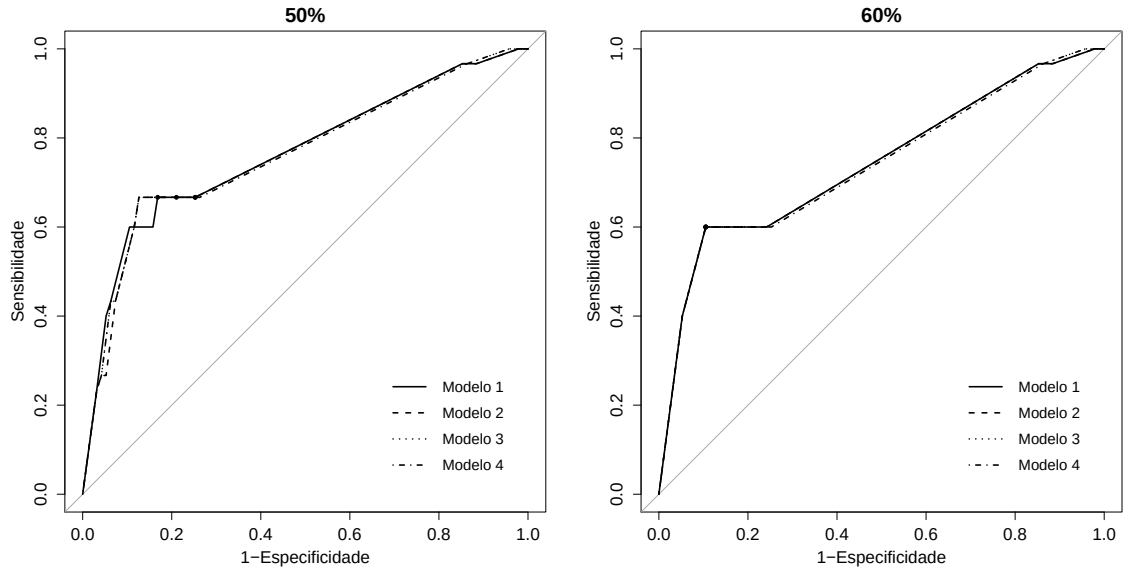

Figura 5.18: Curvas ROC dos modelos logísticos para o Nelfinavir utilizando a codificação binária e o conjunto "Lista da IAS" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.2.1.2 Lista da IAS e Codificação de Eisenberg

Quando aplicado o critério de seleção $\geq 50\%$, as variáveis selecionadas foram as posições D30, I84, N88 e L90. O melhor ponto de corte definido foi igual a 0,469. Quando o critério foi $\geq 60\%$, as posições D30, I84 e L90 foram selecionadas, com exceção da N88. O melhor ponto de corte, neste caso, foi igual a 0,484. Na Figura 5.19 estão representadas as frequências de seleção de cada variável.

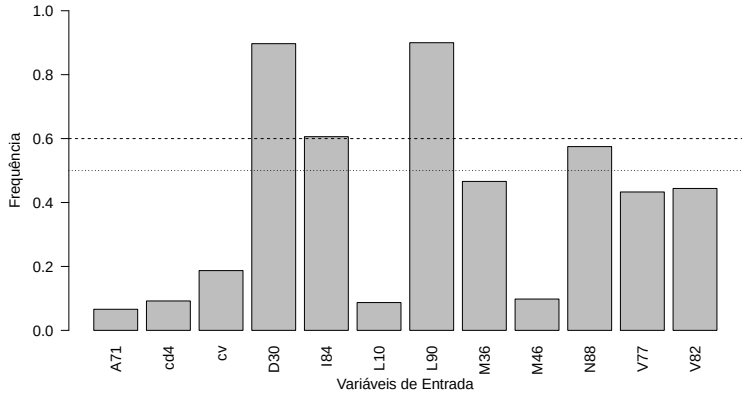

Figura 5.19: Frequência das variáveis nos modelos logísticos para o Nelfinavir utilizando a codificação de Eisenberg e o conjunto "Lista da IAS".

Os quatro modelos obtidos pela RL para o critério $\geq 50\%$ apresentaram as seguintes equações e AICs:

Modelo 1: $Resistência = 0,15 + 1,31D30 + 0,07I84 + 0,09N88 - 0,42L90$, com um AIC igual a 234,5.

Modelo 2: $Resistência = 0,05 + 0,83D30 + 0,30I84 - 0,09N88 - 0,61L90$, com um AIC de 264,8.

Modelo 3: $Resistência = 0,15 + 1,55D30 + 0,39I84 + 0,48N88 - 0,64L90$, e AIC igual a 222,9.

Modelo 4: $Resistência = 0,10 + 1,37D30 + 0,25I84 + 0,30N88 - 0,71L90$, e AIC igual a 237,8.

Os modelos obtidos pela RL, quando selecionadas as posições presentes em pelo menos 60% dos modelos de treinamento, apresentaram as seguintes equações:

Modelo 1: $Resistência = 0,15 + 1,27D30 + 0,08I84 - 0,44L90$, com um AIC igual a 232,8.

Modelo 2: $Resistência = 0,06 + 0,93D30 + 0,30I84 - 0,62L90$, com um AIC de 263,1.

Modelo 3: $Resistência = 0,14 + 1,27D30 + 0,36I84 - 0,56L90$, e AIC igual a 227,3.

Modelo 4: $Resistência = 0,10 + 1,18D30 + 0,22I84 - 0,66L90$, e AIC igual a 238,6.

Na Tabela 5.10 está representado o desempenho médio dos modelos logísticos em termos de AUC, acurácia, sensibilidade e especificidade para os dois critérios de seleção e seus respectivos pontos de corte. Em média, os modelos logísticos para os critérios $\geq 50\%$ e $\geq 60\%$ apresentaram AUCs iguais a 0,73 e 0,71, respectivamente. Para o ponto de corte de 0,469, a acurácia média foi igual a 0,76, sensibilidade e especificidade média de 0,63 e 0,80, respectivamente, e índice *Kappa* médio igual a 0,40 (concordância moderada). Para o ponto de corte igual a 0,484, a acurácia média foi igual a 0,75, sensibilidade e especificidade média de 0,60 e 0,80, respectivamente, e índice *Kappa* médio igual a 0,37 (concordância razoável).

Na Figura 5.20 estão representadas as curvas ROC dos modelos logísticos para o NFV deste grupo.

Tabela 5.10: Desempenho médio dos modelos logísticos para o Nelfinavir utilizando a codificação de Eisenberg e o conjunto "Lista da IAS" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,469	*PC: 0,500*	*PC: 0,484*	*PC: 0,500*
AUC	0,73±0,05	0,73±0,05	0,71±0,02	0,71±0,02
Acurácia	0,76±0,01	0,76±0,01	0,75±0,00	0,75±0,00
Sensibilidade	0,63±0,04	0,63±0,04	0,60±0,00	0,60±0,00
Especificidade	0,80±0,01	0,81±0,01	0,80±0,00	0,80±0,00
Índice Kappa	0,40±0,03	0,40±0,03	0,37±0,00	0,37±0,00

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

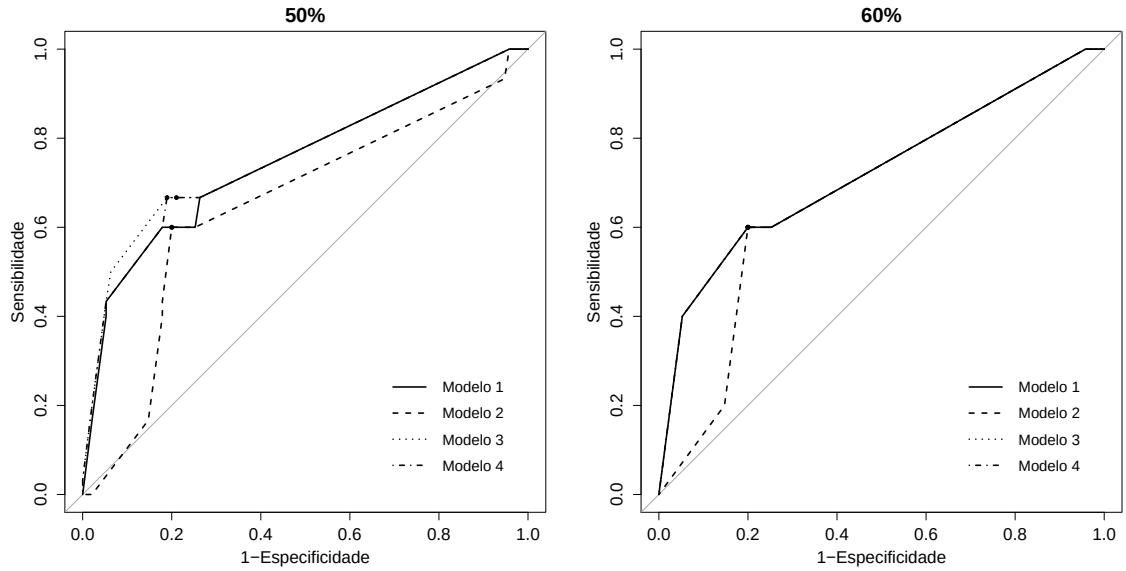

Figura 5.20: Curvas ROC dos modelos logísticos para o Nelfinavir utilizando a codificação de Eisenberg e o conjunto "Lista da IAS" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.2.1.3 Todas as Posições e Codificação Binária

Neste grupo, as variáveis selecionadas, quando assumido o critério de seleção $\geq 50\%$, foram as posições D30, T74, L89 e L90. O melhor ponto de corte encontrado foi igual a 0,454. Quando o critério foi $\geq 60\%$, as variáveis selecionadas foram apenas as posições D30 e L90. O melhor ponto de corte encontrado foi igual a 0,402. As frequências de seleção de cada variável na fase de treinamento estão representadas na Figura 5.19.

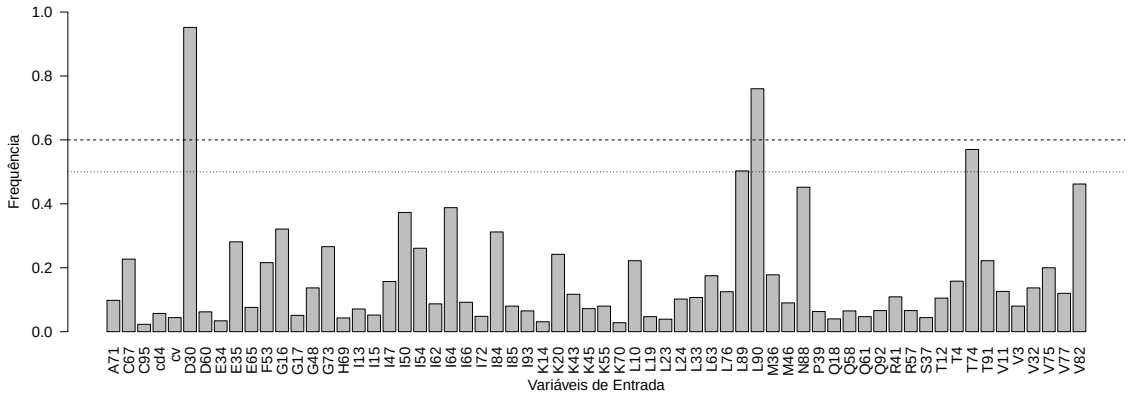

Figura 5.21: Frequência das variáveis nos modelos logísticos para o Nelfinavir utilizando a codificação binária e o conjunto "Todas as Posições".

Os quatro modelos obtidos pela RL para o critério $\geq 50\%$ foram:

Modelo 1: $Resistência = -0,80 + 3,23D30 + 1,68T74 + 0,17L89 + 0,30L90$, com um AIC igual a 222,7.

Modelo 2: $Resistência = -0,84 + 3,51D30 + 1,39T74 + 0,46L89 + 0,64L90$, com um AIC de 208,1.

Modelo 3: $Resistência = -0,76 + 3,23D30 + 1,15T74 + 0,05L89 + 0,37L90$, e AIC igual a 227,2.

Modelo 4: $Resistência = -0,87 + 3,47D30 + 2,00T74 + 0,56L89 + 0,56L90$, e AIC igual a 208,8.

Quando selecionadas apenas as variáveis presentes em pelo menos 60% dos modelos, as equações dos classificadores foram:

Modelo 1: $Resistência = -0,83 + 3,03D30 + 0,83L90$, com um AIC igual a 231.

Modelo 2: $Resistência = -1,04 + 3,24D30 + 1,50L90$, com um AIC de 218,8.

Modelo 3: $Resistência = -0,88 + 3,10D30 + 0,90L90$, e AIC igual a 229,9.

Modelo 4: $Resistência = -1,02 + 3,22D30 + 1,38L90$, e AIC igual a 221,2.

Na Tabela 5.11 pode ser observado o desempenho médio dos modelos. Em média, o critério$\geq 50\%$ e o $\geq 60\%$ apresentaram AUCs iguais a 0,70 e 0,71, respectivamente. Para o ponto de corte de 0,454, o desempenho médio apresentou uma acurácia igual a 0,74, sensibilidade de 0,53, especificidade igual a 0,81 e índice *Kappa* igual a 0,33, indicando uma concordância razoável. Para o ponto de corte de 0,402, o desempenho médio apresentou uma acurácia igual a 0,71, sensibilidade de 0,60, especificidade igual a 0,75 e índice *Kappa* igual a 0,31, uma concordância também razoável segundo a escala de Landis.

Na Figura 5.6 estão representadas as curvas ROC dos modelos logísticos para o NFV.

