UNIVERSIDADE FEDERAL DE GOIÁS – REGIONAL CATALÃO
UNIDADE ACADÊMICA ESPECIAL DE MATEMÁTICA E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM E OTIMIZAÇÃO

Fernando Henrique da Silva

ESTUDO E DESENVOLVIMENTO DE MÉTODOS PARA PREDIÇÃO DE DOADORES DE SANGUE

DISSERTAÇÃO DE MESTRADO

CATALÃO – GO, 2018

sistema de bibliotecas ufg

PRPG
PRÓ-REITORIA DE PÓS-GRADUAÇÃO

UFG

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1. Identificação do material bibliográfico: **[X] Dissertação** **[ ] Tese**

2. Identificação da Tese ou Dissertação:

Nome completo do autor: Fernando Henrique da Silva

Título do trabalho: Estudo e desenvolvimento de Métodos para predição de doadores de sangue.

3. Informações de acesso ao documento:

Concorda com a liberação total do documento [X] SIM [ ] NÃO1

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF da tese ou dissertação.

Fernando Henrique da Silva
Assinatura do(a) autor(a)2

Ciente e de acordo:

Sérgio F. da Silva
Assinatura do(a) orientador(a)²

Data: 16 / 02 / 18

1 Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

2 A assinatura deve ser escaneada.

FERNANDO HENRIQUE DA SILVA

ESTUDO E DESENVOLVIMENTO DE MÉTODOS PARA PREDIÇÃO DE DOADORES DE SANGUE

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre em Modelagem e Otimização pela Universidade Federal de Goiás – Regional Catalão.

Orientador:
Prof. Dr. Sérgio Francisco Silva

CATALÃO – GO

2018

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Silva, Fernando Henrique da
ESTUDO E DESENVOLVIMENTO DE MÉTODOS PARA PREDIÇÃO DE DOADORES DE SANGUE [manuscrito] / Fernando Henrique da Silva. - 2018.
80 f.

Orientador: Prof. Dr. Sérgio Francisco da Silva.
Dissertação (Mestrado) - Universidade Federal de Goiás, Unidade Acadêmica Especial de Matemática e Tecnologia, Catalão, Programa de Pós-Graduação em Modelagem e Otimização, Catalão, 2018.

Inclui siglas, abreviaturas, lista de figuras, lista de tabelas.

1. Classificadores. 2. Doação de Sangue. 3. Sistemas de Recomendação. I. Silva, Sérgio Francisco da, orient. II. Título.

CDU 51

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
REGIONAL CATALÃO
Programa de Pós-Graduação em Modelagem e Otimização - PPGMO

Defesa Nº _____

Ata de Defesa Pública – Dissertação de Mestrado

Aos 16 dias do mês de fevereiro do ano de 2018, às 14 h: 00 min, reuniram-se os componentes da banca examinadora, professores(as) Dr.(a) Sérgio Francisco da Silva (presidente e orientador), Dr. (a) Tércio Alberto dos Santos Filho, Dr. (a) Marcos Aurélio Batista e Dr. (a) _____ para, em sessão pública realizada na Sala 3 (Laboratório de Controle Operacional) - Bloco B, da Regional Catalão (RC), da Universidade Federal de Goiás (UFG), procederem com a avaliação do trabalho intitulado: "Estudo e Desenvolvimento de Métodos para Predição de Doadores de Sangue", em nível de Mestrado, área de concentração *Modelagem e Otimização*, de autoria de Fernando Henrique da Silva, discente do Programa de Pós-Graduação em Modelagem e Otimização (PPGMO) da UFG/RC. A sessão foi aberta pelo presidente da banca, que fez a apresentação formal dos membros da banca. A seguir, a palavra foi concedida ao discente que, dentro do tempo regulamentar, procedeu a apresentação de seu trabalho. Terminada a apresentação, cada membro da banca arguiu o candidato, tendo-se adotado o sistema de diálogo sequencial. Terminada a fase de arguição, procedeu-se a avaliação do trabalho. Os membros da banca consideraram o trabalho final: (X) **Aprovado** ou ( ) **Reprovado**. Cumpridas as formalidades de pauta, às 16 h: 11 min a presidência da mesa encerrou a sessão e para constar, eu Sérgio Francisco da Silva, lavrei a presente Ata que, depois de lida e aprovada, segue assinada pelos membros da banca examinadora e pelo discente e, posteriormente, será homologada pelo Colegiado do PPGMO.

Catalão-GO, 16 de fevereiro de 2018.

Sérgio F. da Silva
Prof.(a) Dr. (a):
Programa de Pós-Graduação em Modelagem e Otimização, UFG/RC.
(Presidente da Banca)

Prof.(a) Dr. (a):
Instituição
Departamento de Ciência da Computação
UFG/RC

Prof.(a) Dr. (a):
Instituição
Programa de Pós Graduação em Modelagem e Otimização
UFG/RC

Fernando Henrique da Silva
Discente:
Programa de Pós-Graduação em Modelagem e Otimização, UFG/RC.

Dedico este trabalho a minha família que foi meu esteio a todo momento, em especial minha esposa, mãe, pai e irmão que não mediram esforços para que chegasse até aqui.

Agradecimentos

Agradeço primeiramente a Deus, pelo dom da vida e por me dar a graça de realizar o sonho de conquistar a titulação de mestre. Agradeço a minha família que deu um imenso e efetivo apoio para para conquistar esse mérito. Deixo também meu agradecimento a UFG nas pessoas do meu orientador Prof. Sérgio Francisco da Silva, do Coordenador da PPGMO, Prof. Celso Vieira Abud e do Diretor da Regional Catalão Thiago Jabur Bittar.

"Any sufficiently advanced technology is indistinguishable from magic"
"Qualquer tecnologia suficientemente avançada é indistinta de magia."
Arthur C. Clarke

RESUMO

SILVA, F. H.. *ESTUDO E DESENVOLVIMENTO DE MÉTODOS PARA PREDIÇÃO DE DOADORES DE SANGUE*. 2018. 80 f. Dissertação (Mestrado em Modelagem e Otimização) – Unidade Acadêmica Especial de Matemática e Tecnologia, Universidade Federal de Goiás – Regional Catalão, Catalão – GO.

Unidades Hemoterápicas encontram dificuldades para otimizar a busca por doadores de sangue em situações de emergência, assim como para manter seus estoques sanguíneos em níveis adequados. Por outro lado, a utilização de técnicas computacionais de predição tem obtido ótimos resultados em várias áreas do conhecimento, podendo ser vista como uma ferramenta fundamental na obtenção de doações de sangue, contudo, são pouco exploradas neste setor. Dado esta lacuna, este trabalho objetivou em analisar e desenvolver técnicas de predição para otimizar a busca por doadores com maior taxa de conversão à doação, com foco em técnicas de mineração de dados. Para isto, primeiramente analisou-se o desempenho de classificadores tradicionais da literatura aplicados a uma base de dados real, o que produziu resultados de predição insatisfatórios. Na busca de resultados de maior qualidade foi proposta uma abordagem de recomendação dos top-k, que utiliza heurísticas para a estimar a confiança em doação. Experimentos computacionais demonstram que a abordagem de recomendação top-k alcança bons resultados para todas as três heurísticas desenvolvidas. A heurística baseada em vetores de suporte obteve taxas de precisão de 94,09% entre os top-10 recomendados, chegando a 99,90% de precisão para o top-1, para a mesma base em que não se obteve sucesso com o uso de classificadores. É esperado que os resultados deste trabalho contribuam para a comunidade acadêmica devido a variedade de classificadores analisados e principalmente pela abordagem de recomendações top-k proposta. Futuramente esta abordagem poderá ser melhor analisada com outras bases de dados e até mesmo aprimorada pelo desenvolvimento de novas heurísticas. Além disso, acredita-se que a abordagem top-k desenvolvida possa ser utilizada em sistemas predição na área da saúde, com foco na predição de doadores de sangue principalmente em situações de emergência.

Palavras-chaves: Classificadores, Doação de Sangue, Sistemas de Recomendação.

ABSTRACT

SILVA, F. H.. *ESTUDO E DESENVOLVIMENTO DE MÉTODOS PARA PREDIÇÃO DE DOADORES DE SANGUE*. 2018. 80 f. Master Thesis in Modelling and Optimization – Unidade Acadêmica Especial de Matemática e Tecnologia, Universidade Federal de Goiás – Regional Catalão, Catalão – GO.

Hemotherapy units has difficulties to optimize the search for blood donors in emergency situations, as well as to keep their blood stocks at adequate levels. On the other hand, the use of computational techniques for prediction has obtained promissing results in several areas of knowledge, and can be seen as a fundamental tool in obtaining blood donations, however, are little explored in this sector. Given this gap, this research aimed to analyze and develop prediction techniques to optimize the search for donors with higher conversion rate to the donation, focusing on data mining techniques. For this, we first analyzed the performance of traditional literature classifiers applied to a real database, which produced unsatisfactory prediction results. Seeking for higher quality results we propose a top-k recommendation approach of blood donors, which uses heuristics to estimate a confidence degree in donation. Computational experiments show that the top-k recommendation approach achieves good results for all three developed heuristics. The support vector-based heuristic achieving 94.09% of precision among the top-10 recommended, and 99.90% of precision for top-1, for the same data set that the classifiers were not successful. It is expected that the results of this research will contribute to the academic community due to the variety of classifiers analyzed and especially due to the proposed top-k recommendations approach. In the future, this approach can be better analyzed with other databases and even improved by the development of new heuristics. In addition, it is believed that the developed top-k approach can be used in health prediction systems, with a focus on predicting blood donors, especially in emergency situations.

Keywords: Blood Donation, Classifiers, Recommendation Systems.