Tabela 5.11: Desempenho médio dos modelos logísticos para o Nelfinavir utilizando a codificação binária e o conjunto "Todas as Posições" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,454	*PC: 0,500*	*PC: 0,402*	*PC: 0,500*
AUC	0,70±0,00	0,70±0,00	0,71±0,00	0,71±0,00
Acurácia	0,74±0,05	0,76±0,05	0,71±0,00	0,71±0,00
Sensibilidade	0,53±0,08	0,50±0,07	0,60±0,00	0,60±0,00
Especificidade	0,81±0,09	0,85±0,08	0,75±0,00	0,75±0,00
Índice Kappa	0,33±0,05	0,35±0,05	0,31±0,00	0,31±0,00

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

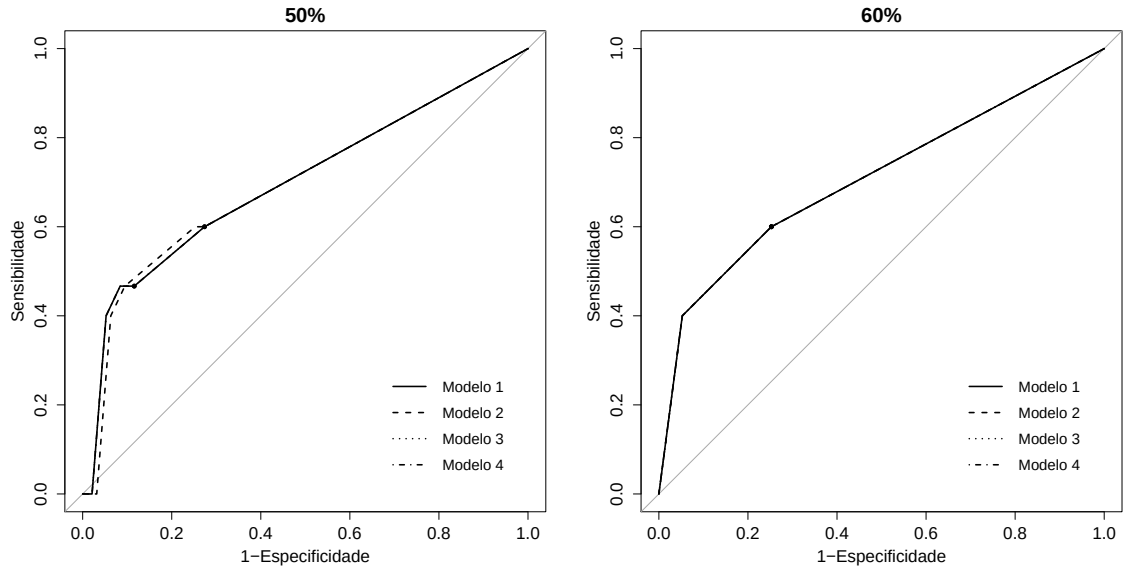

Figura 5.22: Curvas ROC dos modelos logísticos para o Nelfinavir utilizando a codificação binária e o conjunto "Todas as Posições" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.2.1.4 Todas as Posições e Codificação Segundo a Escala de Eisenberg

Quando aplicado o critério de seleção $\geq 50\%$, as variáveis selecionadas foram as posições D30, K20, I54 e L90. O melhor ponto de corte definido foi igual a 0,373. Quando o critério foi $\geq 60\%$, apenas as posições D30 e K20 foram selecionadas. O melhor ponto de corte encontrado foi igual a 0,417.

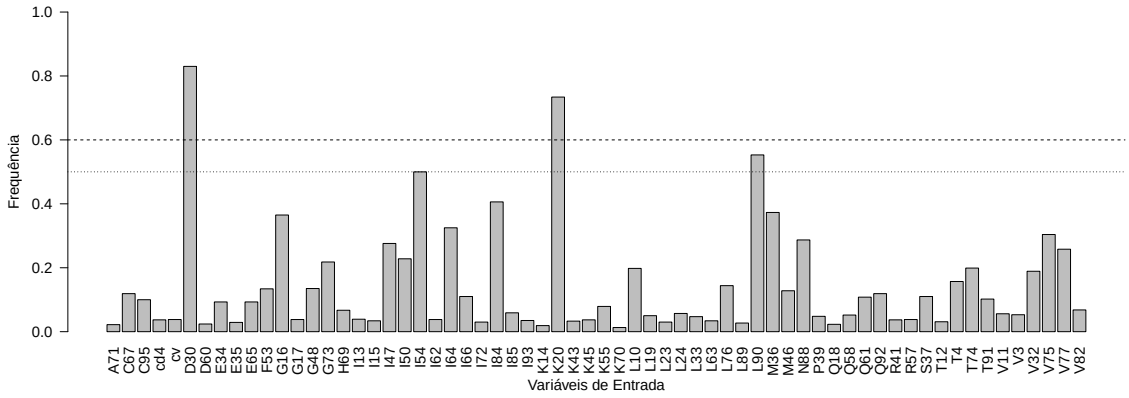

Figura 5.23: Frequência das variáveis nos modelos logísticos para o Nelfinavir utilizando a codificação de Eisenberg e o conjunto "Todas as Posições".

As equações dos modelos de RL e seus respectivos AICs para o critério de seleção $\geq 50\%$ foram:

Modelo 1: $Resistência = 0,13 + 0,33K20 + 1,28D30 + 0,33I54 - 0,37L90$, com um AIC igual a 227,4.

Modelo 2: $Resistência = 0,02 + 0,31K20 + 0,95D30 + 0,66I54 - 0,58L90$, com um AIC de 255,4.

Modelo 3: $Resistência = 0,12 + 0,03K20 + 1,23D30 + 0,69I54 - 0,68L90$, e AIC igual a 220,8.

Modelo 4: $Resistência = 0,11 + 0,39K20 + 1,22D30 + 0,25I54 - 0,63L90$, e AIC igual a 233.

Para o critério de seleção $\geq 60\%$, as equações foram:

Modelo 1: $Resistência = 0,15 + 0,44K20 + 1,27D30$, com um AIC igual a 229,2.

Modelo 2: $Resistência = 0,06 + 0,39K20 + 0,91D30$, com um AIC de 269,5.

Modelo 3: $Resistência = 0,15 + 0,20K20 + 1,23D30$, e AIC igual a 235,9.

Modelo 4: $Resistência = 0,10 + 0,50K20 + 0,13D30$, e AIC igual a 241,9.

Na Tabela 5.12 está representado o desempenho médio dos modelos segundo os critérios de seleção e seus respectivos pontos de corte. Em média, os modelos logísticos para os critérios $\geq 50\%$ e $\geq 60\%$ apresentaram AUCs iguais a 0,77 e 0,82, respectivamente. Para o ponto de corte de 0,373, a acurácia média foi igual a 0,75, sensibilidade e especificidade médias de 0,65 e 0,79, respectivamente, e índice *Kappa* médio igual a 0,39 (concordância razoável). Para o ponto de corte de 0,417, a acurácia média foi igual a 0,82, sensibilidade e especificidade média de 0,73 e 0,84, respectivamente, e índice *Kappa* médio igual a 0,53 (concordância moderada).

Na Figura 5.24 estão representadas as curvas ROC dos modelos logísticos para o NFV.

Tabela 5.12: Desempenho médio dos modelos logísticos para o Nelfinavir utilizando a codificação de Eisenberg e o conjunto "Todas as Posições" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,373	*PC: 0,500*	*PC: 0,417*	*PC: 0,500*
AUC	0,77±0,03	0,77±0,03	0,82±0,00	0,82±0,00
Acurácia	0,75±0,03	0,79±0,03	0,82±0,00	0,82±0,00
Sensibilidade	0,65±0,03	0,62±0,02	0,73±0,00	0,63±0,16
Especificidade	0,79±0,04	0,84±0,03	0,84±0,00	0,88±0,05
Índice Kappa	0,39±0,04	0,45±0,05	0,53±0,00	0,50±0,06

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

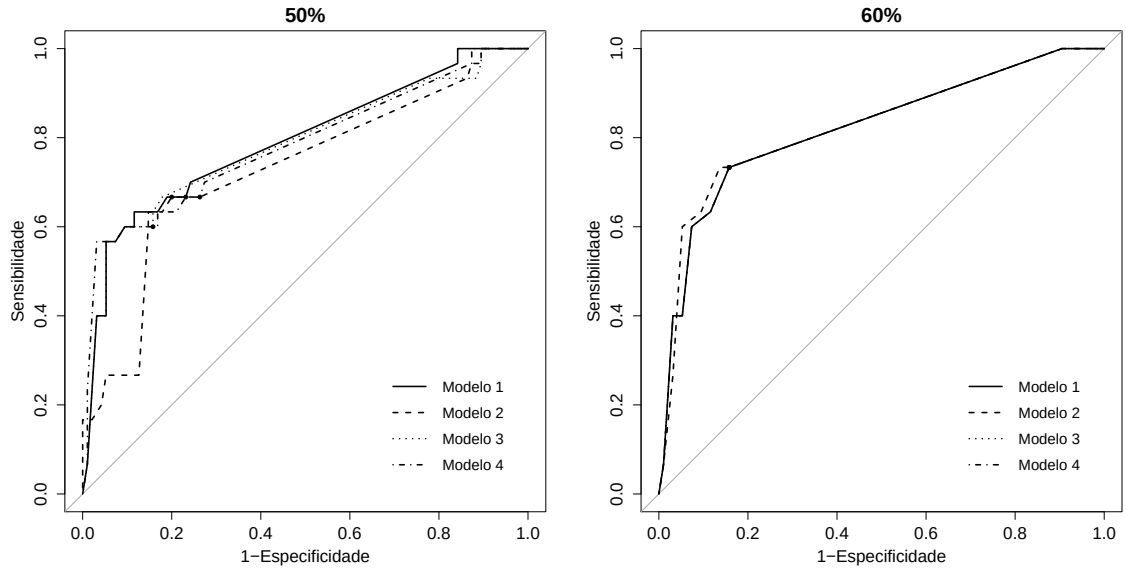

Figura 5.24: Curvas ROC dos modelos logísticos para o Nelfinavir utilizando a codificação de Eisenberg e o conjunto "Todas as Posições" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.2.2 Modelos de Redes Neurais Probabilísticas

5.2.2.1 Lista da IAS e Codificação Binária

Quando aplicado o critério de seleção $\geq 50\%$, as variáveis selecionadas foram todas as posições iniciais, com exceção da A71, assim como as variáveis CD4 e CV. O melhor ponto de corte definido foi igual a 0,500. Quando o critério foi $\geq 60\%$, apenas as posições L10, D30, M36, I84 e L90 foram selecionadas, além das variáveis

CD4 e CV. O melhor ponto de corte, neste caso, foi igual a 0,496. Na Figura 5.25 estão representadas as frequências de seleção de cada variável.

O fator de alisamento encontrado na fase de treinamento e aplicado nas redes PNN foi igual a 0,288.

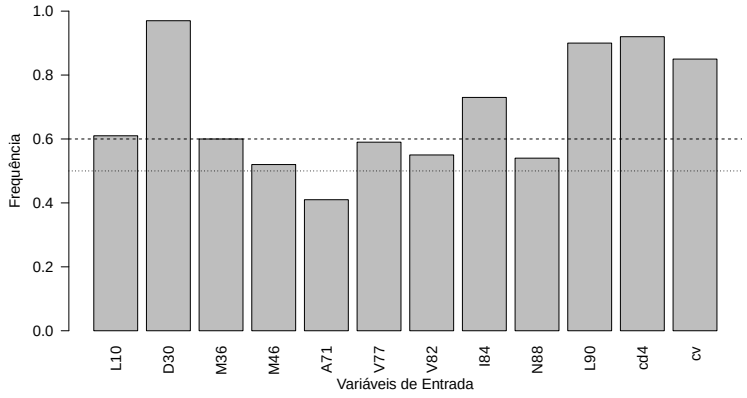

Figura 5.25: Frequência das variáveis nas redes neurais probabilísticas para o Nelfinavir utilizando a codificação binária e o conjunto "Lista da IAS".

Na Tabela 5.13 pode ser observado que o desempenho médio dos modelos. Em média, os modelos deste grupo apresentaram AUC igual a 0,64 e 0,67 para os critérios $\geq 50\%$ e $\geq 60\%$, respectivamente. O critério $\geq 50\%$ apresentou acurácia média igual a 0,60, sensibilidade média de 0,68, especificidade média igual a 0,57 e índice *Kappa médio* igual a 0,19 (concordância baixa). Para o critério $\geq 60\%$ e ponto de corte igual a 0,496, em média, a acurácia foi de 0,61, a sensibilidade e especificidade iguais a 0,71 e 0,58, respectivamente, e índice *Kappa* igual a 0,21, caracterizando uma concordância razoável. Na Figura 5.26 estão representadas as curvas ROC das redes PNN deste grupo para o NFV.

Tabela 5.13: Desempenho médio das redes neurais probabilísticas para o Nelfinavir utilizando a codificação binária e o conjunto "Lista da IAS" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,500	*PC: 0,500*	*PC: 0,496*	*PC: 0,500*
AUC	0,64±0,04	0,64±0,04	0,67±0,02	0,67±0,02
Acurácia	0,60±0,02	0,60±0,02	0,61±0,02	0,63±0,03
Sensibilidade	0,68±0,06	0,68±0,06	0,71±0,04	0,63±0,05
Especificidade	0,57±0,01	0,57±0,01	0,58±0,03	0,63±0,04
Índice Kappa	0,19±0,05	0,19±0,05	0,21±0,03	0,21±0,04

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

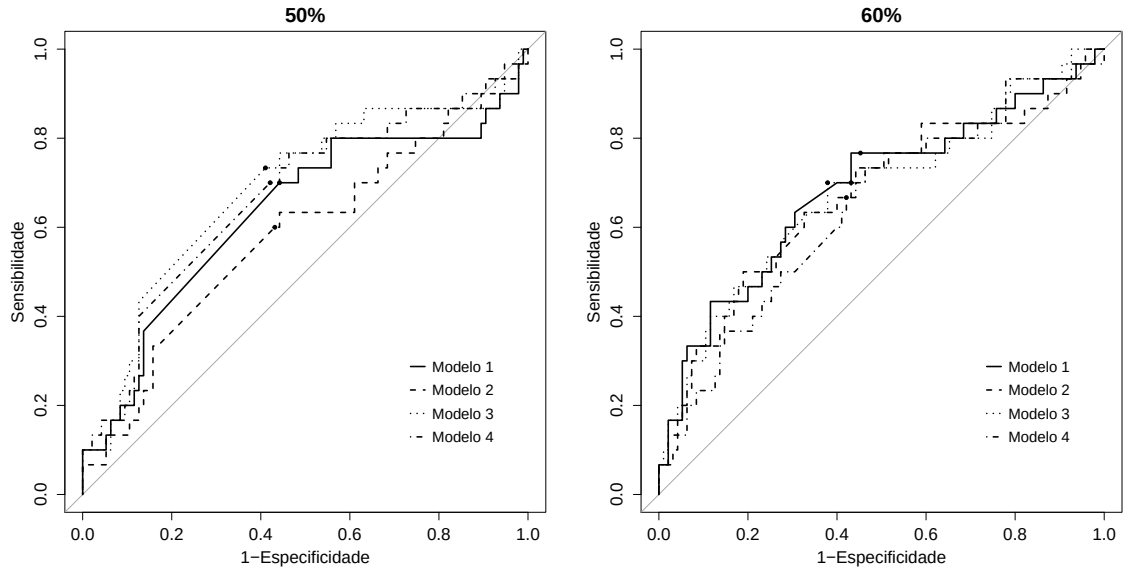

Figura 5.26: Curvas ROC das redes neurais probabilísticas para o Nelfinavir utilizando a codificação binária e o conjunto "Lista da IAS" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.2.2.2 Lista da IAS e Codificação de Eisenberg

Quando aplicada a codificação de Eisenberg, as variáveis eleitas pelo critério de seleção $\geq 50\%$ foram as posições D30, M46, A71, V77, I84, L90 e as variáveis CD4 e CV. O melhor ponto de corte encontrado foi igual a 0,500. Quando o critério foi $\geq 60\%$, apenas as posições D30, I84 e L90 foram selecionadas, assim como as variáveis CD4 e CV. O melhor ponto de corte, neste caso, foi igual a 0,579. Na Figura 5.27 estão representadas as frequências de seleção de cada variável.

O fator de alisamento encontrado na fase de treinamento e aplicado nas redes PNN foi igual a 0,262.

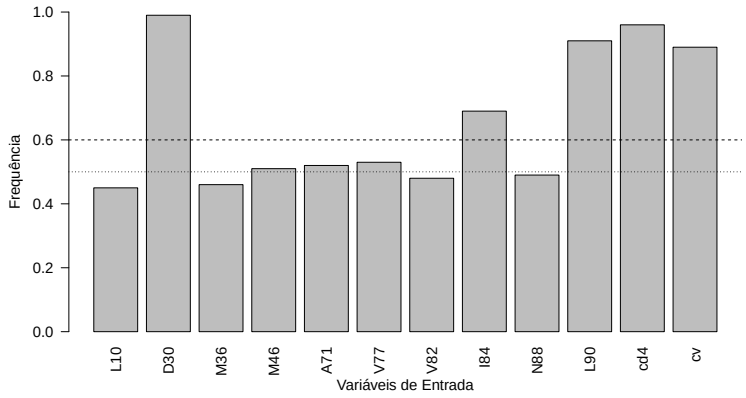

Figura 5.27: Frequência das variáveis nas redes neurais probabilísticas para o Nelfinavir utilizando a codificação de Eisenberg e o conjunto "Lista da IAS".

A Tabela 5.14 apresenta o desempenho médio dos modelos obtidos para este grupo. Os critérios de seleção de variáveis $\geq 50\%$ e $\geq 60\%$ apresentaram AUCs iguais a 0,61 e 0,71, respectivamente. Para o ponto de corte de 0,500 e critério $\geq 50\%$, o desempenho médio apresentou uma acurácia igual a 0,61, sensibilidade de 0,62, especificidade de 0,61 e índice *Kappa* igual a 0,18, indicando uma concordância baixa. Para o ponto de corte de 0,579, o desempenho médio apresentou uma acurácia igual a 0,78, sensibilidade de 0,48, especificidade de 0,88 e índice *Kappa* igual a 0,38, uma concordância razoável segundo a escala de Landis. A AUC média dos modelos foi igual a 0,61 e 0,71 para os critérios $\geq 50\%$ e $\geq 60\%$, respectivamente.

Na Figura 5.28 estão representadas as curvas ROC dos modelos para os dois critérios de seleção.

Tabela 5.14: Desempenho médio das redes neurais probabilísticas para o Nelfinavir utilizando a codificação de Eisenberg e o conjunto "Lista da IAS" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,500	*PC: 0,500*	*PC: 0,579*	*PC: 0,500*
AUC	0,61±0,02	0,61±0,02	0,71±0,01	0,71±0,01
Acurácia	0,61±0,01	0,61±0,01	0,78±0,01	0,70±0,01
Sensibilidade	0,62±0,02	0,62±0,02	0,48±0,02	0,60±0,04
Especificidade	0,61±0,01	0,61±0,01	0,88±0,01	0,74±0,01
Índice Kappa	0,18±0,02	0,18±0,02	0,38±0,03	0,29±0,03

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

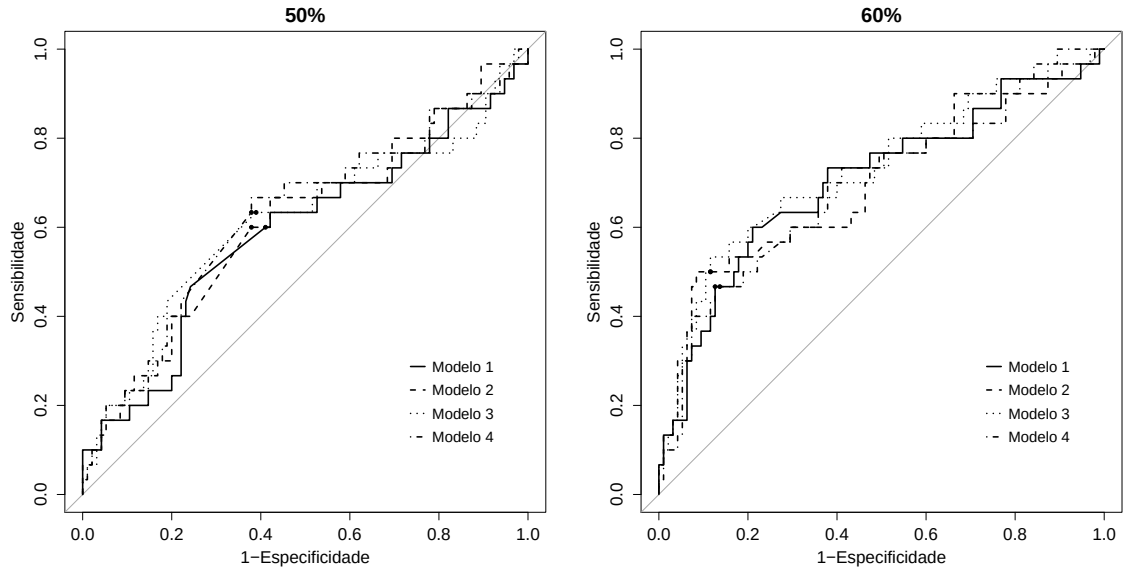

Figura 5.28: Curvas ROC das redes neurais probabilísticas para o Nelfinavir utilizando a codificação de Eisenberg e o conjunto "Lista da IAS" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.2.2.3 Todas as Posições e Codificação Binária

Quando aplicado o critério de seleção $\geq 50\%$, as variáveis selecionadas foram as posições V3, D30, E35, K43, I64, T74, L89 e L90, assim como as variáveis CD4 e CV. O melhor ponto de corte definido foi igual a 0,392. Quando o critério foi $\geq 60\%$, apenas as posições D30, T74, L89 e L90 foram selecionadas, e também as variáveis CD4 e CV. O melhor ponto de corte, neste caso, foi igual a 0,696. Na Figura 5.29 estão representadas as frequências de seleção de cada variável.

O fator de alisamento encontrado na fase de treinamento e aplicado nas redes PNN foi igual a 0,726.

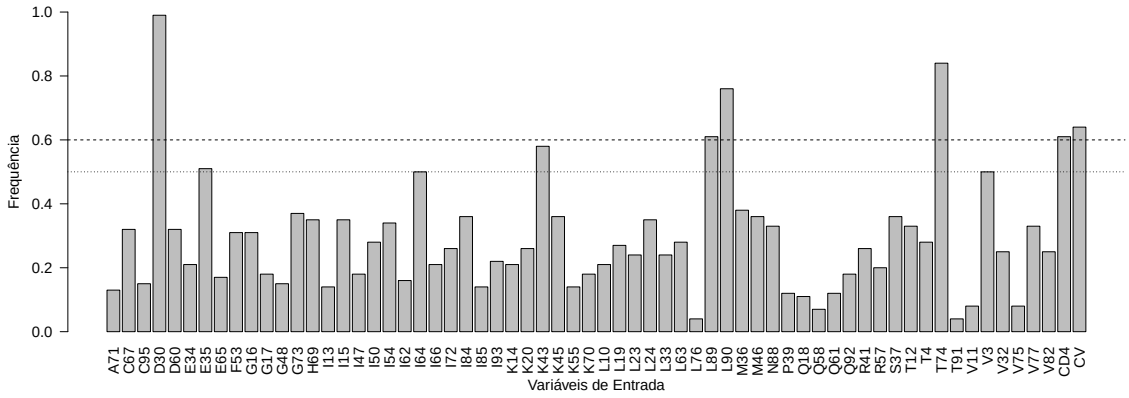

Figura 5.29: Frequência das variáveis nas redes neurais probabilísticas para o Nelfinavir utilizando a codificação binária e o conjunto "Todas as Posições".

Na Tabela 5.15 pode ser observado o desempenho médio dos modelos. Em média, tanto para o critério $\geq 50\%$ quanto para o $\geq 60\%$, os modelos apresentaram AUC igual a 0,72. Para o ponto de corte de 0,392, o desempenho médio apresentou uma acurácia igual a 0,69, sensibilidade de 0,68, especificidade de 0,69 e índice *Kappa* igual a 0,31, indicando uma concordância razoável. Para o ponto de corte de 0,696, o desempenho médio apresentou uma acurácia igual a 0,82, sensibilidade de 0,48, especificidade de 0,93 e índice *Kappa* igual a 0,45, uma concordância moderada segundo a escala de Landis.

Na Figura 5.30 estão representadas as curvas ROC dos modelos para o NFV.

Tabela 5.15: Desempenho médio das redes neurais probabilísticas para o Nelfinavir utilizando a codificação binária e o conjunto "Todas as Posições" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,392	*PC: 0,500*	*PC: 0,696*	*PC: 0,500*
AUC	0,72±0,03	0,72±0,03	0,72±0,01	0,72±0,01
Acurácia	0,69±0,03	0,73±0,03	0,82±0,02	0,76±0,03
Sensibilidade	0,68±0,06	0,53±0,02	0,48±0,02	0,54±0,02
Especificidade	0,69±0,03	0,80±0,01	0,93±0,03	0,83±0,04
Índice Kappa	0,31±0,06	0,31±0,10	0,45±0,04	0,36±0,04

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

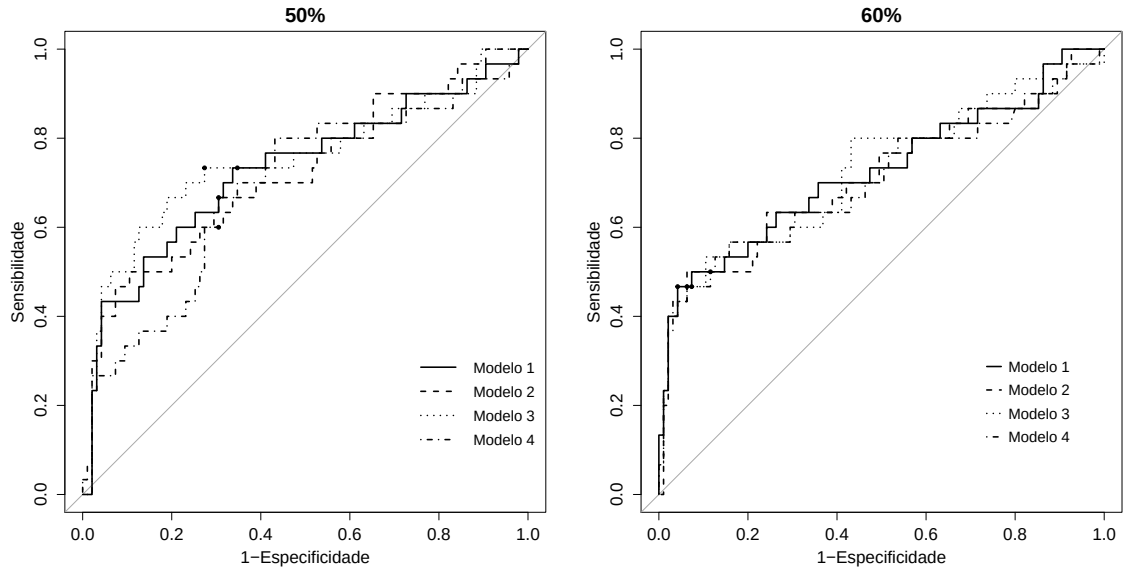

Figura 5.30: Curvas ROC das redes neurais probabilísticas para o Nelfinavir utilizando a codificação binária e o conjunto "Todas as Posições" para os dois critérios de seleção de variáveis.

Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.2.2.4 Todas as Posições e Codificação Segundo a Escala de Eisenberg

Quando aplicado o critério de seleção $\geq 50\%$, as variáveis selecionadas foram as posições K20, D30, E35, K43, T74, L89 e L90 e as variáveis CD4 e CV. O melhor ponto de corte definido foi igual a 0,264. Quando o critério foi $\geq 60\%$, as posições K20, D30, T74, L89 e L90 foram selecionadas. O melhor ponto de corte, neste caso, foi igual a 0,001. Na Figura 5.31 estão representadas as frequências de seleção de cada variável.

O fator de alisamento encontrado na fase de treinamento e aplicado nas redes PNN foi igual a 0,661.

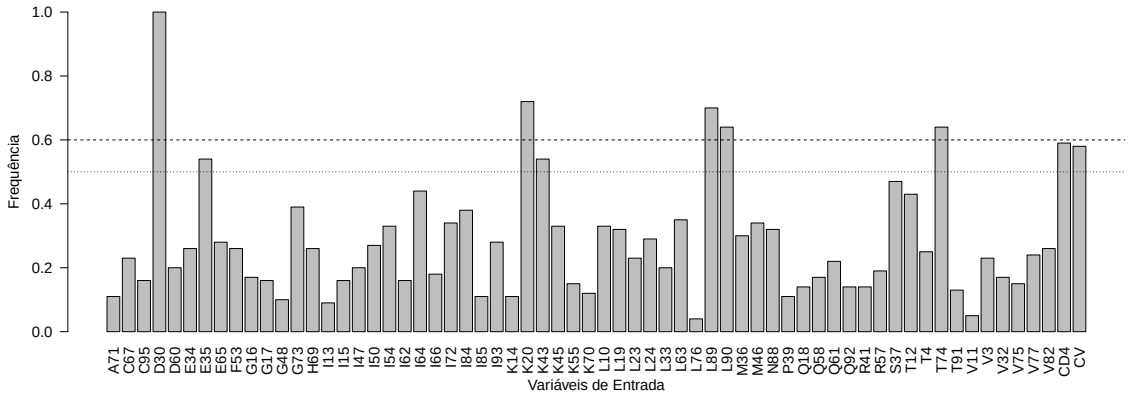

Figura 5.31: Frequência das variáveis nas redes neurais probabilísticas para o Nelfinavir utilizando a codificação de Eisenberg e o conjunto "Todas as Posições".

Na Tabela 5.16 está representado o desempenho médio dos modelos deste grupo em termos de AUC, acurácia, sensibilidade e especificidade para os dois critérios de seleção e seus respectivos pontos de corte. Em média, os modelos obtidos para os critérios $\geq 50\%$ e $\geq 60\%$ apresentaram AUCs iguais a 0,77 e 0,76, respectivamente. Para o ponto de corte de 0,264, a acurácia média foi igual a 0,70, sensibilidade e especificidade médias de 0,75 e 0,68, respectivamente, e índice *Kappa* médio igual a 0,34 (concordância razoável). Para o ponto de corte igual a 0,001, a acurácia média foi igual a 0,71, sensibilidade e especificidade médias de 0,70 e 0,71, respectivamente, e índice *Kappa* médio igual a 0,36 (concordância razoável).