LISTA DE FIGURAS

Figura 2.1 – Classificação dos Grupos e Subgrupos Sanguíneos. 28
Figura 2.2 – Tipagem e Compatibilidade Sanguínea. 29
Figura 2.3 – Etapas para doação de sangue. 31
Figura 4.1 – Exemplo de árvore de decisão 45
Figura 4.2 – Hiperplano de separação SVM de maior margem. 49
Figura 4.3 – Mapeamento de um conjunto de dados não linearmente separável em um linearmente separável. 50
Figura 4.4 – Rede MLP com uma camada oculta. 52
Figura 4.5 – Estrutura usual de uma rede RBF. 57
Figura 5.1 – Distribuição de doadores. 63
Figura 5.2 – Distribuição das doações na campanha de Março de 2007. 63
Figura 5.3 – Fluxograma da abordagem de recomendação dos top-k proposta. 67
Figura 5.4 – Ilustação da heurística baseada nos vizinhos mais próximos. 68

LISTA DE TABELAS

Tabela 4.1 – Exemplos de treinamento para o problema jogar tênis 44
Tabela 5.1 – Classificação RVD 62
Tabela 5.2 – Resultados obtidos pelos classificadores experimentados. 66
Tabela 5.3 – Resultados por classificadores *Top-k Recomendation* ($k = 10$). 70

LISTA DE ABREVIATURAS E SIGLAS

ANVISA — Agência Nacional de Vigilância Sanitária

CART — Classification and Regression Trees

FN — False Negative

FP — False Positive

HEMOCAT — Hemocentro Regional de Catalão

IG — Information Gain

IVD — Iregular Volunteer Donor (Doador Voluntário irregular)

kNN — k-Nearest Neighbor

LR — Logistic Regression

MLP — Multi-Layer Perceptron

NB — Naive Bayes

NVD — New Volunteer Donor (Novo Doador Voluntário)

OMS — Organização Mundial da Saúde

RBF — Radial-Basis Function Networks

RVD — Regular Volunteer Donor (Doador Voluntário Regular)

SMOTE — Synthetic Minority Over-sampling Technique

SVM — Support Vector Machine

TF-IDF — Term Frequency–Inverse Document frequency

TN — True negative

TP — True positive

UH — Unidade Hemoterápica

SUMÁRIO

1 INTRODUÇÃO 23
1.1 Considerações Iniciais 23
1.2 Justificativa 24
1.3 Objetivos 25
1.4 Principais Resultados e Contribuições 25
1.4.1 Análise de classificadores para predição de doação de sangue 25
1.4.2 Desenvolvimento e análise de métodos de recomendação dos top-k items 26
1.5 Organização do trabalho 26

2 DOAÇÃO DE SANGUE 27
2.1 Considerações Iniciais 27
2.2 Grupos Sanguíneos e Compatibilidade 27
2.3 Processo de Doação 29
2.4 Fatores de Percepção e Motivação do Doador 33

3 SISTEMAS DE RECOMENDAÇÃO 35
3.1 Considerações Iniciais 35
3.2 Técnicas Clássicas de Sistemas de Recomendação 36
3.2.1 Filtragem Baseada em Conteúdo 36
3.2.2 Filtragem Colaborativa 37
3.2.2.1 Vizinhos mais próximos baseados em usuário 38
3.2.2.2 Vizinhos mais próximos baseados em item 38
3.2.3 Filtragem Baseada em conhecimento 38
3.2.3.1 Recomendação baseada em restrições 38
3.2.3.2 Recomendação baseada em casos 38
3.2.4 Abordagens Híbridas 39
3.3 Sistemas de Recomendação e Aplicações na Área da Saúde 40

4 CLASSIFICADORES 43
4.1 Considerações Iniciais 43
4.2 Árvores de Decisão 44
4.2.1 ID3 *(Iterative Dichotomiser 3)* 46
4.2.2 C4.5 46

4.2.3 CART . . . 47
4.3 *Naive Bayes* . . . 48
4.4 *Support Vector Machines* . . . 49
4.5 Classificador dos Vizinhos mais Próximos (*k-Nearest Neighbor*) . . . 50
4.6 *Multilayer Perceptron* . . . 51
4.7 Classificador *Radial-Basis Function Network* . . . 53
4.8 Técnicas de amostragem de dados . . . 58
4.9 Balanceamento de classes . . . 58
4.10 Métodos de estimativa da probabilidade de classe . . . 59

5 DESENVOLVIMENTO E EXPERIMENTAÇÃO . . . 61
5.1 Considerações Iniciais . . . 61
5.2 Base de dados . . . 61
5.2.1 Balanceamento da Base de dados . . . 64
5.3 Metodologia de avaliação de modelos de classificação . . . 64
5.4 Resultados dos algoritmos classificadores . . . 65
5.5 Recomendação dos Top-k . . . 66
5.5.1 Heurística baseada nos vizinhos mais próximos . . . 67
5.5.2 Heurística baseada no Teorema de Bayes . . . 68
5.5.3 Heurística baseada no hiperplano de separação das classes . . . 69
5.6 Resultados de recomendação dos top-k . . . 70
5.7 Análise e discussão . . . 70

6 CONCLUSÕES . . . 73
6.1 Considerações Iniciais . . . 73
6.2 Principais Contribuições . . . 73
6.3 Propostas de Trabalhos Futuros . . . 74

REFERÊNCIAS . . . 75

Capítulo 1

INTRODUÇÃO

1.1 Considerações Iniciais

A transfusão de sangue sempre foi uma questão de preocupação mundial por não haver qualquer substância capaz de substituir o sangue humano. Por isso as doações de sangue são de suma importância para suprir as necessidades de pacientes em casos de urgências (acidentes e enfermidades graves), bem como doenças crônicas que exijam transfusões regulares, como anemias e deficiências na coagulação (AGUIAR *et al.*, 2016).

A doação de sangue, apesar de ser um ato comum e muito difundido, traz consigo algumas dificuldades em suas abordagens quanto a fomento e manutenção de doadores de sangue. As principais dificuldades são relacionadas a como obter e manter um número suficiente de doadores, e como obter doações de sangue em situações de emergência. Considerando situações de emergência é importante que as unidades hemoterápicas tenham tecnologias de gerenciamento de dados que possibilitem indicar as pessoas cadastradas que tem as maiores probabilidades de doar sangue em resposta a uma requisição. As aplicações dos métodos computacionais desenvolvidos nesta pesquisa estão mais relacionadas a este último aspecto (SILVA; VALADARES, 2015).

Segundo a Organização Mundial da Saúde (OMS), a qualidade do sangue está ligada a regularidade de doações, tanto pelo altruísmo intrínseco em doadores regulares quanto a maior quantidade de triagens clínicas para garantimento da aptidão, as quais são realizadas em toda doação, esta prerrogativa pode ser observada no estudo de Z., M.R. e M. (2006) sobre segurança do sangue e frequência da doação. Atualmente o Brasil conta com apenas 1.9% de toda sua população como doadores regulares, o que dificulta a manutenção dos bancos de sangue em todo país. Para que seja possível melhorar tais índices diversas pesquisas são realizadas como as de Guiddi *et al.* (2015) e Gillespie e Hillyer (2002), com a finalidade de reverter este quadro e manter o volume de doações sugeridos pela OMS que é de 3% de doadores regulares em um país (World Health Organization, 2010).

O aprendizado de máquina e os sistemas de recomendação são ferramentas utilizadas para melhorar os índices nas buscas pelos itens mais relevantes em grandes quantidades de dados, podendo estes serem amplamente aplicados a área da saúde, incluindo as instituições e atividades hemoterápicas. Estas ferramentas podem ser aplicadas aos bancos de dados de hemocentros em substituição as buscas exaustivas por doadores que são comumente utilizadas.

A predição de doadores é objeto de estudo nos artigos Santhanam e Sundaram (2010) e Darwiche *et al.* (2010) e mostram um excelente ganho de eficiência em seus resultados computacionais. Porém, faz-se necessário observar uma gama ainda maior de métodos para que a literatura avance no auxilio da predição dos doadores de sangue com maior probabilidade de conversão e assim contribuir ao aumento dos índices de doações regulares.

1.2 Justificativa

Foi realizado um estudo de caso no Hemocentro Regional de Catalão - HEMOCAT, estado de Goiás, com o intuito de ampliar o conhecimento sobre os procedimentos e atual situação dos processos de fomento a doação de sangue. Todos os passos desde a busca por informações acerca da ação, coleta de sangue, até a entrega dos resultados dos exames foram analisados. Um dos principais apontamentos diante das observações realizadas é de que existe uma baixa eficiência dos sistemas de gestão no que diz respeito à conversão das buscas por doadores em efetiva doação. As ferramentas utilizadas pelo hemocentro para contactar doadores desfavorecem encontrar os melhores resultados, ter dados atualizados e acesso simplificado a essas informações.

Do ponto de visto científico, este trabalho é justificado pela falta de uma análise abrangente de métodos existentes na literatura para a predição de doação de sangue, assim como pela necessidade de desenvolver novos métodos aplicáveis para a problema de predição em questão. Foi observado que na literatura a maioria dos estudos voltados para predição de doares de sangue elegem apenas um ou poucos classificadores como é o caso de Santhanam e Sundaram (2010) que avalia apenas um algoritmo classificador, sendo de grande relevância uma avaliação mais ampla de métodos clássicos da literatura. Unido a isso foi identificado em diversos estudos como os de Deshpande e Karypis (2004), Karypis (2001) e Lee *et al.* (2016) a utilização da avaliação probabilística dos resultados obtidos pelos classificadores item a item. Com isso é possível ordena-los de acordo com seu grau de confiança, porém não foram encontrados estudos de sua aplicação a predição de doadores de sangue.

1.3 Objetivos

O objetivo deste trabalho é investigar e desenvolver técnicas de predição para encontrar os doadores com maior probabilidade de conversão a doação analisando características de uma base de dados de uma Unidade Hemoterápica (UH). Para isso foram inicialmente avaliados métodos de classificação, os quais obtiveram resultados insatisfatórios. A partir deste resultado, buscou-se criar técnicas de predição mais aprimoradas, o que levou ao desenvolvimento de técnicas de recomendação dos top-k, as quais utilizam uma heurística para estimar, com base nas características de um indivíduo, um grau de confiança de que este irá doar sangue em resposta a uma solicitação.