Na Figura 5.32 estão representadas as curvas ROC dos modelos para os dois critérios de seleção.

Tabela 5.16: Desempenho médio das redes neurais probabilísticas para o Nelfinavir utilizando a codificação de Eisenberg e o conjunto "Todas as Posições" segundo os pontos de corte e os critérios de seleção de variáveis.

	Critério $\geq 50\%$		*Critério* $\geq 60\%$	
	PC: 0,264	*PC: 0,500*	*PC: 0,001*	*PC: 0,500*
AUC	0,77±0,01	0,77±0,01	0,76±0,01	0,76±0,01
Acurácia	0,70±0,01	0,75±0,03	0,71±0,01	0,74±0,02
Sensibilidade	0,75±0,03	0,58±0,03	0,70±0,07	0,57±0,06
Especificidade	0,68±0,01	0,80±0,03	0,71±0,00	0,80±0,03
Índice Kappa	0,34±0,02	0,35±0,06	0,36±0,04	0,34±0,04

PC: Ponto de Corte; AUC: Área sob a Curva ROC.

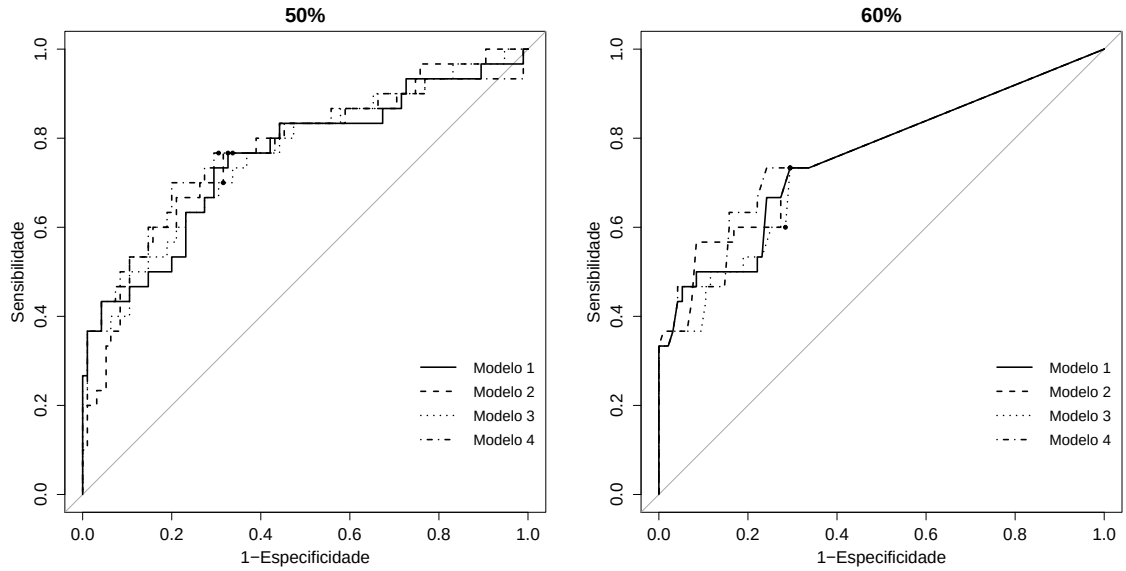

Figura 5.32: Curvas ROC das redes neurais probabilísticas para o Nelfinavir utilizando a codificação de Eisenberg e o conjunto "Todas as Posições" para os dois critérios de seleção de variáveis.
Os círculos representam os pontos de corte encontrados para cada critério de seleção de variáveis.

5.2.3 Tabelas - Resumo

Este item apresenta um resumo informando os resultados encontrados para cada ARV. Nas Tabelas 5.17 e 5.18 estão listadas as variáveis selecionadas por cada metodologia e critério de seleção final ($\geq 50\%$ e $\geq 60\%$). Nas Tabelas 5.19 e 5.20 estão os desempenhos médios dos modelos obtidos pelo estudo.

Tabela 5.17: Variáveis selecionadas pelos modelos segundo os critérios de seleção, as codificações dos aminoácidos e as técnicas de modelagem para o Lopinavir.

Variáveis Iniciais	*Codificação*	*Critério de Seleção*	*Regressão Logística*	*Rede Neural Probabilística*
Lista da IAS	*Binária*	$\geq 50\%$	I47, I54, L90 e CV	I54, L63, CD4 e CV
		$\geq 60\%$	I47, I54, L90 e CV	I54, CD4 e CV
Lista da IAS	*Eisenberg*	$\geq 50\%$	V32, I47, I54, L90 e CV	I54, CD4 e CV
		$\geq 60\%$	V32, I47, I54, L90 e CV	I54, CD4 e CV
Todas as Posições	*Binária*	$\geq 50\%$	I13, E35, R41, K45, I47, I54, I62, G73, L90, CD4 e CV	I54, K70 e CV
		$\geq 60\%$	I47, I54, L90, CD4 e CV	I54 e CV
Todas as Posições	*Eisenberg*	$\geq 50\%$	I13, K14, D30, V32, E35, R41, K45, I47, I54, R57, I62, I64, K70, G73, I84, I85, N88, L90, T91, CD4 e CV	K14, S37, I54, K70 e CV
		$\geq 60\%$	K45, I47, I54, G73, L90, T91 e CV	K14, I54, K70 e CV

Tabela 5.18: Variáveis selecionadas pelos modelos segundo os critérios de seleção, as codificações dos aminoácidos e as técnicas de modelagem para o Nelfinavir.

Variáveis Iniciais	*Codificação*	*Critério de Seleção*	*Regressão Logística*	*Rede Neural Probabilística*
Lista da IAS	*Binária*	$\geq 50\%$	D30, V82, I84, N88 e L90	L10, D30, M36, M46, V77, V82, I84, N88, L90, CD4 e CV
		$\geq 60\%$	D30, V82, I84 e L90	L10, D30, M36, I84, L90, CD4 e CV
Lista da IAS	*Eisenberg*	$\geq 50\%$	D30, I84, N88 e L90	D30, M46, A71, V77, I84, L90, CD4 e CV
		$\geq 60\%$	D30, I84 e L90	D30, I84, L90, CD4 e CV
Todas as Posições	*Binária*	$\geq 50\%$	D30, T74, L89 e L90	V3, D30, E35, K43, I64, T74, L89, L90, CD4 e CV
		$\geq 60\%$	D30 e L90	D30, T74, L89, L90, CD4 e CV
Todas as Posições	*Eisenberg*	$\geq 50\%$	K20, D30, I54 e L90	K20, D30, E35, K43, T74, L89, L90, CD4 e CV
		$\geq 60\%$	K20 e D30	K20, D30, T74, L89 e L90

Tabela 5.19: Comparação entre os desempenhos médios dos modelos para o Lopinavir.

			Regressão Logística					*Rede Neural Probabilística*				
Variáveis	*Codificação*	*Seleção*	*AUC*	*A*	*S*	*E*	*K*	*AUC*	*A*	*S*	*E*	*K*
Lista da IAS	*Binária*	$\geq 50\%$	0,79	0,70	0,71	0,70	0,26	0,64	0,71	0,56	0,74	0,21
		$\geq 60\%$	0,79	0,70	0,71	0,70	0,26	0,64	0,69	0,61	0,71	0,21
	Eisenberg	$\geq 50\%$	0,82	0,68	0,80	0,66	0,27	0,60	0,64	0,63	0,64	0,16
		$\geq 60\%$	0,82	0,68	0,80	0,66	0,27	0,60	0,64	0,63	0,64	0,16
Todas as Posições	*Binária*	$\geq 50\%$	0,80	0,70	0,75	0,69	0,28	0,76	0,79	0,58	0,82	0,33
		$\geq 60\%$	0,82	0,66	0,91	0,62	0,29	0,75	0,66	0,73	0,65	0,22
	Eisenberg	$\geq 50\%$	0,82	0,74	0,64	0,76	0,28	0,66	0,67	0,62	0,68	0,19
		$\geq 60\%$	0,78	0,67	0,73	0,66	0,23	0,74	0,78	0,59	0,82	0,33

AUC: Área sob a Curva ROC; A: Acurácia; S: Sensibilidade; E: Especificidade; K: Índice *Kappa*.

Tabela 5.20: Comparação entre os desempenhos médios dos modelos para o Nelfinavir.

			Regressão Logística					*Rede Neural Probabilística*				
Variáveis	*Codificação*	*Seleção*	*AUC*	*A*	*S*	*E*	*K*	*AUC*	*A*	*S*	*E*	*K*
Lista da IAS	*Binária*	$\geq 50\%$	0,76	0,75	0,67	0,78	0,40	0,64	0,60	0,68	0,57	0,19
		$\geq 60\%$	0,74	0,82	0,60	0,89	0,51	0,67	0,61	0,71	0,58	0,21
	Eisenberg	$\geq 50\%$	0,73	0,76	0,63	0,80	0,40	0,61	0,61	0,62	0,61	0,18
		$\geq 60\%$	0,71	0,75	0,60	0,80	0,37	0,71	0,78	0,48	0,88	0,38
Todas as Posições	*Binária*	$\geq 50\%$	0,70	0,74	0,53	0,81	0,33	0,72	0,69	0,68	0,69	0,31
		$\geq 60\%$	0,71	0,71	0,60	0,75	0,31	0,72	0,82	0,48	0,93	0,45
	Eisenberg	$\geq 50\%$	0,77	0,75	0,65	0,79	0,39	0,77	0,70	0,75	0,68	0,34
		$\geq 60\%$	0,82	0,82	0,73	0,84	0,53	0,76	0,71	0,70	0,71	0,36

AUC: Área sob a Curva ROC; A: Acurácia; S: Sensibilidade; E: Especificidade; K: Índice *Kappa*.

5.2.4 Algoritmos de Interpretação *Stanford HIVdb* e *Rega*

Para o LPV, o algoritmo *Stanford HIVdb* classificou os dados em cinco níveis de resistência: suscetível, resistência potencialmente baixa, resistência baixa, resistência intermediária e resistência elevada, enquanto que, para o algoritmo *Rega*, os dados foram classificados em três níveis de resistência: suscetível, resistência intermediária e resistência elevada.

Para avaliar o desempenho destes programas, os resultados do *Stanford HIVdb* foram classificados em três critérios: (1) em que as amostras suscetíveis constituíram a classe dos não resistentes e as restantes formaram a classe dos resistentes; (2) em

que os pacientes suscetíveis e os de resistência potencialmente baixa e baixa constituíram a classe dos não resistentes e os de resistência intermediária e elevada formaram a classe dos resistentes e (3) em que os pacientes suscetíveis, os de resistência potencialmente baixa e baixa e os de resistência intermediária constituíram a classe dos não resistentes e os de resistência elevada formaram a classe dos resistentes. Em relação aos resultados do algoritmo *Rega*, estes foram classificados em dois critérios de maneira diferente: (1) em que as amostras suscetíveis constituíram a classe dos não resistentes e as de resistência intermediária e elevada formaram a classe dos resistentes e (2) em que os pacientes suscetíveis e os de resistência intermediária constituíram a classe dos não resistentes e apenas os de resistência elevada formaram a classe dos resistentes.

O desempenho dos algoritmos segundo os seus critérios de classificação para o LPV está representado na Tabela 5.21. O algoritmo *Stanford HIVdb* apresentou melhor desempenho, em termos de sensibilidade, quando assumido o critério 1, com AUC igual a 0,63, acurácia igual a 0,62 e sensibilidade e especificidade iguais a 0,63 e 0,62, respectivamente. O índice *Kappa* foi igual a 0,15, indicando uma concordância baixa. O algoritmo *Rega* também apresentou melhor desempenho quando utilizado o critério 1, com AUC igual a 0,63, acurácia igual a 0,70, sensibilidade e especificidade iguais a 0,53 e 0,74, respectivamente, e índice *Kappa* igual a 0,19, indicando também uma concordância baixa. O desempenho dos algoritmos para o LPV segundo os dois critérios está representado na Tabela 5.21.

Tabela 5.21: Desempenho dos algoritmos *Stanford HIVdb* e *Rega* para o Lopinavir.

	Stanford HIVdb			*Rega*	
	Critério 1	*Critério 2*	*Critério 3*	*Critério 1*	*Critério 2*
AUC	0,63	0,67	0,66	0,63	0,66
Acurácia	0,62	0,73	0,82	0,70	0,83
Sensibilidade	0,63	0,58	0,42	0,53	0,42
Especificidade	0,62	0,75	0,90	0,74	0,91
Kappa	0,15	0,24	0,32	0,19	0,33

Para o NFV, os algoritmos *Stanford HIVdb* e *Rega* classificaram os dados em três níveis de resistência: suscetível, resistência intermediária e resistência elevada.

Neste caso, os resultados foram classificados em dois critérios: (1) em que as amostras suscetíveis constituíram a classe dos não resistentes e, as de resistência intermediária e elevada formaram a classe dos resistentes e, (2) em que os pacientes suscetíveis e os de resistência intermediária constituíram a classe dos não resistentes e, apenas os de resistência elevada formaram a classe dos resistentes.

Para o NFV, o algoritmo *Stanford HIVdb* apresentou melhor desempenho, em

termos de sensibilidade, quando assumido o critério 1, com AUC igual a 0,65, acurácia igual a 0,62, sensibilidade e especificidade iguais a 0,70 e 0,60, respectivamente, e índice *Kappa* igual a 0,23 (concordância razoável). O algoritmo *Rega* também apresentou melhor desempenho, em termos de sensibilidade, quando utilizado o critério 1, com AUC igual a 0,58, acurácia igual a 0,61 e sensibilidade e especificidade iguais a 0,53 e 0,63, respectivamente. O índice *Kappa* foi igual a 0,13, indicando uma concordância baixa. O desempenho dos algoritmos para o NFV segundo os dois critérios está representado na Tabela 5.22.

Tabela 5.22: Desempenho dos algoritmos *Stanford HIVdb* e *Rega* para o Nelfinavir.

	Stanford HIVdb		*Rega*	
	Critério 1	*Critério 2*	*Critério 1*	*Critério 2*
AUC	0,65	0,69	0,58	0,49
Acurácia	0,62	0,70	0,61	0,62
Sensibilidade	0,70	0,67	0,53	0,23
Especificidade	0,60	0,71	0,63	0,75
Kappa	0,23	0,31	0,13	-0,02

Capítulo 6

Discussão

Neste estudo, foram desenvolvidos quatro grupos de classificadores de resistência aos inibidores da HIV-protease LPV e NFV baseando-se em duas metodologias (RL e PNN) e em dois tipos de codificações dos aminoácidos (binária e de Eisenberg). Para o desenvolvimento destes modelos foram utilizados dados da Rede Nacional de Genotipagem (RENAGENO) aplicados pela primeira vez em um estudo cujo enfoque foi o da modelagem preditiva.