1.4 Principais Resultados e Contribuições

Este estudo tem grande relevância não somente a área da saúde onde pode ser aplicado, mas também para que faça parte da literatura e assim garantir maior evolução de pesquisas voltadas a classificação, mineração e predição de dados atingindo também outras áreas. A utilização de uma maior gama de classificadores bem como de métodos para otimiza-los torna a visão mais ampla o que auxilia na tomada de decisão e amplia as formas de solucionar o problema.

1.4.1 Análise de classificadores para predição de doação de sangue

Inicialmente o problema de predição de doação de sangue foi modelado em termos de uma tarefa de classificação, onde dado um potencial doador, deseja-se prever se este doará sangue ou não em uma data (campanha) específica. Para esta tarefa, analisamos o desempenho dos principais classificadores clássicos da literatura, a saber, Naive Bayes (NB), *Support Vector Machine* (SVM), *Multi-Layer Perceptron* (MLP), *Radial Basis Function Network* (RBFNet), *k-Nearest Neighbor* (kNN), e as árvores de decisão C4.5 e CART. Os resultados desta comparação mostram uma grande variação quando se diz respeito as métricas utilizadas, as quais foram *accuracy*, *precision* e *recall*. Houveram picos nos resultados individuais para cada métrica. Em termos de *accuracy* o melhor classificador foi o kNN com k=19 (78,73%), em *precision* C4.5 (82,14%) e em *recall* o melhor foi o SVM (96,11%), porém o classificador que teve o mais alto índice de desempenho pela média de todas as métricas foi o CART. Em resumo, foi concluído que nenhum dos classificadores tiveram desempenho acima da média por não haver linearidade dos resultados. Adicionalmente, como a base de dados utilizada é altamente desbalanceada, verifica-se que um classificador que sempre prediz em favor da classe majoritária, reteria desempenho não muito distante do melhor dos classificadores analisados. Para corrigir o problema do desbalanceamento entre as classes, foi aplicada uma abordagem de super-amostragem da classe minoritária denominada *Synthetic Minority Over-sampling Technique* (SMOTE), contudo, os resultados continuaram

aquém do desejável para aplicações práticas, tendo para o melhor classificador neste caso o algoritmo CART, que com base na média geral das métricas obteve, 73,80%, 74,17% e 73,84% de *accuracy*, *precision* e *recall*, respectivamente.

1.4.2 Desenvolvimento e análise de métodos de recomendação dos top-k items

Dado o desempenho relativamente insatisfatório de classificadores para a tarefa de predição de doação de sangue, foi proposta uma abordagem de recomendação que calcula uma espécie de grau de confiança ou probabilidade, de que um determinado indivíduo irá doar sangue. Com base neste valor de confiança (ou probabilidade), recomendamos os top-k itens. Conforme esta abordagem, neste trabalho foram exploradas três heurísticas para estimativa de grau de confiança, sendo: heurística baseada nos vizinhos mais próximos, heurística baseada no Teorema de Bayes e heurística baseada em hiperplanos de separação. Os resultados obtidos mostram que qualquer uma destas heurísticas produzem melhor avaliação do que a analise apenas dos classificadores. Além disso a heurística baseada em hiperplanos de separação produziu resultados altamente precisos, com valores de *precision* acima de 96% para todas as top-5 recomendações e acima de 84% para top-10. Com estes resultados, acredita-se que a abordagem desenvolvida pode agregar à literatura e também ao desenvolvimento de ferramentas para auxilio às atividades hemoterápicas na busca de doadores, assim como no aumento da regularidade das doações.

1.5 Organização do trabalho

O restante deste documento é organizado da seguinte forma: Nos Capítulos 2, 3, 4 são apresentadas revisões bibliográficas abordando respectivamente os conceitos sobre doação de sangue, métodos de sistemas de recomendação e algoritmos classificadores. O Capítulo 5 primeiramente modela o problema de recomendação de doadores como um problema de classificação e apresenta os resultados obtidos pelos vários classificadores experimentados; em uma segunda etapa são apresentados e avaliados os métodos de recomendação top-k propostos, os quais obtiveram resultados bastante superiores aos classificadores. Por fim, no Capítulo 6, são apresentadas as conclusões obtidas pela pesquisa e são discutidas propostas de trabalhos futuros.

Capítulo 2

DOAÇÃO DE SANGUE

2.1 Considerações Iniciais

Para que os objetivos deste trabalho sejam atingidos, faz-se necessária uma analise detalhada sobre a doação de sangue, portanto, neste capítulo são abordados conceitos e especificidades advindos da literatura sobre a doação de sangue com seus processos e os fatores de motivação para o doador.

O sangue humano é um tecido vivo que circula pelo corpo, levando oxigênio e nutrientes a todos os órgãos (OLIVEIRA, 2016). Atualmente não existe nada que substitua a função do sangue no corpo humano e como este não é produzido artificialmente, as pessoas que necessitam, contam com a doação e solidariedade de outras pessoas (RODRIGUES; REIBNITZ, 2011).

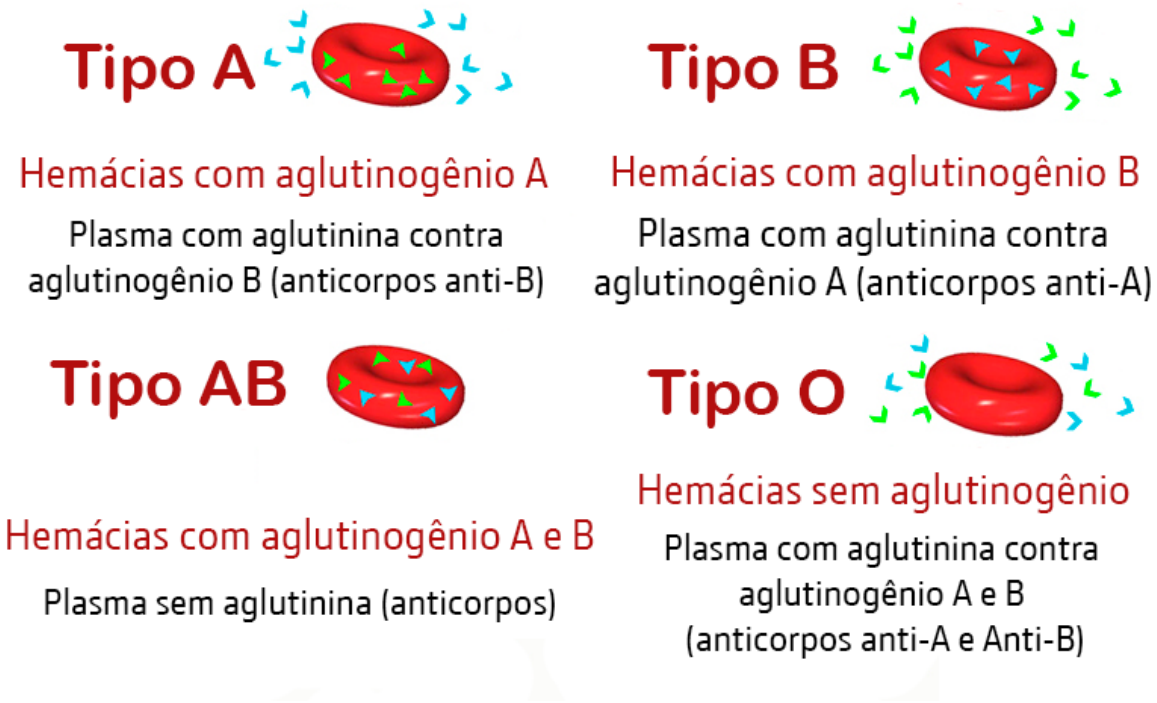
Segundo World Health Organization (2010), a capacidade de um país para atender consistentemente às demandas clínicas de sangue e seus derivados, depende da doação de sangue por um número suficiente de indivíduos saudáveis. Os doadores voluntários bem informados que estão empenhados em doar sangue regularmente formam a base de um grupo sustentável de doadores de sangue que pode fornecer um suprimento de sangue seguro, confiável e suficiente em todas as situações e em todos os momentos.

2.2 Grupos Sanguíneos e Compatibilidade

O sangue é classificado nos grupos (ou tipos) A, B, AB e O, e ainda em subgrupos conforme o "Rh", que pode ser positivo ou negativo. Todos os tipos sanguíneos podem ter o "Rh"positivo ou "Rh"negativo (SUN; LU; JIN, 2016). A classificação em grupos depende do tipo de proteína presente no sangue, conforme ilustrado na Figura 2.1. Para cada tipo sanguíneo existe uma proteína específica, sendo que apenas o sangue tipo O não apresenta tal proteína. Já a classificação em subgrupos está relacionada aos diferentes componentes

(antígenos) que podem haver no plasma – componente líquido do sangue no qual os demais componentes e as proteínas que alimentam as células ficam suspensos. Em caso de haver os antígenos o fator é considerado positivo "Rh+"e na sua ausência, negativo "Rh-" (GARRAUD; TISSOT, 2016). Na Figura 2.1 é possível verificar como esses componentes, chamados de aglutininas[1] interagem com os grupos sanguíneos.

Figura 2.1 – Classificação dos Grupos e Subgrupos Sanguíneos.

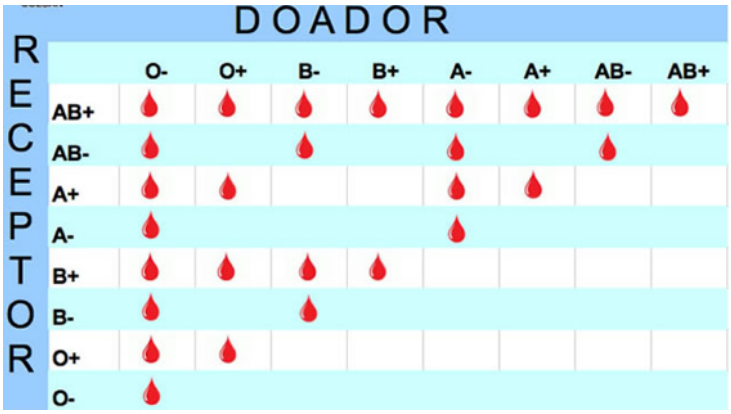


Fonte: Expedição e Vida (2013).