Além da avaliação em relação à técnica de modelagem e à codificação dos aminoácidos, foram aplicados dois critérios de seleção de variáveis; o critério $\geq 50\%$, em que as variáveis presentes em pelo menos 50% dos modelos de treinamento foram selecionadas e o critério $\geq 60\%$, em que foram selecionadas apenas aquelas variáveis presentes em pelo menos 60% dos mesmos modelos.

A seleção das variáveis significativas para compor os modelos finais foi realizada por meio das técnicas de *bootstrap* e *stepwise* na RL e por meio do *bootstrap*, validação cruzada e SFS na PNN. Esses procedimentos foram aplicados com o objetivo de escolher as variáveis mais significantes em cada método, possibilitando uma melhor seleção, uma vez que essas metodologias geram modelos diferentes quando utilizados conjuntos de dados diferentes. Essa escolha das variáveis deve ser realizada cuidadosamente, visto que a inclusão de variáveis não relevantes ao problema em estudo poderá prejudicar o desempenho do modelo, assim como o erro de classificação.

Outra característica deste estudo foi a utilização de dados balanceados em relação ao desfecho (resistentes e não resistentes), cujo objetivo foi o de evitar o surgimento de uma maior tendência dos modelos em responder bem para as classes majoritárias em detrimento das minoritárias. Geralmente, a construção de modelos com dados desbalanceados apresenta o risco de se obter soluções com um bom desempenho apenas em áreas com maior abundância de observação nos dados de entrada, prejudicando características importantes do modelo, como a robustez e capacidade de generalização.

Quando avaliado os quatro grupos de modelos de RL desenvolvidos para o LPV em relação às variáveis selecionadas, todos os modelos da RL selecionaram as posições I47, I54, L90 e também a carga viral (CV). Na rede PNN, a concordância ocorreu apenas em relação à variável CV. A posição I54 apareceu em três dos quatro grupos.

Estudos mostram que as mutações I47A/V são do tipo não polimórfica comumente selecionadas pelo LPV [53–55], sendo a mutação I47A responsável por conferir alto nível de resistência ao LPV, apesar de ser menos comum [56, 57]. Geralmente, acompanhando esta mutação, está presente a mutação V32I, que atua em sinergia para reduzir a susceptibilidade ao LPV [54]. No grupo em que foi utilizada a codificação por Eisenberg e o conjunto "Lista da IAS", a posição V32 foi selecionada, representando a única variável eleita que diferiu entre os grupos da RL que utilizaram as posições definidas pelas IAS.

Em relação à posição I54, a mutação I54V, não polimórfica, é selecionada principalmente pelo LPV [53–55], contribuindo na redução de suscetibilidade a este ARV [58]. Outras mutações não polimórficas desta posição são a I54M/L, também selecionadas pelo LPV e as mutações I54A/T/S, que ocorrem quase que exclusivamente em pacientes que receberam múltiplos PIs, reduzindo a suscetibilidade ao LPV [58].

A mutação L90M, cuja posição também foi selecionada pela RL, tem sido observada em níveis elevados em pacientes com falha ao LPV [55].

A única posição relevante para a resistência ao LPV selecionada pela PNN foi a posição I54, presente em todos os grupos.

Para este estudo, foi definido como melhor modelo aquele que apresenta maior valor de sensibilidade, uma vez que esta característica atribui ao classificador a capacidade de detectar mais casos positivos (resistentes). Ao apresentar maior sensibilidade, o modelo classificará corretamente um resultado negativo, sendo útil no encaminhamento de pacientes com resultado positivo para uma avaliação com melhor desempenho, como um teste fenotípico, neste caso.

Quando comparados os desempenhos médios dos modelos logísticos dos grupos que utilizaram a "Lista da IAS", para o melhor ponto de corte, a codificação de Eisenberg apresentou uma sensibilidade média maior, igual a 0,80. A especificidade foi igual a 0,66, a AUC igual a 0,82, a acurácia igual a 0,68 e o índice *Kappa* igual a 0,27 .

Não ocorreu diferença na seleção de variáveis nos grupos "Lista da IAS e Codificação Binária" e "Lista da IAS e Codificação de Eisenberg" quando assumidos os dois critérios de seleção.

Para os grupos da RL que utilizaram "Todas as Posições", ocorreu diferença na seleção das variáveis quando variou-se o critério. Além das posições já mencionadas, duas mutações presentes na "Lista da IAS" foram selecionadas na RL e possuem

efeitos sobre o LPV: a G73 em ambos os grupos e a I84 no grupo que utilizou a codificação de Eisenberg. A mutação G73S, além de ser selecionada na ausência de outras mutações de resistência, apresenta uma forte correlação com outras mutações de resistência aos PIs (isto é, L10I, K20I, M46I, ou L90M), sugerindo que ela possa ser parte de um padrão que confere resistência ao LPV [53]. Em relação à posição I84, a mutação I84V tem sido associada a falhas terapêuticas, o que contribui para a resistência ao LPV [53, 56]

Para os grupos da RL que aplicaram "Todas as Posições", o grupo com maior sensibilidade foi aquele que utilizou a codificação binária, com critério de seleção $\geq 60\%$. A sensibilidade média foi igual a 0,91, especificidade igual a 0,62, AUC igual a 0,82, acurácia igual a 0,66 e índice *Kappa* igual a 0,29.

Em relação aos modelos obtidos pela rede PNN, quando utilizada a "Lista da IAS", a maior sensibilidade média foi encontrada na codificação de Eisenberg para ambos os critérios de seleção de variáveis, com valor igual a 0,63. A especificidade foi igual a 0,64, AUC igual a 0,60, acurácia igual a 0,64 e índice *Kappa* igual 0,16. Para os grupos que utilizaram "Todas as Posições", a melhor sensibilidade foi igual a 0,73 para a codificação binária e critério $\geq 60\%$, com especificidade igual a 0,65, AUC igual a 0,75, acurácia igual a 0,66 e índice *Kappa* igual 0,22.

Avaliando todos os grupos de modelos para o LPV, pode-se observar que as variáveis I47 e I54 foram as mais importantes para a predição de resistência a este ARV na RL, enquanto que, apenas a I54 se destacou nas redes PNN.

Já em relação ao NFV, quando avaliado os resultados obtidos pelos quatro grupos, apenas a posição D30 foi selecionada por todos e pelos dois critérios de seleção de variáveis, tanto para a RL quanto para a PNN. A mutação D30N é uma alteração não polimórfica altamente associada aos casos de resistência ao NFV [59]. Ela ocorre apenas em pacientes que receberam NFV e não confere resistência cruzada a outro PI [60]. Além disso, essa mutação reduz a suscetibilidade ao NFV em 5 a 20 vezes, sendo frequentemente seguida pelo desenvolvimento da mutação N88D, cuja combinação reduz a suscetibilidade ao NFV em cerca de 50 vezes [61].

A posição L90 também foi outra que apareceu em todos os grupos, porém, não foi selecionada pelo critério $\geq 60\%$ no grupo "Todas as Posições e Codificação de Eisenberg" da RL. A L90M é uma mutação bem característica de resistência ao NFV [61], sendo responsável por reduzir a suscetibilidade do vírus à droga [58].

A posição I84 foi selecionada nos dois grupos de modelos quando utilizada a "Lista da IAS", tanto pela RL quanto pela PNN. A I84V é uma mutação não polimórfica selecionada por muitos PIs, relacionada à resistência ao NFV [58].

A posição N88, apesar de aparecer em alguns modelos quando utilizado o critério $\geq 50\%$, não foi selecionada quando essa porcentagem foi alterada para 60%, tanto nos modelos logísticos quanto nas redes PNN. A mutação N88D é selecionada pelo

NFV e em combinação com a mutação D30N reduz a susceptibilidade a esse ARV.

Na RL, as posições K20, T74, V82 e L89 foram selecionadas uma única vez pelos modelos. Com exceção da K20, essas outras posições não apareceram nos modelos quando o critério de seleção foi $\geq 60\%$.

As mutações K20I/M/T/V são comumente selecionadas pelo NFV. A mutação K20R é uma mutação acessória altamente polimórfica que contribui tanto para a redução da susceptibilidade aos PIs quanto no aumento da replicação do vírus em que contêm mutações de resistência (Nijhuis, 1999).

Apesar das mutações da posição 82 (V82A/T/S/F) serem selecionadas principalmente pelo IDV e pelo LPV, elas também reduzem a suscetibilidade ao NFV [55, 58].

As mutações T74P/S são mutações acessórias responsáveis por reduzir a suscetibilidade do vírus aos ARVs. Nos subtipos B, a mutação T74S é selecionada pelo NFV [62].

A L89V é uma mutação acessória não polimórfica selecionada por vários PIs, entre eles o NFV, contribuindo na redução da suscetibilidade a essas drogas [61].

Nos modelos obtidos pela PNN, quando utilizada a "Lista da IAS", a maioria das posições foram selecionadas pelo critério $\geq 50\%$. Essas posições são consideradas importantes no processo de resistência ao NFV. Dentre as selecionadas e que não foram mencionadas anteriormente, as M46I/L são mutações primárias não polimórficas selecionadas pelo NFV.

Quando comparados os desempenhos médios dos modelos de RL dos grupos que utilizaram a "Lista da IAS", ao assumir o melhor ponto de corte, o conjunto que utilizou a codificação binária e o critério $\geq 50\%$ apresentou uma sensibilidade média maior, igual a 0,67, especificidade igual a 0,78, AUC igual a 0,76, acurácia igual a 0,75 e índice *Kappa* igual a 0,40. Esse aumento da sensibilidade pode estar relacionado à inclusão da posição N88 ao modelo, cuja mutação, em combinação com a D30N, reduz a susceptibilidade ao NFV, podendo ter ajudado na melhor identificação dos casos resistentes.

Para os grupos da RL que aplicaram "Todas as Posições", os modelos com maiores sensibilidades foram aqueles que utilizaram a codificação de Eisenberg com critério de seleção $\geq 60\%$. A sensibilidade média foi igual a 0,73, especificidade de 0,84, AUC de 0,82, acurácia de 0,82 e índice *Kappa* igual a 0,53.

Em relação aos modelos obtidos pela rede PNN, quando utilizada a "Lista da IAS", a maior sensibilidade média foi encontrada na codificação binária para o critério de seleção de variáveis $\geq 60\%$, com valor igual a 0,71. A especificidade foi igual a 0,58, a AUC igual a 0,67, a acurácia igual a 0,61 e o índice Kappa igual a 0,21. Para os grupos que utilizaram "Todas as Posições", a melhor sensibilidade foi igual a 0,75 para a codificação de Eisenberg e critério $\geq 50\%$, com especificidade

igual a 0,68, AUC igual a 0,77, acurácia de 0,70 e índice Kappa igual a 0,34.

Avaliando os modelos para o NFV, apenas a posição D30 foi selecionada em todos os modelos. Essa posição, por estar sempre presente, corrobora com os achados em outros estudos, indicando que esta posição seja uma grande definidora de resistência ao NFV.

Quando comparado o desempenho de todos os grupos, a RL apresentou melhor desempenho geral tanto para o LPV quanto para o NFV.

Em termos de acurácia, analisando as duas técnicas de modelagem, os valores médios variaram de 0,64 a 0,79 para o LPV e de 0,60 a 0,82 para o NFV. No estudo de Bonet *et al.* (2007) [10], a energia dos aminoácidos da protease foi utilizada como característica para predizer a resistência aos ARVs, utilizando SVM, redes neurais MLP e redes neurais BRNN. Para o LPV, as redes MLPs apresentaram uma acurácia média de 90,10%, enquanto que para o NFV, o valor foi igual a 86,88%, ambos superiores ao encontrado neste estudo. Os modelos obtidos por SVM apresentaram uma acurácia média de 86,51% e 79,30% para o LPV e NFV, respectivamente, com resultado semelhante ao do presente estudo para o NFV. As BRNNs apresentaram acurácia média igual a 94,40% e 93,57% para o LPV e NFV, respectivamente. Em termos de sensibilidade e especificidade, as redes BRNNs apresentaram sensibilidades iguais a 96,25% e 94,44% e especificidades iguais a 77,78% e 89,66% para o LPV e NFV, respectivamente. No presente estudo, as sensibilidades médias encontradas para o LPV, analisando as duas técnicas de modelagem, foram de 0,56 a 0,91, com especificidades médias variando de 0,62 a 0,82. Para o NFV, as sensibilidades médias encontradas foram de 0,48 a 0,75 com especificidades médias variando de 0,57 a 0,93.

Pasomsub *et al.* (2010) [12] também desenvolveu redes neurais artificiais para predizer o fenótipo a partir das sequências genotípicas. Para o LPV, a AUC foi igual a 0,92 (IC: 0,88 - 0,95) e para o NFV foi igual a 0,94 (IC: 0,92 - 0,97). No estudo de Vermeiren *et al.* (2007) [33] foi aplicada a técnica de regressão linear na predição de resistência aos ARVs. Para o LPV, a AUC encontrada foi igual a 0,77 e para o NFV, a AUC foi igual a 0,76. Neste presente estudo, em termos de AUC, os modelos apresentaram valores variando de 0,60 a 0,82 para o LPV e de 0,61 a 0,82 para o NFV.

Os resultados obtidos por Pasomsub *et al.* (2010) são bem superiores quando comparados ao dos modelos desse estudo. No entanto, a ausência de outros indicadores de desempenho, tais como a sensibilidade e especificidade, limita a interpretação do desempenho do modelo e a comparação com os resultados deste estudo. Ao selecionar um classificador, além da acurácia, é importante conhecer outros parâmetros de desempenho como a sensibilidade e a especificidade. Muitos trabalhos não apresentam esses indicadores de desempenho, reduzindo a interpretação de seus resultados.

No estudo de Rhee *et al.* (2006) [47], cinco métodos de aprendizagem (árvores

de decisão, redes neurais do tipo *feed-forward*, regressão por vetores-suporte, regressão por mínimos quadrados e regressão de ângulo mínimo) foram utilizados na construção de seus modelos, fazendo uso de um conjunto completo com 70 posições da HIV-protease e um conjunto de mutações selecionadas pela lista da IAS. Para o LPV, a acurácia foi igual a 0,76 para o conjunto completo de posições e 0,73 para o conjunto lista da IAS, ambos para as redes neurais. Para o NFV, a acurácia foi igual a 0,73 tanto para o conjunto completo de posições quanto para o conjunto lista da IAS, ambas para as redes neurais.