Durante uma transfusão de sangue é necessário que haja compatibilidade sanguínea, ou seja, que o sangue doado possua aglutinogênios compatíveis com as aglutininas do receptor. Caso o sangue doado não seja compatível com o do receptor, o sangue coagulará formando uma massa sólida que obstrui os capilares sanguíneos e prejudica a circulação do sangue, podendo levar o receptor à morte. Como o grupo sanguíneo tipo “O” não possui aglutinogênio ele é considerado doador universal, bem como o tipo “AB” é considerado receptor universal, pois não possui nenhuma aglutinina. Em relação ao Fator Rh, os Rh+ podem receber tanto sangue com o mesmo tipo, como sangue com Rh negativo (GARRAUD; TISSOT, 2016). A Figura 2.2 ilustra a compatibilidade sanguínea, considerando os grupos e subgrupos.

[1] Aglutininas: anticorpos do plasma sanguíneo específicos contra determinados aglutinogênios. Aglutinogênios: glicoproteína existentes nas hemácias sanguíneas.

Figura 2.2 – Tipagem e Compatibilidade Sanguínea.



Fonte: Colsan (2016).

O recomendado é que o paciente receba sangue do mesmo tipo que o seu. Apenas em urgências, pelo fato de que muitas vezes não há tempo para tipificar o sangue do paciente, é que se usa o sangue universal, que é o sangue tipo O- (RODRIGUES; REIBNITZ, 2011).

2.3 Processo de Doação

A lei brasileira da Constituição Federal nº 10.205, de 21 de março de 2001, regulamenta os processos relativos à coleta, processamento, estocagem, distribuição e transfusão do sangue, seus componentes e derivados, e ainda estabelece o ordenamento institucional indispensável à execução adequada dessas atividades. Atualmente a Agência Nacional de Vigilância Sanitária (ANVISA) é quem atua como órgão fiscalizador das doações de sangue no país (BRASIL, 2001).

Os órgãos, institutos e organizações de saúde responsáveis pela gestão de hemocomponentes no Brasil, denominadas Unidades Hemoterápicas (UHs), possuem 9 diferentes denominações definidas quanto a abrangência e atividades especificas como coleta, processamento, estocagem, distribuição e aplicação dos hemocomponentes segundo BRASIL (1995). Essas definições são:

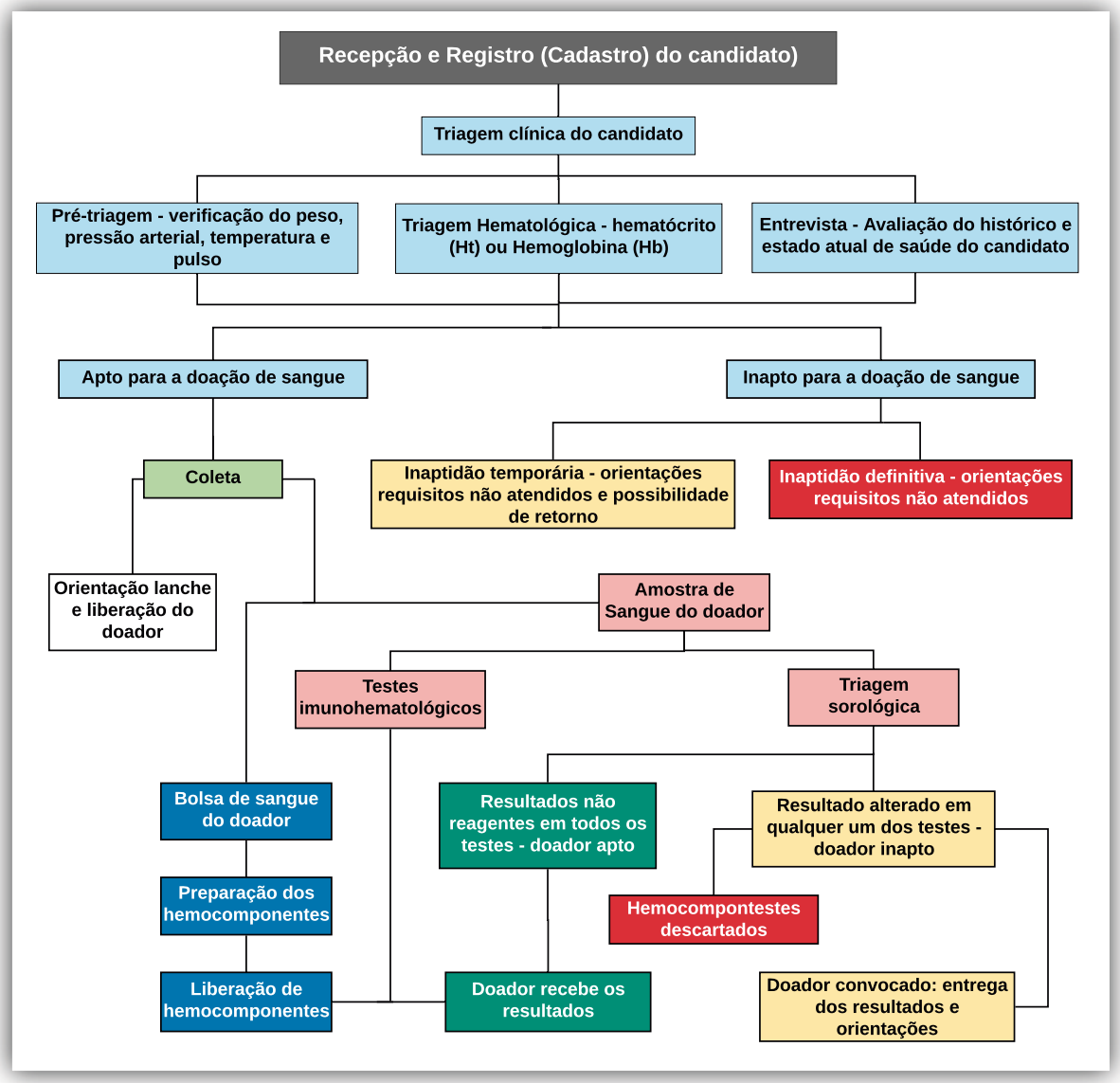
- **HEMOCENTRO**: estrutura centralizada, preferencialmente localizada na capital prestando assistência e apoio hemoterápico à rede de serviço de saúde com maior complexidade e tecnologia do que as demais. Os hemocentros prestam serviços de ensino e pesquisa, de controle de qualidade, suporte técnico, formação e integração de instituições públicas e filantrópicas. Eles também promovem junto com as Secretarias Estaduais de Saúde e através da ANVISA, mecanismos para desenvolver as atividades relativas a gestão de hemocomponentes de forma adequada e segura.

- **HEMOCENTRO REGIONAL:** estrutura de complexidade intermediária com atuação macro-regional em área hematológica/hemoterápica, apoiando e assistindo à rede de serviços de saúde. Eles coordenam e desenvolvem ações na macro-região em descentralização ao Hemocentro.
- **HEMONÚCLEO:** caracteriza-se pela descentralização do Hemocentro Regional, sendo sua localização preferencialmente extra-hospitalar. Estes prestam assistência hemoterápica e/ou hematológica, em nível local.
- **UNIDADES SOROLÓGICAS:** são laboratórios públicos ou privados que desenvolvem o controle sorológico do sangue a ser transfundido apoiando as entidades de assistência que necessitam de diagnóstico sorológico.
- **SERVIÇO DE HEMOTERAPIA:** preferencialmente intra-hospitalar, público ou privado, tem a função de prestar assistência hemoterápica/hematológica, o qual recruta doadores, processa o sangue, realiza os testes necessários, armazena e o prepara para transfusão. Este distribui o sangue, exclusivamente para apenas um hospital, podendo ou não prestar atendimento ambulatorial.
- **UNIDADE DE COLETA E TRANSFUSÃO:** unidade de atendimento de coleta e transfusão localizada em hospitais isolados ou pequenos municípios, onde a demanda não justifique a instalação de uma estrutura mais complexa. Esta UH envia o sangue para ser processado em outra unidade de maior estrutura.
- **AGÊNCIA TRANSFUSIONAL:** localização obrigatoriamente intra-hospitalar, com a função hemoterápica. O suprimento de sangue a essas agências é realizado através das unidades de maior estrutura.
- **POSTO DE COLETA:** unidade especial para coleta de sangue, podendo ser intra ou extra-hospitalar, fixo ou móvel enviando o produto para outra unidade de maior complexidade para etapas pós coleta.
- **SERVIÇO DE HEMOTERAPIA DISTRIBUIDOR:** com características similares ao **SERVIÇO DE HEMOTERAPIA**, com exceção da preferencia em ser extra-hospitalar e realizar distribuição para mais de um hospital.

As doações são realizadas em UHs, sendo que o candidato deve comparecer levando documento de identificação para que seja feito o cadastro ou confirmação dos dados, para ser encaminhado para as fases de triagem clínica como ilustrado na Figura 2.3. Na fase de pré-triagem, mede-se a pressão arterial, a temperatura, o peso e o pulso. Logo após candidato é submetido a triagem hematológica para coleta de sangue, afim de realizar um exame denominado hematócrito, que determina a quantidade total de hemácias no sangue. Na fase seguinte é realizada a entrevista, cuja finalidade é avaliar os antecedentes patológicos

e os possíveis fatores de risco daquele candidato à doação. Com esses dados é possível saber se a pessoa pode doar sangue naquele momento (RAZOUK; REICHE, 2004; GARRAUD; TISSOT, 2016)

Figura 2.3 – Etapas para doação de sangue.



Fonte: (PINHO *et al.*, 2001).

Com a entrevista realizada, o candidato apto segue para a sala de coleta de sangue, onde além do sangue que irá para a bolsa é também coletado o sangue para realizar os primeiros exames de verificação. Nesta etapa são feitos testes para doenças infecciosas como a *Acquired Immunodeficiency Syndrome* (AIDS), hepatites C e B, doença de Chagas e para o *Human T-cell Lymphotropic Virus* (HTLV) [2].

Existem diferentes processos de doação de sangue que podem ser divididas com relação a quantidade de hemocomponentes coletados: a Doação de Sangue Total Voluntária

[2] O Vírus T-linfotrópico humano (HTLV) atinge os linfócitos que são as células de defesa do organismo

e Doação por Aférese Voluntária. Os três processos são descritos a seguir (VANE; GANEM, 2006):

- **Doação de Sangue Total:** é a forma mais comum, com a coleta de até 450ml de sangue com todos seus hemocomponentes e colocado em uma bolsa com materiais e soluções para conservá-lo.
- **Doação por Aférese:** a doação por aférese concentra-se em retirar apenas as plaquetas do sangue, representando uma quantidade até 7 vezes maior do que a doação de sangue total. A reposição das plaquetas é feita naturalmente dentro de 48hs pelo corpo sem afetar a saúde do individuo que poderá doar novamente dentro de 72h.

Outras subdivisões podem ser feitas segundo a intenção e/ou destinação dos hemocomponentes, sendo eles a doação espontânea, de reposição, por convocação e autóloga, descritos nos tópicos seguintes:

- **Doação espontânea:** é a doação voluntária feita por doadores motivados e com finalidade altruísta para manter o estoque da UH em níveis adequados (PINHO *et al.*, 2001).
- **Doação de reposição:** tem interesse em atender uma necessidade específica de um paciente. No ato da coleta o doador informa os dados do paciente e uma bolsa de sangue é encaminhada. Por conta do tempo necessário para testar os componentes sanguíneos, geralmente uma outra coleta é encaminhada, e após os testes da doação destinada ao paciente, esta repõem a que foi de fato utilizada (PINHO *et al.*, 2001).
- **Doação por convocação:** estas são realizadas por doadores registrados na UH e receberam chamados por estas unidades através de seus meios de comunicação (PINHO *et al.*, 2001).
- **Doação autóloga:** trata-se da doação para si mesmo, como em caso de cirurgias planejadas onde serão necessárias transfusões, é possível realizar coleta e armazenamento do sangue do próprio paciente em quantidade necessária para o procedimento. Esta prática é muito utilizada por não haver riscos de contaminação do paciente mas necessita de prescrição médica (PINHO *et al.*, 2001).

Os impedimentos para doação de sangue podem ser temporários, como em casos de gravidez, resfriado, tatuagem nos últimos doze meses, amamentação, peso abaixo de 50Kg entre outros, implicando que a inaptidão pode revertida em determinado tempo ou ação do candidato. A inaptidão pode ser também definitiva como nos casos em que o candidato é portador de doenças infecciosas crônicas, malária, tenha mais de 69 anos, apresentou hepatite, entre outros (BRASIL, 2013).

Para realizar a coleta das informações necessárias para indicar a aptidão ou inaptidão de um candidato a doação, a entrevista mostra-se de grande importância para o processo. Os Anexos na portaria BRASIL (2011) destacam as tabelas de triagem clínica para doenças, cirurgias, procedimentos invasivos e uso de medicamentos que definem se um candidato é inapto definitivo ou temporário. Essas informações podem apenas ser coletadas através de entrevista com a análise do histórico médico do doador.

O intervalo mínimo para doação de sangue é de dois meses para homens e de três meses para mulheres, com a exceção das mulheres entre 60 e 69 anos, onde o prazo é de seis meses (LYLE; SMITH; SULLIVAN, 2009). Os impedimentos relativos a idade foram alterados em 2013 pelo Ministério da Saúde, através da publicação da Portaria 2.712 de 12 de novembro, que reduziu a idade mínima para doação de 18 para 16 anos e aumentou a idade máxima de 67 para 69 anos; com isso houve a inclusão de 8,7 milhões de pessoas nos cadastros para novos doadores voluntários de sangue no Brasil (BRASIL, 2013).

No Brasil, a logística das bolsas de sangue entre as diversas UHs públicas e contratadas pelo Sistema Único de Saúde (SUS), acontece através das hemorredes que podem ser definidas como o conjunto das unidades públicas que atuam na área de sangue e hemoderivados, visando atender a demanda de sangue. Assim, dentro de um mesmo estado brasileiro, é possível transferir bolsas de sangue com maior facilidade entre cidades que pertencem a mesma hemorrede. Apenas em casos excepcionais é que bolsas de sangue são transferidas entre hemorredes diferentes (PORTAL BRASIL, 2016).

2.4 Fatores de Percepção e Motivação do Doador

Segundo o estudo realizado por Gillespie e Hillyer (2002) a motivação primária para doação de sangue é principalmente devido ao comportamento pró-social, ideologia da qual se infunde que não há vantagens óbvias para o doador, contudo, há benéfico ao receptor. Os autores apresentam ainda que outros motivos para a doação estão ligados a responsabilidade social, comportamento afetivo e psicológico, e ressaltam que é na realidade um conjunto desses e outros fatores que levam a pessoa a doar. Silva e Valadares (2015) aponta ainda haver uma transformação de ideologias, em que a motivação do indivíduo deve ser a vontade de agir em prol do outro. Outros fatores que levam o cidadão a doar são: pessoas que já precisaram de sangue ou que acreditam que um dia possam precisar, outras acreditam que os testes médicos realizados são um benefício ao doador, tem-se também os que são motivados devido a estarem envolvidos com grupos de amizade que participam de campanhas de doação (GILLESPIE; HILLYER, 2002).

Por outro lado, fatores como medo de agulha, de hematomas, de ver sangue, dor, desconforto, ansiedade, falta de consciência, apatia (indiferença perante a questão) e não adesão ao processo de doação são fatores que desmotivam potenciais doadores. Outra dificul-

dade é quando um doador é acometido por situações que lhe impedem temporariamente, ou mesmo definitivamente de doar sangue (RINGWALD; ZIMMERMANN; ECKSTEIN, 2010).

Segundo o Ministério da Saúde (2016), 1.9% da população brasileira doa sangue regularmente. Apesar da taxa está dentro dos parâmetros mundiais, que é entre 1% e 3%, ela pode e precisa melhorar. Os números de doação no Brasil refletem a necessidade de novas ferramentas para auxiliar o processo de doação de sangue.

Capítulo 3

SISTEMAS DE RECOMENDAÇÃO

3.1 Considerações Iniciais

Atualmente, a Internet dispõe de um grande fluxo de informações que podem ser utilizadas para diversos fins. No entanto, essas informações nem sempre estão organizadas de forma viável para seu uso, sendo necessário a aplicação de métodos que permitam filtrar as informações relevantes para um interesse específico. Nesse contexto, o aprimoramento de consultas por similaridade e os sistemas de recomendação têm se destacado (SHARDANAND; MAES, 1995).

Goldberg *et al.* (1992) que são os autores do primeiro sistema de recomendação que se conhece, o *Tapestry*, fizeram também as primeiras menções ao termo "Filtragem Colaborativa", frequentemente utilizado para designar os sistemas de recomendação, sendo a abordagem mais utilizada. O *Tapestry* era baseando em um sistema de email utilizando técnicas de sistema de recomendação para que, com base nas ações de seus usuários fosse filtrada a informação com mais eficiência. As técnicas utilizadas no *Tapestry* incluem filtragem baseada em conteúdo e filtragem colaborativa.

Os sistemas de recomendação têm um papel crucial na escolha de alternativas possíveis, pois muitas pessoas podem não ter conhecimento ou experiência para escolher dentre uma grande quantidade de opções. Além disso, a quantidade de dados podem ser muito grande para uma análise manual. Adicionalmente, mesmo que a quantidade de dados seja gerenciável manualmente, as decisões sobre os dados podem ser complexas, assim como pode ser altamente indesejável um procedimento de tentativa e erro. Estes sistemas aumentam a eficácia e a capacidade no processo de indicação que ocorre de modo mais ameno e imparcial do que em relações sociais humanas (RESNICK; VARIAN, 1997).

3.2 Técnicas Clássicas de Sistemas de Recomendação

As várias técnicas para produzir recomendações são classificadas com base no tipo de informações adquiridas dos usuários, dos itens e na estrutura do sistema em si. Uma forma clássica de distinguir as diferentes técnicas é feita por Burke (2007), sendo divididas em seis classes de técnicas: Filtragem Baseada em Conteúdo, Filtragem Colaborativa, Filtragem Baseada em Conhecimento, Filtragem Demográfica, Filtragem Baseada em Comunidade e Abordagens Híbridas:

3.2.1 Filtragem Baseada em Conteúdo

A abordagem baseada em conteúdo tem por objetivo gerar automaticamente descrições dos conteúdos de itens e comparar estas descrições com os interesses dos usuários, visando verificar se o item é ou não relevante para cada indivíduo (HERLOCKER; KONSTAN; RIEDL, 2000). Seu principal foco é recomendar itens contendo informações textuais como sites, blogs, artigos científicos dentre outros, mas pode ser empregado em outros tipos de dados, como fotos, vídeos, músicas (áudio), etc.

O sistema utiliza de um histórico das ações do usuário, recomendando itens semelhantes aos que ele gostou anteriormente. A similaridade entre os itens é calculada com base em suas características em comum. Considerando, por exemplo, a recomendação de filmes, caso o usuário tenha gostado de um filme de ação o sistema pode recomendar outros de mesmo gênero.

Adicionalmente, este sistemas podem extrair características intrínsecas dos dados de forma a, por exemplo, estimar se um filme é de ação ou não, com base em características computadas da sequência de *frames* (imagens), que compõem o filme. Procedimentos similares para estimar similaridade de itens podem ser feitos para áudios, textos, imagens, grupos em redes sociais, dentre outros itens alvo de recomendações.

Os dados utilizados podem ser estruturados, tais como datas de publicação e categorias de filmes, entretanto, é comum a utilização de formatos semi-estruturados de dados, tais como textos de sinopses de filme, descrições de itens, entre outros, que implicam em uma nova forma de representar os itens, o que é fundamental na construção de um recomendador baseado em conteúdo.