Neste presente trabalho, ao utilizar o conjunto "Todas as Posições", a acurácia variou de 0,66 a 0,79 para o LPV e de 0,69 a 0,82 para o NFV. No caso da "Lista da IAS", os resultados foram de 0,64 a 0,71 para o LPV e de 0,60 a 0,82 para o NFV. Os resultados deste estudo foram muito próximos aos encontrados no trabalho de Rhee *et al.* (2006).

Utilizando os dados deste trabalho, dois algoritmos de interpretação foram avaliados em relação ao desfecho definido pelo Algoritmo Brasileiro: *Stanford HIVdb* e *Rega*. Para o LPV, o melhor desempenho do *Stanford HIVdb*, quando avaliada a AUC, foi igual a 0,67. Para o *Rega*, o melhor desempenho teve uma AUC igual a 0,66. Quando avaliados os casos para o NFV, o algoritmo *Stanford HIVdb* apresentou melhor desempenho quando a AUC foi igual a 0,69. Para o *Rega*, o valor foi inferior, com AUC igual a 0,58.

No mesmo estudo de Pasomsub *et al.* (2010) [12], que também compara sua metodologia com alguns algoritmos de interpretação, o valor da AUC para o algoritmo *Stanford HIVdb* foi de 0,88 (IC: 0,84 - 0,92) para o LPV e igual a 0,90 (IC: 0,87 - 0,93) para o NFV, apresentando resultado superior em relação ao obtido neste estudo.

Uma possível explicação dos resultados relativamente baixos encontrados neste trabalho para os algoritmos *Stanford HIVdb* e *Rega* é a ausência de concordância nas classificações de determinados casos com o Algoritmo Brasileiro.

Neste estudo, as classificações em resistente e não resistente foram provenientes dos resultados fornecidos pelo Algoritmo Brasileiro, um programa com a função de localizar as mutações genéticas do HIV e, por meio de um conjunto de regras pré-estabelecidas, indicar a quais ARVs o vírus está resistente. Apesar de ser amplamente utilizado no país, nenhuma publicação oficial a respeito do desempenho deste algoritmo encontra-se disponível na literatura.

A ausência de dados que comprovem o bom desempenho do Algoritmo Brasileiro coloca em questão os resultados alcançados neste estudo, podendo até mesmo serem superiores aos obtidos. Torna-se, portanto, necessário um estudo que avalie o desempenho do Algoritmo Brasileiro em termos de poder discriminatório, assim como a sua comparação com os algoritmos altamente difundidos na literatura.

Capítulo 7

Conclusão

Nesse estudo, tratamos o problema da seleção das posições de mutação em sequências da protease e na previsão de resistência à terapia antirretroviral nos subtipos B, um problema de difícil solução devido a grande complexidade mutacional do HIV-1.

A contribuição desse trabalho consistiu no desenvolvimento de modelos de RL e PNN, combinando a seleção de variáveis dos modelos por meio das técnicas de *bootstrap*, validação cruzada, *stepwise* e SFS, e previsão de resistência em sequências da protease de pacientes portadores do HIV-1 dos subtipos B em falha terapêutica no Brasil para os inibidores Lopinavir e Nelfinavir.

De maneira geral, a seleção de variáveis permitiu observar aquelas posições mais significativas para cada ARV. Algumas posições que foram selecionadas, além das citadas pela lista da IAS, não foram significativas para aumentar o desempenho dos modelos, com exceção de apenas alguns casos pontuais. Em relação à codificação dos aminoácidos, ambas apresentaram desempenhos próximos, com alguns modelos apresentando melhores resultados para a codificação binária e outros para a codificação de Eisenberg.

A determinação dos melhores pontos de corte para a classificação pelos modelos logísticos e pelas redes neurais probabilísticas foram importantes na obtenção de modelos com melhores valores de sensibilidade, principal enfoque dos nossos modelos.

Os modelos de RL para o LPV apresentaram, para os pontos de corte encontrados, os seguintes valores médios: AUCs variando de 0,78 a 0,82, acurácias de 0,66 a 0,74, sensibilidades de 0,64 a 0,91, especificidades de 0,62 a 0,76 e índices Kappa variando de 0,23 a 0,29. Nas PNNs, esses valores foram: AUCs variando de 0,60 a 0,76, acurácias de 0,64 a 0,79, sensibilidades de 0,56 a 0,73, especificidades de 0,64 a 0,82 e índices Kappa variando de 0,16 a 0,33.

Para o NFV, os valores médios encontrados para os modelos de RL foram: AUCs variando de 0,70 a 0,82, acurácias de 0,71 a 0,82, sensibilidades de 0,53 a 0,73, especificidades de 0,75 a 0,89 e índices Kappa variando de 0,31 a 0,53. Nas PNNs, esses

valores foram: AUCs variando de 0,61 a 0,77, acurácias de 0,60 a 0,82, sensibilidades de 0,48 a 0,75, especificidades de 0,57 a 0,93 e índices Kappa variando de 0,18 a 0,45.

Neste presente estudo, pode-se observar que a RL apresentou melhor desempenho médio quando comparada à PNN. Os modelos obtidos apresentaram desempenhos próximos ao dos algoritmos *Stanford HIVdb* e *Rega*, expondo, em alguns modelos, desempenhos superiores.

Apesar das limitações, os modelos propostos neste trabalho representam uma ferramenta auxiliar na classificação de novos indivíduos em relação ao desenvolvimento de resistência aos inibidores da HIV-protease Lopinavir e Nelfinavir, podendo se tornar úteis na escolha da melhor prática terapêutica para cada indivíduo HIV+.

Como trabalhos futuros, pretende-se: avaliar o desempenho do Algoritmo Brasileiro a partir de resultados fenotípicos provenientes de bancos de dados internacionais disponibilizados na internet e comparar o seu desempenho com os sistemas internacionais *Stanford HIVdb* e *Rega*; identificar padrões de mutações de acordo com o subtipo do HIV e verificar se existem diferentes mutações entre os subtipos que sejam significantes no desenvolvimento de resistência aos ARVs e desenvolver um modelo capaz de reconhecer a presença de resistência aos ARVs independente do subtipo.

Referências Bibliográficas

[1] RAMBAUT, A., POSADA, D., CRANDALL, K. A., et al. "The causes and consequences of HIV evolution." *Nat Rev Genet*, v. 5, n. 1, pp. 52–61, Jan 2004. 1, 2.1.1

[2] KANTOR, R., KATZENSTEIN, D. "Drug resistance in non-subtype B HIV-1." *J Clin Virol*, v. 29, n. 3, pp. 152 – 159, Mar 2004. 1, 2.1.1

[3] BONGERTZ, V., BOU-HABIB, D. C., BRÍGIDO, L. F., et al. "HIV-1 diversity in Brazil: genetic, biologic, and immunologic characterization of HIV-1 strains in three potential HIV vaccine evaluation sites. Brazilian Network for HIV Isolation and Characterization." *J Acquir Immune Defic Syndr*, v. 23, n. 2, pp. 184 – 193, Fev 2000. 1, 2.1.1

[4] UNAIDS. "How to get to zero: Faster. Smarter. Better." Dez 2011. 1

[5] MINISTÉRIO DA SAÚDE, SECRETARIA DE VIGILÂNCIA EM SAÚDE, D. D. D. A. E. H. V. "Boletim Epidemiológico Aids e DST". 2012. Disponível em: <`http://www.aids.gov.br/publicacao/2012/boletim-epidemiologico-aids-e-dst-2012`>. 1

[6] DOURADO, I., VERAS, M. A. D. S. M., BARREIRA, D., et al. "Tendências da epidemia de Aids no Brasil após a terapia anti-retroviral", *Revista de Saúde Pública*, v. 40, pp. 9 – 17, Abr 2006. ISSN: 0034-8910. 1

[7] ZOLOPA, A. R., SHAFER, R. W., WARFORD, A., et al. "HIV-1 genotypic resistance patterns predict response to saquinavir-ritonavir therapy in patients in whom previous protease inhibitor therapy had failed." *Ann Intern Med*, v. 131, n. 11, pp. 813 – 821, Dez 1999. 1

[8] PROSPERI, M. C., ALTMANN, A., ROSEN-ZVI, M., et al. "Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment", *Antivir Ther*, v. 14, n. 3, pp. 433–42, 2009. 1

[9] VAN DER BORGHT, K., VERHEYEN, A., FEYAERTS, M., et al. "Quantitative prediction of integrase inhibitor resistance from genotype through consensus linear regression modeling", *Virology journal*, v. 10, n. 1, pp. 8, 2013. 1

[10] BONET, I., GARCÍA, M. M., SAEYS, Y., et al. "Predicting Human Immunodeficiency Virus (HIV) Drug Resistance Using Recurrent Neural Networks." In: Mira, J., Álvarez, J. R. (Eds.), *IWINAC (1)*, v. 4527, *Lecture Notes in Computer Science*, pp. 234 – 243. Springer, 2007. 1, 2.3, 3, 6

[11] DRAGHICI, S., POTTER, R. B. "Predicting HIV drug resistance with neural networks." *Bioinformatics*, v. 19, n. 1, pp. 98 – 107, Jan 2003. 2.3, 3

[12] PASOMSUB, E., SUKASEM, C., SUNGKANUPARPH, S., et al. "The application of artificial neural networks for phenotypic drug resistance prediction: evaluation and comparison with other interpretation systems." *Jpn J Infect Dis*, v. 63, n. 2, pp. 87 – 94, Mar 2010. 2.4.2, 3, 6

[13] WANG, D., LARDER, B. "Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks." *J Infect Dis*, v. 188, n. 5, pp. 653 – 660, Set 2003. 1, 2.3.2, 2.4.2, 3

[14] BARRÉ-SINOUSSI, F., CHERMANN, J. C., REY, F., et al. "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)." *Science*, v. 220, n. 4599, pp. 868 – 871, Mai 1983. 2.1.1

[15] GALLO, R. C., SARIN, P. S., GELMANN, E. P., et al. "Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS)." *Science*, v. 220, n. 4599, pp. 865 – 867, Mai 1983. 2.1.1

[16] BARIN, F., M'BOUP, S., DENIS, F., et al. "Serological evidence for virus related to simian T-lymphotropic retrovirus III in residents of west Africa." *Lancet*, v. 2, n. 8469-70, pp. 1387 – 1389, 1985. 2.1.1

[17] PEETERS, M., SHARP, P. M. "Genetic diversity of HIV-1: the moving target." *AIDS*, v. 14 Suppl 3, pp. S129 – S140, 2000. 2.1.1

[18] PEETERS, M. "The genetic variability of HIV-1 and its implications." *Transfus Clin Biol*, v. 8, n. 3, pp. 222 – 225, Jun 2001. 2.1.1, 2.2

[19] VARMUS, H. "Retroviruses." *Science*, v. 240, n. 4858, pp. 1427 – 1435, Jun 1988. 2.1.2

[20] NIAID. "NIAID's HIV/AIDS Research Program". 2012. Disponível em: <http://www.niaid.nih.gov/topics/hivaids/Pages/Default.aspx>. 2.1.2, 2.1.3

[21] COFFIN JM, HUGHES SH, V. H. E. *Retroviruses*. Cold Spring Harbor (NY), 1997. Disponível em: <http://www.ncbi.nlm.nih.gov/books/NBK19376/>. 2.1.3

[22] BUSHMAN, F., LANDAU, N. R., EMINI, E. A. "New developments in the biology and treatment of HIV." *Proc Natl Acad Sci U S A*, v. 95, n. 19, pp. 11041 – 11042, Set 1998. 2.1.3, 2.2.1, 2.3

[23] CLAPHAM, P. R., MCKNIGHT, A. "HIV-1 receptors and cell tropism." *Br Med Bull*, v. 58, pp. 43 – 59, 2001. 2.1.3

[24] BEERENWINKEL, N., SING, T., LENGAUER, T., et al. "Computational methods for the design of effective therapies against drug resistant HIV strains." *Bioinformatics*, v. 21, n. 21, pp. 3943 – 3950, Nov 2005. 2.1.3

[25] SOUZA, M. V. N. D., ALMEIDA, M. V. D. "Drogas anti-VIH: passado, presente e perspectivas futuras", *Química Nova*, v. 26, pp. 366 – 372, Mai 2003. ISSN: 0100-4042. 2.1.3, 2.2.1

[26] BAROUCH, D. H. "Challenges in the development of an HIV-1 vaccine." *Nature*, v. 455, n. 7213, pp. 613 – 619, Out 2008. 2.2

[27] FUMERO, E., PODZAMCZER, D. "New patterns of HIV-1 resistance during HAART." *Clin Microbiol Infect*, v. 9, n. 11, pp. 1077 – 1084, Nov 2003. 2.2

[28] MENÉNDEZ-ARIAS, L. "Molecular basis of human immunodeficiency virus drug resistance: an update." *Antiviral Res*, v. 85, n. 1, pp. 210 – 231, Jan 2010. 2.2, 2.2.1

[29] MINISTÉRIO DA SAÚDE, SECRETARIA DE VIGILÂNCIA EM SAÚDE, P. N. D. D. E. A. "Recomendações para Terapia Antirretroviral em Crianças e Adolescentes Infectados pelo HIV: manual de bolso". 2009. 2.2

[30] AIDSINFO. "HIV and Its Treatment - FDA-Approved Anti-HIV Medications". Ago 2012. Disponível em: <http://aidsinfo.nih.gov/contentfiles/ApprovedMedstoTreatHIV_FS_en.pdf>. 2.2.1

[31] ALI, A., BANDARANAYAKE, R. M., CAI, Y., et al. "Molecular Basis for Drug Resistance in HIV-1 Protease." *Viruses*, v. 2, n. 11, pp. 2509 – 2535, Nov 2010. 2.2.1

[32] HIRSCH, M. S., GÜNTHARD, H. F., SCHAPIRO, J. M., et al. "Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel." *Top HIV Med*, v. 16, n. 3, pp. 266 – 285, 2008. 2.3.1

[33] VERMEIREN, H., VAN CRAENENBROECK, E., ALEN, P., et al. "Prediction of HIV-1 drug susceptibility phenotype from the viral genotype using linear regression modeling." *J Virol Methods*, v. 145, n. 1, pp. 47 – 55, Out 2007. 2.3.2, 6

[34] KRZANOWSKI, W. *An Introduction to Statistical Modelling*. Arnold Texts in Statistics Series. John Wiley & Sons, 1998. ISBN: 9780340691854. Disponível em: <http://books.google.com.br/books?id=tXpuQgAACAAJ>. 2.4.1, 2.4.1, 2.5.1

[35] HOSMER, D., LEMESHOW, S., STURDIVANT, R. *Applied Logistic Regression*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN: 9781118548356. Disponível em: <http://books.google.com.br/books?id=bRoxQBIZRd4C>. 2.4.1

[36] GURNEY, K., GURNEY, K. *Neural Networks*. Taylor & Francis Group, 1997. ISBN: 9781857285031. Disponível em: <http://books.google.com.br/books?id=HOsvllRMMP8C>. 2.4.2

[37] GALLANT, S. *Neural Network*. Bradford Bks. Mit Press, 1993. ISBN: 9780262071451. Disponível em: <http://books.google.com.br/books?id=N8i6pTafq1kC>. 2.4.2

[38] SPECHT, D. F. "Probabilistic neural networks", *Neural Netw.*, v. 3, n. 1, pp. 109 – 118, Jan 1990. ISSN: 0893-6080. 2.4.2, 2.4.2

[39] VAFAIE, H., DE JONG, K. "Genetic algorithms as a tool for feature selection in machine learning". In: *Tools with Artificial Intelligence, 1992. TAI'92, Proceedings., Fourth International Conference on*, pp. 200–203. IEEE, 1992. 2.5

[40] MILLER, A. *Subset selection in regression*. CRC Press, 2002. 2.5

[41] AKAIKE, H. "A new look at the statistical model identification", *IEEE Trans. Automatic Control*, v. AC-19, pp. 716 – 723, 1974. ISSN: 0018-9286. System identification and time-series analysis. 2.5.1

[42] EFRON, B. "Bootstrap methods: another look at the jackknife", *The annals of Statistics*, pp. 1–26, 1979. 2.6.1

[43] ZWEIG, M. H., CAMPBELL, G. "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." *Clin Chem*, v. 39, n. 4, pp. 561 – 577, Abr 1993. 2.7.4

[44] LANDIS, J. R., KOCH, G. G. "The measurement of observer agreement for categorical data", *Biometrics*, v. 33, n. 1, pp. 159 – 74., Mar 1977. 2.7.5

[45] KOHAVI, R. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". pp. 1137–1143. Morgan Kaufmann, 1995. 2.7.6.1, 2.7.6.2

[46] WANG, K., JENWITHEESUK, E., SAMUDRALA, R., et al. "Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance", *Antiviral Therapy*, v. 9, n. 3, pp. 343–352, 2004. 3

[47] RHEE, S.-Y., TAYLOR, J., WADHERA, G., et al. "Genotypic predictors of human immunodeficiency virus type 1 drug resistance", *Proceedings of the National Academy of Sciences*, v. 103, n. 46, pp. 17355–17360, 2006. 3, 6

[48] LARDER, B., WANG, D., REVELL, A., et al. "The development of artificial neural networks to predict virological response to combination HIV therapy", *Antiviral therapy*, v. 12, n. 1, pp. 15, 2007. 3

[49] SILVA, R. M. *Algoritmo Genético e Kernel Discriminante de Fisher Aplicado a Identificação de Mutações de Resistência do HIV-1 aos Inibidores Antiretrovirais da Protease*. Tese de d. sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2009. 3

[50] JOHNSON, V. A., CALVEZ, V., GUNTHARD, H., et al. "Update of the drug resistance mutations in HIV-1: March 2013", *Top Antivir Med*, v. 21, pp. 4–12, 2013. 4.3

[51] STANFORD, U. "HIVdb Program - Genotypic Resistance Interpretation Algorithm - Version 6.2.0". 2013. Disponível em: <`http://sierra2.stanford.edu/sierra/servlet/JSierra`>. 4.5

[52] VAN LAETHEM, K., DE LUCA, A., ANTINORI, A., et al. "A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients", *Antiviral therapy*, v. 7, n. 2, pp. 123–9, 2002. 4.5

[53] GARRIGA, C., PÉREZ-ELÍAS, M. J., DELGADO, R., et al. "Mutational patterns and correlated amino acid substitutions in the HIV-1 protease after

virological failure to nelfinavir-and lopinavir/ritonavir-based treatments", *Journal of medical virology*, v. 79, n. 11, pp. 1617–1628, 2007. 6

[54] MO, H., KING, M. S., KING, K., et al. "Selection of resistance in protease inhibitor-experienced, human immunodeficiency virus type 1-infected subjects failing lopinavir-and ritonavir-based therapy: mutation patterns and baseline correlates", *Journal of virology*, v. 79, n. 6, pp. 3329–3338, 2005. 6

[55] SANTOS, J. R., LLIBRE, J. M., IMAZ, A., et al. "Mutations in the protease gene associated with virological failure to lopinavir/ritonavir-containing regimens", *Journal of antimicrobial chemotherapy*, v. 67, n. 6, pp. 1462–1469, 2012. 6

[56] ŠAŠKOVÁ, K. G., KOŽÍŠEK, M., LEPŠÍK, M., et al. "Enzymatic and structural analysis of the I47A mutation contributing to the reduced susceptibility to HIV protease inhibitor lopinavir", *Protein Science*, v. 17, n. 9, pp. 1555–1564, 2008. 6

[57] DE MENDOZA, C., VALER, L., BACHELER, L., et al. "Prevalence of the HIV-1 protease mutation I47A in clinical practice and association with lopinavir resistance", *AIDS*, v. 20, pp. 1071–1074, 2006. 6

[58] RHEE, S.-Y., TAYLOR, J., FESSEL, W. J., et al. "HIV-1 protease mutations and protease inhibitor cross-resistance", *Antimicrobial agents and chemotherapy*, v. 54, n. 10, pp. 4253–4261, 2010. 6

[59] SAAH, A. J., HAAS, D. W., DINUBILE, M. J., et al. "Treatment with indinavir, efavirenz, and adefovir after failure of nelfinavir therapy", *Journal of Infectious Diseases*, v. 187, n. 7, pp. 1157–1162, 2003. 6

[60] SANTOS, A. F., SOARES, M. A. "The impact of the nelfinavir resistance-conferring mutation D30N on the susceptibility of HIV-1 subtype B to other protease inhibitors", *Memórias do Instituto Oswaldo Cruz*, v. 106, n. 2, pp. 177–181, 2011. 6

[61] RHEE, S.-Y., GONZALES, M. J., KANTOR, R., et al. "Human immunodeficiency virus reverse transcriptase and protease sequence database", *Nucleic acids research*, v. 31, n. 1, pp. 298–303, 2003. 6

[62] DEFORCHE, K., SILANDER, T., CAMACHO, R., et al. "Analysis of HIV-1 pol sequences using Bayesian networks: implications for drug resistance", *Bioinformatics*, v. 22, n. 24, pp. 2975–2979, 2006. 6

Anexo

Este item apresenta os resultados encontrados para cada cada modelo obtido para os inibidores da HIV-protease LPV e NFV.

Nas tabelas abaixo AUC indica a área sob a Curva ROC, A indica a acurácia, S a sensibilidade, E a representa a especificidade *e* K o índice *Kappa.*

Desempenho dos modelos logísticos para o Lopinavir utilizando o critério de seleção de variáveis $\geq 50\%$, a codificação binária e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,345					*Ponto de Corte: 0,500*				
	AUC	*A*	*S*	*E*	*K*	*AUC*	*A*	*S*	*E*	*K*
Modelo 1	0,83	0,67	0,89	0,63	0,29	0,83	0,80	0,53	0,85	0,33
Modelo 2	0,80	0,62	0,89	0,57	0,24	0,80	0,76	0,53	0,80	0,26
Modelo 3	0,86	0,69	1,00	0,63	0,34	0,86	0,79	0,68	0,81	0,38
Modelo 4	0,78	0,63	0,84	0,59	0,23	0,78	0,79	0,53	0,84	0,31
Modelo 5	0,84	0,70	0,89	0,66	0,32	0,84	0,79	0,663	0,82	0,36
Média	0,82	0,66	0,91	0,62	0,29	0,82	0,79	0,58	0,82	0,33

Desempenho dos modelos logísticos para o Lopinavir utilizando o critério de seleção de variáveis $\geq 60\%$, a codificação binária e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,383					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,80	0,70	0,63	0,72	0,23	0,80	0,792	0,47	0,85	0,29
Modelo 2	0,78	0,66	0,79	0,64	0,25	0,78	0,73	0,47	0,77	0,19
Modelo 3	0,80	0,72	0,68	0,73	0,27	0,81	0,81	0,58	0,85	0,36
Modelo 4	0,79	0,70	0,68	0,70	0,25	0,79	0,81	0,53	0,86	0,34
Modelo 5	0,80	0,72	0,74	0,72	0,29	0,80	0,79	0,63	0,82	0,36
Média	0,79	0,70	0,71	0,70	0,26	0,79	0,79	0,54	0,83	0,31

Desempenho dos modelos logísticos para o Lopinavir utilizando o critério de seleção de variáveis $\geq 50\%$ e $\geq 60\%$, a codificação de Eisenberg e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,367					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,82	0,69	0,84	0,66	0,29	0,82	0,82	0,53	0,88	0,37
Modelo 2	0,78	0,63	0,84	0,59	0,23	0,78	0,73	0,58	0,75	0,24
Modelo 3	0,83	0,70	0,74	0,69	0,26	0,83	0,82	0,53	0,88	0,37
Modelo 4	0,84	0,68	0,84	0,65	0,28	0,84	0,82	0,53	0,88	0,37
Modelo 5	0,82	0,72	0,74	0,72	0,29	0,82	0,82	0,63	0,85	0,40
Média	0,82	0,68	0,80	0,66	0,27	0,82	0,80	0,56	0,85	0,35

Desempenho dos modelos logísticos para o Lopinavir utilizando o critério de seleção de variáveis $\geq 50\%$, a codificação binária e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,389					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,77	0,76	0,63	0,76	0,28	0,77	0,82	0,37	0,90	0,27
Modelo 2	0,80	0,68	0,79	0,66	0,27	0,80	0,78	0,68	0,79	0,35
Modelo 3	0,86	0,75	0,89	0,73	0,39	0,86	0,80	0,79	0,80	0,43
Modelo 4	0,81	0,67	0,68	0,67	0,22	0,81	0,80	0,63	0,83	0,37
Modelo 5	0,78	0,66	0,74	0,64	0,22	0,78	0,77	0,68	0,78	0,34
Média	0,80	0,70	0,75	0,69	0,28	0,80	0,79	0,63	0,82	0,35

Desempenho dos modelos logísticos para o Lopinavir utilizando o critério de seleção de variáveis $\geq 60\%$, a codificação binária e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,345					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,83	0,66	0,89	0,63	0,29	0,83	0,80	0,53	0,85	0,33
Modelo 2	0,80	0,63	0,89	0,58	0,24	0,80	0,76	0,53	0,80	0,26
Modelo 3	0,86	0,69	1,00	0,63	0,34	0,86	0,79	0,68	0,81	0,38
Modelo 4	0,78	0,64	0,84	0,59	0,23	0,78	0,79	0,53	0,84	0,31
Modelo 5	0,84	0,70	0,89	0,66	0,32	0,84	0,79	0,63	0,82	0,36
Média	0,82	0,66	0,91	0,62	0,29	0,82	0,79	0,58	0,82	0,33

Desempenho dos modelos logísticos para o Lopinavir utilizando o critério de seleção de variáveis $\geq 50\%$, a codificação de Eisenberg e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,447					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,74	0,73	0,47	0,77	0,19	0,74	0,78	0,47	0,84	0,27
Modelo 2	0,80	0,70	0,79	0,69	0,29	0,80	0,70	0,68	0,71	0,25
Modelo 3	0,87	0,75	0,63	0,77	0,30	0,87	0,79	0,58	0,83	0,34
Modelo 4	0,83	0,76	0,68	0,77	0,33	0,83	0,78	0,63	0,80	0,33
Modelo 5	0,84	0,75	0,63	0,77	0,30	0,84	0,76	0,58	0,79	0,29
Média	0,82	0,74	0,64	0,76	0,28	0,82	0,76	0,59	0,79	0,30

Desempenho dos modelos logísticos para o Lopinavir utilizando o critério de seleção de variáveis $\geq 60\%$, a codificação de Eisenberg e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,355					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,76	0,66	0,74	0,64	0,22	0,76	0,82	0,47	0,87	0,35
Modelo 2	0,77	0,65	0,74	0,63	0,21	0,77	0,74	0,58	0,77	0,26
Modelo 3	0,80	0,70	0,68	0,70	0,24	0,80	0,82	0,53	0,88	0,37
Modelo 4	0,81	0,67	0,79	0,65	0,26	0,81	0,83	0,53	0,89	0,39
Modelo 5	0,77	0,66	0,68	0,66	0,21	0,77	0,82	0,53	0,87	0,36
Média	0,78	0,67	0,73	0,66	0,23	0,78	0,81	0,53	0,86	0,34

Desempenho dos modelos logísticos para o Nelfinavir utilizando o critério de seleção de variáveis $\geq 50\%$, a codificação binária e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,428					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	Kappa	*AUC*	*Ac*	*Se*	*Es*	Kappa
Modelo 1	0,77	0,79	0,67	0,83	0,47	0,77	0,82	0,6	0,88	0,49
Modelo 2	0,76	0,73	0,67	0,75	0,36	0,76	0,79	0,67	0,83	0,47
Modelo 3	0,77	0,76	0,67	0,79	0,41	0,77	0,82	0,67	0,87	0,53
Modelo 4	0,76	0,73	0,67	0,75	0,36	0,76	0,76	0,67	0,79	0,41
Média	0,76	0,75	0,67	0,78	0,40	0,76	0,80	0,65	0,84	0,47

Desempenho dos modelos logísticos para o Nelfinavir utilizando o critério de seleção de variáveis$\geq 60\%$, a codificação binária e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,558					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	Kappa	*AUC*	*Ac*	*Se*	*Es*	Kappa
Modelo 1	0,74	0,82	0,60	0,89	0,51	0,74	0,82	0,60	0,89	0,51
Modelo 2	0,74	0,82	0,60	0,89	0,51	0,74	0,72	0,60	0,76	0,32
Modelo 3	0,74	0,82	0,60	0,89	0,51	0,74	0,82	0,60	0,89	0,51
Modelo 4	0,74	0,82	0,60	0,89	0,51	0,74	0,72	0,60	0,76	0,32
Média	0,74	0,82	0,60	0,89	0,51	0,74	0,77	0,60	0,83	0,41

Desempenho dos modelos logísticos para o Nelfinavir utilizando o critério de seleção de variáveis$\geq 50\%$, a codificação de Eisenberg e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,469					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,75	0,75	0,60	0,80	0,37	0,75	0,77	0,60	0,82	0,40
Modelo 2	0,65	0,75	0,60	0,80	0,37	0,65	0,75	0,60	0,80	0,37
Modelo 3	0,76	0,78	0,67	0,81	0,44	0,76	0,78	0,67	0,81	0,44
Modelo 4	0,75	0,76	0,67	0,79	0,41	0,75	0,76	0,67	0,79	0,41
Média	0,73	0,76	0,63	0,80	0,40	0,73	0,76	0,63	0,81	0,40

Desempenho dos modelos logísticos para o Nelfinavir utilizando o critério de seleção de variáveis≥ 60%, a codificação de Eisenberg e o conjunto "Lista da IAS" segundo os pontos de corte.