O processo de construção de qualquer das representações dos itens pode ser dividido em etapas, a primeira denominada extração de características. Aqui o foco será na descrição de itens representados na forma textual, pois são mais correlacionados com o problema desta pesquisa. Neste caso, um texto será caracterizado por um conjunto de atributos que sumariza o seu conteúdo, sendo necessárias etapas para se obter tal caracterização. Primeiramente é comum relizar uma etapa de pré-processamento que tem forte dependência do idioma em que foi escrito o texto, sendo removidas palavras bastante comuns a todos os

textos, denominadas de *stopwords* e faz-se uma coleta dos radicais das palavras que não são *stopwords*. Tal processo de se obter os radicais de palavras é chamado de *stemming* (WANG; WANG, 2005). Na segunda etapa os atributos (termos) selecionados para caracterizar os textos são mapeados em um vetor de características. Associado a cada atributo (termo) normalmente se estabelece um peso e alguns dos mais comuns são (SEBASTIANI, 2002):

- *Term Frequency–Inverse Document frequency* (TF-IDF) – estima o quão é importante um termo de um documento em uma dada coleção. O valor TF-IDF aumenta proporcionalmente ao número de vezes que um termo aparece no documento, mas é compensado pela frequência do termo na coleção, o que ajuda a ajustar o fato de que alguns termos aparecem frequentemente ao longo de toda a coleção.

- *Information Gain* (IG) – estima o ganho de informação ao adicionar o termo em questão para predição de categoria do item.

3.2.2 Filtragem Colaborativa

Este tipo de filtragem recomenda itens ao usuário com base em seu histórico, como na filtragem baseada em conteúdo, porém, cruza-se os perfis de usuários, fazendo com que a escolha seja com base em usuários de gostos semelhantes. Tal abordagem é referida por Schafer, Konstan e Riedl (2001) como "correlação pessoa-a-pessoa"e é considerada a técnica mais popular e amplamente utilizada principalmente em *e-commerces*.

A obtenção das preferencias do usuário por um item específico podem ser determinadas de forma explícita e implícita. Na forma explicita o usuário é questionado ou preenche um formulário contendo uma nota (*rating*) para cada item, o que exige mais esforço ao usuário. Já na forma implícita os dados são coletados pela observação do seu comportamento, como a compra de um determinado item ou a frequência em que o visita, podendo assim o sistema assumir uma avaliação positiva ao item. Essa forma de obtenção das preferências (perfil) do usuário pode não ser totalmente confiável por haver garantia que as ações foram interpretadas corretamente, exigindo estratégias mais avançadas para analisar as preferências.

O grande volume de dados em sistemas de recomendação é de relevante preocupação uma vez que as matrizes de avaliação são muito esparsas contendo um grande número de usuários e itens, e poucos itens avaliados pelos usuários. Outro problema dos sistemas de filtragem colaborativa é o *cold-start*, que se refere ao surgimento de novos usuários sem que haja informação suficiente sobre estes para encontrar perfis parecidos e realizar as recomendações, aumentando a imprecisão das recomendações (SU; KHOSHGOFTAAR, 2009).

3.2.2.1 Vizinhos mais próximos baseados em usuário

A abordagem baseia-se no conjunto de avaliações de um dado usuário para encontrar outros usuários que apresentam maior similaridade de preferencias, sendo eles os vizinhos mais próximos. Com base nesses usuários vizinhos o sistema produz recomendações dos itens que eles gostaram e que usuário atual ainda não avaliou (FELFERING *et al.*, 2011).

Em resumo, esta abordagem assume que se os usuários X e Y avaliam n itens similarmente, irão avaliar outros itens da mesma forma; então se um usuário e seu(s) vizinho(s) mais próximo(s) tiverem preferências similares no passado, futuramente apresentarão preferencias similares em outros itens.

3.2.2.2 Vizinhos mais próximos baseados em item

Nesta abordagem, o objetivo é realizar recomendações com base nos itens dada a similaridade de avaliação pelos usuários. Se um usuário demostrou preferência por alguns itens, o sistema irá recomendar itens semelhantes a estes, que o usuário ainda não avaliou. Em síntese, se um usuário gostou de determinados itens e estes receberam avaliações similares a outro item X, então será provável que o usuário goste também de X.

3.2.3 Filtragem Baseada em conhecimento

Sistemas baseados em conhecimento sugerem itens com base no conhecimento de domínio sobre as características do item, recomendando itens os quais suas características satisfazem, até um determinado grau, as necessidades e preferências do usuário. A função de similaridade, portanto, encarrega-se de estimar o quanto as necessidades do usuário coincidem com os itens. Neste caso, a pontuação gerada pode ser vista com o nível de utilidade do item para determinado usuário. Existem duas técnicas básicas de filtragem baseada em conhecimento: baseada em restrições e baseada em casos.

3.2.3.1 Recomendação baseada em restrições

As tarefas baseadas em restrições são definidas pela tupla *(V,D,C)*, em que V é o conjunto de variáveis, D é o conjunto dos domínios destas variáveis e C é o conjunto de restrições que explora requisitos do usuário sobre as característica dos itens. Tal abordagem pode ser resolvida por algoritmos para problemas de satisfações de restrições tais como busca com retrocesso *(backtracking)*, propagação de restrições e busca local (RUSSELL; NORVIG, 2009).

3.2.3.2 Recomendação baseada em casos

A recuperação dos itens é realizada de acordo com uma medida de similaridade calculada entre as propriedades de um item e os requisitos fornecidos pelo usuário. O sistema

solicita os requisitos ao usuário e realiza uma pesquisa baseada em casos utilizando a medida de similaridade para encontrar os itens com menor distância até as preferências informadas (RICCI; ROKACH; SHAPIRA, 2011; FELFERING *et al.*, 2011).

3.2.4 Abordagens Híbridas

É possível criar uma abordagem que possibilite a combinação das características das abordagens básicas, que são as abordagens híbridas. As técnicas assim denominadas têm diferentes características, e assim, vantagens e desvantagens. Utilizando-se das abordagens híbridas torna-se possível unir os pontos fortes e excluir os pontos fracos. No trabalho de Shih e Liu (2005), por exemplo, é possível ver uma aplicação da abordagem híbrida que através de dados de compras frequentes de consumidores analisa a demanda a fim de predizer, maximizando qualidade dos resultados, um *ranking* de produtos candidatos a um usuário unificando os métodos de Filtragem Colaborativa e Baseada em Conteúdo. Conforme Jannach *et al.* (2010), a literatura evidencia três propostas para abordagens híbridas: a monolítica, a paralela e a *pipeline*.

Na proposta monolítica, é realizado todo o processo em uma única implementação. Em Melville, Mooney e Nagarajan (2002), por exemplo, foi realizada a hibridização monolítica das técnicas de recomendação colaborativa e baseada em conteúdo. A abordagem de filtragem baseada em conteúdo resolve o problema de matrizes de avaliações esparsas, preenchendo-a utilizando os atributos dos usuários e dos itens, e então, a técnica colaborativa realiza recomendações.

A proposta paralela realiza o processamento em implementações distintas gerando resultados independentes, que depois são combinados. Em Bostandjiev, O'Donovan e Höllerer (2012) foi criada uma ferramenta para recomendação de músicas com um painel explicativo, onde o usuário pode ajustar os algoritmos em tempo real. Este sistema realiza a hibridização paralela de técnicas para três diferentes APIs sociais populares: Wikipédia com abordagem baseada em conteúdo, buscando o contexto da música (categorias, estilos, álbuns etc.), Facebook com abordagem colaborativa utilizando dados dos perfis de amigos, e o Twitter com uma abordagem baseada em peritos, que é uma generalização da colaborativa, porém, buscando as contas mais influentes vinculadas às pesquisas realizadas.

Por fim a *pipeline* consiste de fases sequenciais onde um método gera as entradas para o próximo. Através da combinação da técnica colaborativa e da baseada em conteúdo, em Balabanović e Shoham (1997) é construído um sistema de recomendação que, primeiro, gera perfis para os usuários por filtragem baseada em conteúdo e a partir destes são realizadas recomendações utilizando a filtragem colaborativa, analisando a semelhança entre os perfis dos usuários.

3.3 Sistemas de Recomendação e Aplicações na Área da Saúde

A utilização de meios eletrônicos para a prestação de cuidados com a saúde, conhecido como *e-healt*, é uma nova perspectiva tecnológica global para aumentar a eficácia na disseminação de informações e de procedimentos médicos beneficiando diretamente os pacientes (KAUR; GUPTA, 2006; MOGHADDASI *et al.*, 2012). A união das necessidades na área da saúde e de soluções advindas de tecnologias desempenham um importante papel na medicina atual, principalmente quando relacionado a análise informações médicas e sistemas de recomendação. Sistemas de recomendação têm sido usados em vários setores da área médica com diversos objetivos, tais como: aumentar a precisão de tratamentos de pacientes, auxílio ao diagnóstico médico, entre outros (KIM *et al.*, 2009).

Em Pradhan, Gay e Nepal (2014) é realizado um estudo visando a melhoria do processo de correspondência de critérios subjetivos, buscando uma harmonização entre profissionais da saúde na área odontológica e seus pacientes. São aplicados questionários aos clientes, sendo um deles voltado para identificar seu perfil pelo comportamento, atitudes e personalidade, e outro aplicado as impressões que tiveram dos profissionais que os atenderam. A partir dai é feita a recomendação do profissional que mais se harmoniza aos critérios dos pacientes.

A tecnologia agregada ao conhecimento do paciente ou usuário e proximidade às informações referentes a eles tem um grande aliado que são os dispositivos móveis por sua capacidade cognitiva e difusão da utilização mundial. Com isso observa-se na literatura alguns projetos desenvolvidos em plataforma *mobile*. O aplicativo FitYou (WING; YANG, 2014) utiliza das tendências em desenvolvimento *mobile*, serviços baseados em localização e sistemas de recomendação a fim de recomendar práticas para uma vida saudável. A solução está focada nos problemas advindos da obesidade que afetam uma proporção significativa da população mundial, gerando problemas como alta pressão arterial, problemas respiratórios, diabetes e colesterol alto. A principal característica do aplicativo está na funcionalidade de realizar recomendações dinamicamente de acordo com os serviços baseados em localização e saúde do usuário em tempo real, utilizando preferências determinadas a partir da história do usuário e informações de saúde, a partir de um perfil biométrico.