	PC = 0,484					*PC = 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,72	0,75	0,60	0,80	0,37	0,72	0,75	0,60	0,80	0,37
Modelo 2	0,67	0,75	0,60	0,80	0,37	0,67	0,75	0,60	0,80	0,37
Modelo 3	0,72	0,75	0,60	0,80	0,37	0,72	0,75	0,60	0,80	0,37
Modelo 4	0,72	0,75	0,60	0,80	0,37	0,72	0,75	0,60	0,80	0,37
Média	0,71	0,75	0,60	0,80	0,37	0,71	0,75	0,60	0,80	0,37

Desempenho dos modelos logísticos para o Nelfinavir utilizando o critério de seleção de variáveis≥ 50%, a codificação binária e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,454					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,70	0,78	0,47	0,88	0,37	0,70	0,78	0,47	0,88	0,37
Modelo 2	0,70	0,70	0,60	0,73	0,28	0,70	0,70	0,60	0,73	0,28
Modelo 3	0,70	0,78	0,47	0,88	0,37	0,70	0,79	0,47	0,89	0,39
Modelo 4	0,70	0,70	0,60	0,73	0,28	0,70	0,78	0,47	0,88	0,37
Média	0,70	0,74	0,53	0,81	0,33	0,70	0,76	0,50	0,85	0,35

Desempenho dos modelos logísticos para o Nelfinavir utilizando o critério de seleção de variáveis≥ 60%, a codificação binária e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,402					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,71	0,71	0,60	0,75	0,31	0,71	0,71	0,60	0,75	0,31
Modelo 2	0,71	0,71	0,60	0,75	0,31	0,71	0,71	0,60	0,75	0,31
Modelo 3	0,71	0,71	0,60	0,75	0,31	0,71	0,71	0,60	0,75	0,31
Modelo 4	0,71	0,71	0,60	0,75	0,31	0,71	0,71	0,60	0,75	0,31
Média	0,71	0,71	0,60	0,75	0,31	0,71	0,71	0,60	0,75	0,31

Desempenho dos modelos logísticos para o Nelfinavir utilizando o critério de seleção de variáveis≥ 50%, a codificação de Eisenberg e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,373					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,79	0,77	0,67	0,80	0,42	0,79	0,82	0,63	0,88	0,52
Modelo 2	0,73	0,72	0,67	0,74	0,34	0,73	0,78	0,63	0,82	0,43
Modelo 3	0,78	0,79	0,60	0,84	0,43	0,78	0,78	0,60	0,84	0,43
Modelo 4	0,78	0,74	0,67	0,77	0,38	0,78	0,78	0,63	0,82	0,43
Média	0,77	0,75	0,65	0,79	0,39	0,77	0,79	0,62	0,84	0,45

Desempenho dos modelos logísticos para o Nelfinavir utilizando o critério de seleção de variáveis≥ 60%, a codificação de Eisenberg e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,417					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,82	0,82	0,73	0,84	0,53	0,82	0,82	0,63	0,88	0,52
Modelo 2	0,83	0,82	0,73	0,84	0,53	0,83	0,82	0,73	0,84	0,53
Modelo 3	0,82	0,82	0,73	0,84	0,53	0,82	0,82	0,40	0,95	0,41
Modelo 4	0,82	0,82	0,73	0,84	0,53	0,82	0,82	0,73	0,84	0,53
Média	0,82	0,82	0,73	0,84	0,53	0,82	0,82	0,63	0,88	0,50

Desempenho das redes neurais probabilísticas para o Lopinavir utilizando o critério de seleção de variáveis≥ 50%, a codificação binária e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,501					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,66	0,73	0,53	0,73	0,18	0,66	0,67	0,58	0,69	0,17
Modelo 2	0,59	0,74	0,47	0,74	0,15	0,59	0,64	0,47	0,67	0,09
Modelo 3	0,69	0,75	0,58	0,74	0,24	0,69	0,70	0,68	0,71	0,25
Modelo 4	0,54	0,73	0,53	0,73	0,18	0,54	0,61	0,58	0,61	0,11
Modelo 5	0,73	0,75	0,68	0,75	0,30	0,73	0,70	0,74	0,69	0,27
Média	0,64	0,71	0,56	0,74	0,21	0,64	0,66	0,61	0,67	0,18

Desempenho das redes neurais probabilísticas para o Lopinavir utilizando o critério de seleção de variáveis≥ 60%, a codificação binária e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,502					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,65	0,68	0,58	0,70	0,18	0,65	0,66	0,58	0,67	0,16
Modelo 2	0,60	0,66	0,58	0,68	0,17	0,60	0,64	0,58	0,65	0,14
Modelo 3	0,69	0,74	0,68	0,75	0,30	0,69	0,72	0,68	0,73	0,27
Modelo 4	0,55	0,67	0,58	0,69	0,17	0,55	0,64	0,58	0,65	0,14
Modelo 5	0,68	0,71	0,63	0,73	0,24	0,68	0,69	0,63	0,70	0,21
Média	0,63	0,69	0,61	0,71	0,21	0,63	0,67	0,61	0,68	0,19

Desempenho das redes neurais probabilísticas para o Lopinavir utilizando o critério de seleção de variáveis ≥ 50% e≥ 60%, a codificação de Eisenberg e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,500				
	AUC	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,60	0,66	0,63	0,67	0,19
Modelo 2	0,60	0,62	0,68	0,60	0,16
Modelo 3	0,63	0,68	0,68	0,68	0,23
Modelo 4	0,56	0,59	0,63	0,58	0,12
Modelo 5	0,60	0,63	0,53	0,65	0,11
Média	0,60	0,64	0,63	0,64	0,16

Desempenho das redes neurais probabilísticas para o Lopinavir utilizando o critério de seleção de variáveis≥ 50%, a codificação binária e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,480					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,76	0,78	0,58	0,81	0,31	0,76	0,78	0,58	0,81	0,31
Modelo 2	0,75	0,73	0,63	0,75	0,26	0,75	0,73	0,63	0,75	0,26
Modelo 3	0,75	0,80	0,53	0,85	0,33	0,75	0,80	0,53	0,85	0,33
Modelo 4	0,76	0,82	0,58	0,86	0,38	0,76	0,81	0,53	0,86	0,34
Modelo 5	0,77	0,82	0,58	0,86	0,38	0,77	0,82	0,58	0,86	0,38
Média	0,76	0,79	0,58	0,82	0,33	0,76	0,79	0,57	0,82	0,32

Desempenho das redes neurais probabilísticas para o Lopinavir utilizando o critério de seleção de variáveis≥ 60%, a codificação binária e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,001					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,76	0,67	0,74	0,66	0,24	0,76	0,82	0,53	0,87	0,36
Modelo 2	0,74	0,64	0,84	0,60	0,24	0,74	0,74	0,58	0,76	0,25
Modelo 3	0,74	0,72	0,58	0,75	0,23	0,74	0,82	0,53	0,87	0,36
Modelo 4	0,76	0,64	0,74	0,62	0,20	0,76	0,82	0,53	0,87	0,36
Modelo 5	0,76	0,64	0,74	0,62	0,20	0,76	0,82	053	0,87	0,36
Média	0,75	0,66	0,73	0,65	0,22	0,75	0,80	0,54	0,85	0,33

Desempenho das redes neurais probabilísticas para o Lopinavir utilizando o critério de seleção de variáveis≥ 50%, a codificação de Eisenberg e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,429					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,65	0,67	0,63	0,68	0,20	0,65	0,66	0,47	0,70	0,12
Modelo 2	0,65	0,68	0,58	0,70	0,18	0,65	0,70	0,53	0,74	0,19
Modelo 3	0,67	0,70	0,58	0,72	0,20	0,67	0,74	0,53	0,78	0,24
Modelo 4	0,68	0,71	0,63	0,73	0,24	0,68	076	0,53	0,80	0,26
Modelo 5	0,66	0,59	0,68	0,57	0,14	0,66	0,64	0,63	0,64	0,16
Média	0,66	0,67	0,62	0,68	0,19	0,66	0,70	0,54	0,73	0,19

Desempenho das redes neurais probabilísticas para o Lopinavir utilizando o critério de seleção de variáveis≥ 60%, a codificação de Eisenberg e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,462					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,74	0,78	0,58	0,81	0,31	0,74	0,77	0,53	0,81	0,27
Modelo 2	0,73	0,75	0,63	0,77	0,30	0,73	0,75	0,63	0,77	0,27
Modelo 3	0,73	0,78	0,58	0,81	0,31	0,73	0,78	0,58	0,82	0,32
Modelo 4	0,75	0,82	0,58	0,86	0,38	0,75	0,81	0,53	0,86	0,34
Modelo 5	0,75	0,80	0,58	0,84	0,35	0,75	0,80	0,58	0,84	0,35
Média	0,74	0,78	0,59	0,82	0,33	0,74	0,78	0,57	0,82	0,32

Desempenho das redes neurais probabilísticas para o Nelfinavir utilizando o critério de seleção de variáveis≥ 50%, a codificação binária e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,500				
	AUC	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,63	0,59	0,70	0,56	0,19
Modelo 2	0,58	0,58	0,60	0,57	0,13
Modelo 3	0,69	0,62	0,73	0,59	0,24
Modelo 4	0,67	0,61	0,70	0,58	0,21
Média	0,64	0,60	0,68	0,57	0,19

Desempenho das redes neurais probabilísticas para o Nelfinavir utilizando o critério de seleção de variáveis≥ 60%, a codificação binária e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,496					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,68	0,60	0,77	0,55	0,22	0,68	0,62	0,70	0,60	0,23
Modelo 2	0,67	0,60	0,70	0,57	0,20	0,67	0,66	0,60	0,67	0,22
Modelo 3	0,68	0,64	0,70	0,62	0,25	0,68	0,66	0,63	0,66	0,24
Modelo 4	0,64	0,60	0,67	0,58	0,18	0,64	0,60	0,60	0,60	0,15
Média	0,67	0,61	0,71	0,58	0,21	0,67	0,63	0,63	0,63	0,21

Desempenho das redes neurais probabilísticas para o Nelfinavir utilizando o critério de seleção de variáveis≥ 50%, a codificação de Eisenberg e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,500				
	AUC	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,59	0,59	0,60	0,59	0,14
Modelo 2	0,61	0,62	0,60	0,62	0,17
Modelo 3	0,61	0,62	0,63	0,61	0,19
Modelo 4	0,62	0,62	0,63	0,62	0,20
Média	0,61	0,61	0,62	0,61	0,18

Desempenho das redes neurais probabilísticas para o Nelfinavir utilizando o critério de seleção de variáveis≥ 60%, a codificação de Eisenberg e o conjunto "Lista da IAS" segundo os pontos de corte.

	Ponto de Corte: 0,579					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,70	0,77	0,47	0,86	0,34	0,70	0,70	0,63	0,73	0,31
Modelo 2	0,70	0,79	0,50	0,88	0,40	0,70	0,71	0,57	0,76	0,29
Modelo 3	0,73	0,79	0,50	0,88	0,40	0,73	0,71	0,63	0,74	0,32
Modelo 4	0,69	0,78	0,47	0,87	0,36	0,69	0,69	0,57	0,73	0,26
Média	0,71	0,78	0,48	0,88	0,38	0,71	0,70	0,60	0,74	0,29

Desempenho das redes neurais probabilísticas para o Nelfinavir utilizando o critério de seleção de variáveis≥ 50%, a codificação binária e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,392					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,73	0,67	0,73	0,65	0,30	0,73	0,74	0,60	0,78	0,34
Modelo 2	0,72	0,67	0,60	0,65	0,25	0,72	0,73	0,50	0,80	0,29
Modelo 3	0,75	0,73	0,73	0,73	0,38	0,75	0,77	0,63	0,81	0,41
Modelo 4	0,69	0,69	0,67	0,69	0,30	0,69	0,70	0,37	0,81	0,18
Média	0,72	0,69	0,68	0,69	0,31	0,72	0,73	0,53	0,80	0,31

Desempenho das redes neurais probabilísticas para o Nelfinavir utilizando o critério de seleção de variáveis≥ 60%, a codificação binária e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,696					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,72	0,84	0,47	0,96	0,49	0,72	0,76	0,53	0,83	0,36
Modelo 2	0,71	0,79	0,50	0,88	0,40	0,71	0,72	0,57	0,77	0,30
Modelo 3	0,73	0,82	0,47	0,94	0,46	0,73	0,78	0,53	0,86	0,40
Modelo 4	0,71	0,82	0,47	0,93	0,44	0,71	0,77	0,53	0,84	0,37
Média	0,72	0,82	0,48	0,93	0,45	0,72	0,76	0,54	0,83	0,36

Desempenho das redes neurais probabilísticas para o Nelfinavir utilizando o critério de seleção de variáveis$\geq$ 50%, a codificação de Eisenberg e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,264					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,76	070	0,77	0,67	0,35	0,76	0,71	0,53	0,77	0,28
Modelo 2	0,77	0,69	0,77	0,66	0,33	0,77	0,77	0,60	0,82	0,40
Modelo 3	0,76	0,69	0,70	0,68	0,31	0,76	0,74	0,57	0,79	0,33
Modelo 4	0,78	0,71	0,77	0,69	0,37	0,78	0,77	0,60	0,82	0,40
Média	0,77	0,70	0,75	0,68	0,34	0,77	0,75	0,58	0,80	0,35

Desempenho das redes neurais probabilísticas para o Nelfinavir utilizando o critério de seleção de variáveis$\geq$ 60%, a codificação de Eisenberg e o conjunto "Todas as Posições" segundo os pontos de corte.

	Ponto de Corte: 0,001					*Ponto de Corte: 0,500*				
	AUC	*Ac*	*Se*	*Es*	*Kappa*	*AUC*	*Ac*	*Se*	*Es*	*Kappa*
Modelo 1	0,76	0,71	0,73	0,71	0,36	0,76	0,76	0,50	0,84	0,34
Modelo 2	0,76	0,71	0,73	0,71	0,36	0,76	0,74	0,60	0,79	0,36
Modelo 3	0,75	0,69	0,60	0,72	0,27	0,75	0,72	0,53	0,78	0,29
Modelo 4	0,77	0,71	0,73	0,71	0,36	0,77	0,74	0,63	0,78	0,37
Média	0,76	0,71	0,70	0,71	0,34	0,76	0,74	0,57	0,80	0,34