O crescimento da tecnologia *mobile* alcança positivamente em inúmeras áreas da sociedade, sendo a saúde uma delas. O *m-healt*, termo que define os meios eletrônicos de apoio a prestação de cuidados à saúde para a plataforma *mobile*, foi incluido na *Global Strategy for Women's and Children's Health* – Estratégia Global pela Saúde das Mulheres e Crianças, em 2010, apontado com uma das principais ferramentas no desenvolvimento da tecnologia na saúde, difusão de conhecimento à população e acessibilidade, mesmo de países mais pobres (KAY; SANTOS; TAKANE, 2011). Na literatura é possível encontrar diversas aplicações *mobile* focadas na área da saúde como o *framework* desenvolvido por Chomutare,

Arsand e Hartvigsen (2010).

Chomutare, Arsand e Hartvigsen (2010) apresentam um *framework* baseado na junção de ferramentas de rede social, aplicação *mobile* e sistemas de recomendação, para que seja possível encontrar pares de perfis com aspectos de saúde semelhantes, onde alguns desses tenham obtido algum avanço na melhoria da qualidade de vida. O foco do estudo é em pessoas com diabetes. A aplicação também fornece um canal de comunicação possibilitando a troca de experiência entre os usuários. O sistema de recomendação contextualiza os dados pessoais do usuário, dados de saúde e preferências entre os usuários e promove a recomendação dos perfis de saúde que mais se assemelham nesses aspectos.

Ouhbi *et al.* (2015) faz uma revisão de características e funcionalidades de aplicativos para dispositivos móveis (*apps*) gratuitos para auxílio na doação de sangue. O estudo buscou nas maiores lojas de aplicativos existentes, Google Play, Apple Store, Blackberry App World e Windows Mobile App store, sendo identificados 188 aplicativos e selecionados 169, pois, conforme os autores, alguns não puderam ser acessados ou instalados. Foi possível verificar em suas conclusões que na maioria dos casos os aplicativos utilizavam o sistema operacional Android, e apresentavam como funcionalidade principal a busca de doadores.

Capítulo 4

CLASSIFICADORES

4.1 Considerações Iniciais

A classificação é uma das tarefas mais empregadas em mineração de dados e visão computacional. Um sistema de classificação é utilizado para prever a classe de novos exemplos (objetos) baseando-se em suas características. No desenvolvimento de um classificador é criado um modelo computacional com base nas características dos exemplos de treinamento, para prever a classe de novos exemplos. Os dados disponíveis são divididos em dois conjuntos mutuamente exclusivos: um conjunto de treinamento, usado para a criação do modelo de classificação, e um conjunto de teste, usado para estimar a qualidade do modelo. O conjunto de treinamento fica disponível para o classificador, que analisa as relações entre as características e as classes. Os relacionamentos descobertos a partir desses exemplos (modelo), são então utilizados para prever a classe dos exemplos presentes no conjunto de teste, que fica indisponível ao classificador durante a fase de treinamento. Após o classificador prever a classe dos exemplos do conjunto de teste, as classes previstas são então comparadas com as classes reais dos exemplos. Se a classe prevista for igual à real, a previsão foi correta; caso contrário, a previsão foi incorreta. Deste modo, é possível avaliar o desempenho de predição do classificador.

O conhecimento descoberto pelo classificador por meio dos exemplos de treinamento, isto é, o modelo, pode ser representado de várias formas, por exemplo: árvores de decisão (QUINLAN, 1993), redes neurais (HAYKIN, 2009), modelos bayesianos (CONGDON, 2006) e máquinas de vetores de suporte (*Support Vector Machine* – SVM) (SCHÖLKOPF; SMOLA, 2002). Existem também os classificadores que não constroem um modelo para representar o conhecimento descoberto, o que são chamados de classificadores preguiçosos (DUDA; HART; STORK, 2001). O exemplo mais conhecido de classifcador preguiçoso é o *k-Nearest Neighbor* (kNN), também conhecido como classificador dos vizinhos mais próximos, que será apresentado adiante neste texto.

A seguir são apresentados alguns classificadores tradicionais, utilizados nos experi-

mentos deste trabalho.

4.2 Árvores de Decisão

As árvores de decisão classificam padrões com base em uma sequência de testes e decisões. Em geral, uma árvore de decisão representa uma disjunção de conjunções de restrição sobre os valores de característica dos padrões. Cada caminho da raiz da árvore até uma folha corresponde a uma conjunção de testes sobre características, e a árvore como um todo corresponde a uma disjunção destas conjunções. Os padrões são classificados seguindo um caminho na árvore da raiz até uma das folhas, a qual provê a classe do padrão. Cada nó interno da árvore corresponde a um teste sobre alguma característica dos dados, e cada ramo descendente a partir de um nó corresponde a uma possibilidade de valor para a característica testada.

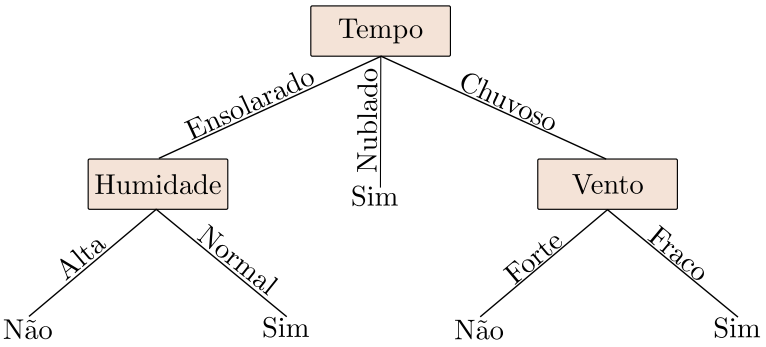
Na Figura 4.1 é fornecido um exemplo de árvore de decisão para o problema "jogar tênis", considerando os dados apresentados na Tabela 4.1. A construção de uma árvore de decisão pode ser vista como um particionamento recursivo do conjunto de dados. No nó raiz todas as instância são consideradas e em cada nó filho considera-se somente o conjunto de dados que satisfaz a condição testada. Este processo é repetido recursivamente até que seja satisfeita uma das seguintes condições de parada: Todos os dados de um mesmo ramo pertencem a uma mesma classe; Não há mais características (testes) a serem adicionadas à árvore; Não há mais dados de treinamento.

Tabela 4.1 – Exemplos de treinamento para o problema "jogar tênis".

Tempo	Temperatura	Humidade	Vento	Jogar Tênis
Ensolarado	Alta	Alta	Fraco	Não
Ensolarado	Alta	Alta	Forte	Não
Nublado	Alta	Alta	Fraco	Sim
Chuvoso	Média	Alta	Fraco	Sim
Chuvoso	Baixa	Normal	Fraco	Sim
Chuvoso	Baixa	Normal	Forte	Não
Nublado	Baixa	Normal	Forte	Sim
Ensolarado	Média	Alta	Fraco	Não
Ensolarado	Baixa	Normal	Fraco	Sim
Chuvoso	Média	Normal	Fraco	Sim
Ensolarado	Média	Normal	Forte	Sim
Nublado	Média	Alta	Forte	Sim
Nublado	Alta	Normal	Fraco	Sim
Chuvoso	Média	Alta	Forte	Não

Fonte: adaptado de (WITTEN; FRANK, 2005).

Figura 4.1 – Árvore de decisão para o exemplo jogar tênis (Tabela 4.1).



Fonte: Autor.

O aspecto mais importante na construção de árvore de decisão é a escolha da característica corrente de teste, que fará a divisão dos dados. O princípio empregado é o de que árvores simples e compactas são preferíveis em relação às complexas. Para este fim, é aplicado um procedimento baseado em um critério de impureza, tal como entropia, que efetua partições resultando em subconjuntos de amostras o mais homogêneas possíveis, em cada ramo da árvore. No decorrer da construção da árvore, uma folha com amostras heterogêneas é substituída por um nó teste que divide o conjunto heterogêneo em subgrupos minimamente heterogêneos, de acordo com o critério de impureza. Em outras palavras, a característica mais informativa em um estágio particular é usada para dividir os dados, pois é a que reduz mais a incerteza.

Como consequência, a operação fundamental de um algoritmo de indução de árvore de decisão é o cálculo de impureza, que determina a divisão a ser realizada em um determinado nó. Existem várias medidas de impureza, todavia, as mais utilizadas são o ganho de informação e a taxa de ganho. Ambas utilizam o conceito de entropia no sentido de teoria da informação, denominado Entropia de Shannon (SHANNON, 1948). Um dado conjunto de padrões $\mathbf{S}$ pode ser descrito em termos de sua distribuição de rótulos de classe, e sua entropia pode ser calculada como:

$$H(\mathbf{S}) = -\sum_{i=1}^{l} P(c_i)\log_2 P(c_i), \tag{4.1}$$

onde $P(c_i)$ corresponde à proporção de padrões em $\mathbf{S}$ pertencente à classe c_i, e l é o número de classes em $\mathbf{S}$.

O ganho de informação $IG(\mathbf{S}, D)$ representa a redução da entropia (incerteza) esperada quando o conjunto $\mathbf{S}$ é dividido com base na característica D, sendo calculado por:

$$IG(\mathbf{S}, D) = H(\mathbf{S}) - H(\mathbf{S}|D) = H(\mathbf{S}) - \sum_{j \in V(D)} \frac{\mathbf{S}_j}{\mathbf{S}} H(\mathbf{S}_j) \tag{4.2}$$

onde $V(D)$ denota os valores possíveis para a característica D, e $\mathbf{S}_j$ é o subconjunto de $\mathbf{S}$ para o qual a característica D tem valor j. A característica mais adequada a ser usada como crité-

rio de decisão é aquela que resulta no valor máximo de $IG(\mathbf{S}, D)$, pois, maximizando o ganho de informação, minimiza-se o grau de impureza. Contudo, o uso do ganho de informação como critério tem uma desvantagem inerente da entropia, favorecendo características com um alto número de valores possíveis. Para evitar este inconveniente, o ganho de informação deve ser normalizado pela entropia de **S** em relação aos valores da característica D, resultando em um outro critério denominado taxa de ganho (*gain ratio*):

$$GainRation(\mathbf{S}, D) = \frac{IG(\mathbf{S}, D)}{-\sum_{j \in V(D)} \frac{|\mathbf{S}_j|}{|\mathbf{S}|} \log_2 \frac{|\mathbf{S}_j|}{|\mathbf{S}|}} \tag{4.3}$$

Os classificadores mais conhecidos baseados em árvore de decisão são ID3, C4.5 e CART, os quais são descritos a seguir.

4.2.1 ID3 *(Iterative Dichotomiser 3)*

Criado por Quinlan (1986), o algoritmo ID3 *(Iterative Dichotomiser 3)* foi o pioneiro na indução de árvores de decisão, sendo recursivo e baseado na busca gulosa, procurando, sobre um conjunto de atributos, aqueles que melhor dividem os exemplos, gerando subárvores. O ID3 não trata valores desconhecidos, ou seja, todos os exemplos do conjunto de treinamento devem ter valores conhecidos para todos os seus atributos e caso os conjuntos de dados apresentem valores desconhecidos é necessário uma etapa de pré-processamento. Além disso lida apenas com atributos categóricos não-ordinais, sendo necessário discretizar os atributos contínuos para processá-los.

O ID3 utiliza o ganho de informação para selecionar a melhor divisão (teste). No entanto, esse critério não considera o número de divisões (número de arestas), e isso pode acarretar em árvores mais complexa. Tal complexidade poderia ser amenizada por métodos pós-poda, porém, não estão presentes no algoritmo original.

4.2.2 C4.5

O algoritmo C4.5 (QUINLAN, 1993) representa uma significativa evolução do ID3 Quinlan (1986). As principais contribuições em relação ao ID3 são:

- Lida tanto com atributos categóricos (ordinais ou não-ordinais) como com atributos contínuos. Para lidar com atributos contínuos, o algoritmo C4.5 define um limiar e então divide os exemplos de forma binária: aqueles cujo valor do atributo é maior que o limiar e aqueles cujo valor do atributo é menor ou igual ao limiar;

- O algoritmo C4.5 permite que os valores desconhecidos para um determinado atributo sejam representados como '?', e o algoritmo trata esses valores de forma especial, sendo estes desconsiderados nos cálculos de ganho e entropia;

- Utiliza a medida de taxa de ganho, vista da Equação 4.3, para selecionar o atributo que melhor divide os exemplos. Essa medida se mostrou superior ao ganho de informação, gerando árvores mais precisas e menos complexas;

- Apresenta um método de pós-poda das árvores geradas. O algoritmo C4.5 faz uma busca na árvore, da raiz para a folha, e transforma em nós folha aqueles ramos que não apresentam nenhum ganho significativo. Tal poda é um processo de controle de *overfitting* (super-ajuste) do algoritmo.

4.2.3 CART

Desenvolvido por Breiman *et al.* (1984), o algoritmo CART , do inglês *Classification and Regression Trees*, tem sua base fundamentada no método de árvores de decisão. Uma de suas principais características é a capacidade de pesquisa de relações entre os dados, compreendendo a simplificação e construção das árvores de decisão, escolhendo a melhor variável para dividir os dados em nós, aplicando recursivamente o procedimento de divisão aos dados em cada um dos nós filhos (HAND; MANNILA; SMYTH, 2001).

Na construção de uma árvore de decisão através do algoritmo CART existem 4 elementos: um conjunto de perguntas binárias, critérios de divisão dos nós, associação de uma classe a folha e a poda. Cada nó **t** é divido pelo conjunto de perguntas binárias **Q**; em casos positivos segue-se para o nó esquerdo e em casos negativos para o nó direito.

O critério para a divisão baseia-se na construção de uma árvore de decisão pelo CART, sendo que a primeira tarefa é descobrir qual dos atributos realiza a melhor divisão, onde cada atributo é levado em consideração, sendo testado como possível divisor. Na etapa de seleção do atributo, onde se procura diminuir a impureza dos nós, pode-se definir o ganho de uma divisão como sendo a diferença entre a impureza do nó pai e a soma desta para os nós filhos. Se a divisão s separar o nó **t** em dois subconjuntos **tL** e **tR** com as proporções **pL** e **pR**, o ganho da impureza será dado por:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{4.4}$$

O algoritmo CART normalmente considera o critério Gini para a divisão dos dados. O critério Gini mede a heterogeneidade dos dados e pode ser entendido como a probabilidade condicional de erro, dado um conjunto de treinamento selecionado de forma aleatória que é dividido em um nó **t**, onde cada classe **j** tem uma probabilidade **p**(**j**|**t**). Segundo o critério de Gini, a impureza de um nó é dada por:

$$i(s,t) = 1 - \sum_j p^2(j|t) \tag{4.5}$$

onde **P**(**j**|**t**) é a probabilidade a priori da classe **j** se formar no nó **t**.

Diferente de outros algoritmos, o CART não utiliza a pré-poda mas sim expande exaustivamente a árvore relizando a pós-poda por meio da redução do fator custo-complexidade. Segundo Breiman *et al.* (1984) a técnica de poda do algoritmo CART é muito eficiente produzindo árvores mais simples e com boa capacidade de generalização.

4.3 *Naive Bayes*

O algoritmo classificador *Naive Bayes* é um classificador estatístico baseado no teorema de Bayes (THEODORIDIS; KOUTROUMBAS, 1999). O teorema de Bayes é definido do seguinte modo: seja $C = \{c_1, c_2, \ldots, c_l\}$ o conjunto de classes dos dados e $\mathbf{x}$ uma instância de classe desconhecida. Considerando que $\mathbf{x}$ pertença a uma das classes do conjunto C, deseja-se determinar $P(c_i|\mathbf{x})$, $1 \leq i \leq l$, ou seja, a probabilidade da classe c_i dada a instância $\mathbf{x}$. O cálculo da probabilidade *a posteriori* da classe c_i condicionada a $\mathbf{x}$, $P(c_i|\mathbf{x})$, é dado pela regra de Bayes:

$$P(c_i|\mathbf{x}) = \frac{P(\mathbf{x}|c_i)P(c_i)}{P(\mathbf{x})}, \tag{4.6}$$

onde $P(c_i)$ é a probabilidade *a priori* da classe c_i, $P(\mathbf{x})$ é a probabilidade *a priori* de $\mathbf{x}$ e $P(\mathbf{x}|c_i)$ é a probabilidade *a posteriori* de $\mathbf{x}$ condicionada a classe c_i. As probabilidades $P(c_i)$, $P(\mathbf{x})$ e $P(\mathbf{x}|c_i)$ são estimadas a partir das instâncias de treinamento.

Dado um exemplo $\mathbf{x}$ de classe desconhecida, um classificador bayesiano prediz que $\mathbf{x}$ pertence a classe que tem a maior probabilidade *a posteriori* $P(c_i|\mathbf{x})$, i.e., $\arg_{c_i} \max P(c_i|\mathbf{x})$. Considerando $P(\mathbf{x})$ constante para todas as classes tem-se que:

$$P(c_i|\mathbf{x}) = P(\mathbf{x}|c_i)P(c_i) \tag{4.7}$$

Este classificador é denominado ingênuo (*naive*) por assumir que as características são condicionalmente independentes, ou seja, que a informação de um evento não é informativa sobre nenhum outro. Assumindo que as características são condicionalmente independentes dada a classe, tem-se que:

$$P(\mathbf{x}|c_i) = \prod_{k=1}^{m} P(x_k|c_i), \tag{4.8}$$

sendo m o número de características dos exemplos e $P(x_k|c_i)$ é estimada dos exemplos de treinamento do seguinte modo:

- Se x_k for categórico, $P(x_k|c_i) = s_{ik}/s_i$, onde s_{ik} é o número de exemplos de treino da classe c_i que têm o valor x_k para a característica A_k e s_i é o número de exemplos de treino da classe c_i.

- Se a característica A_k for contínua, é assumido que ela possui uma distribuição gaussiana sendo calculada a probabilidade como:

$$P(x_k|c_i) = \frac{1}{\sigma_{c_i}\sqrt{2\pi}} e^{-\frac{(x_k-\mu_{c_i})^2}{2\sigma_{c_i}^2}}, \tag{4.9}$$

onde μ_{c_i} e σ_{c_i} são, respectivamente, a média e o desvio padrão dos valores da característica de índice k para os exemplos da classe c_i.

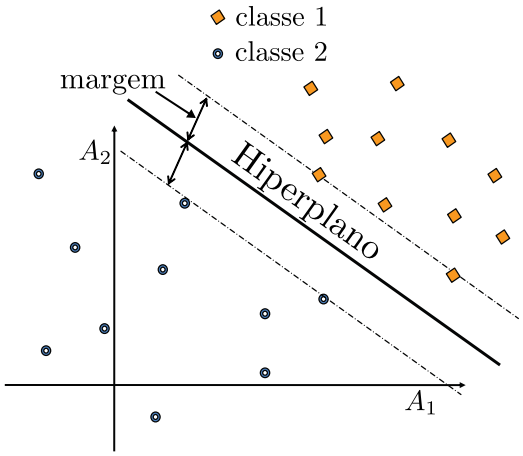
O classificador *Naive Bayes* é simples e, geralmente, apresenta alta precisão preditiva e escalabilidade em grandes bases de dados de alta dimensionalidade (MITRA; ACHARYA, 2003).

4.4 *Support Vector Machines*

As *Support Vector Machines* (SVMs) foram originalmente formuladas para lidar com problemas de classificação binários (duas classes). Atualmente, existe uma série de técnicas que podem ser empregadas na generalização das SVMs para a resolução de problemas multiclasses (VAPNIK, 1995).

Dado um conjunto de treinamento composto por n amostras, denominadas vetores no contexto das SVMs, pertencentes a duas classes linearmente separáveis, o objetivo é definir um hiperplano que separe os vetores. Entre os muitos hiperplanos possíveis, o hiperplano separador ótimo é o plano que maximiza a margem, ou seja, a distância entre o hiperplano e o vetor mais próximo de cada classe. A Figura 4.2 ilustra um hiperplano separador ótimo.

Figura 4.2 – Hiperplano de separação SVM de maior margem.

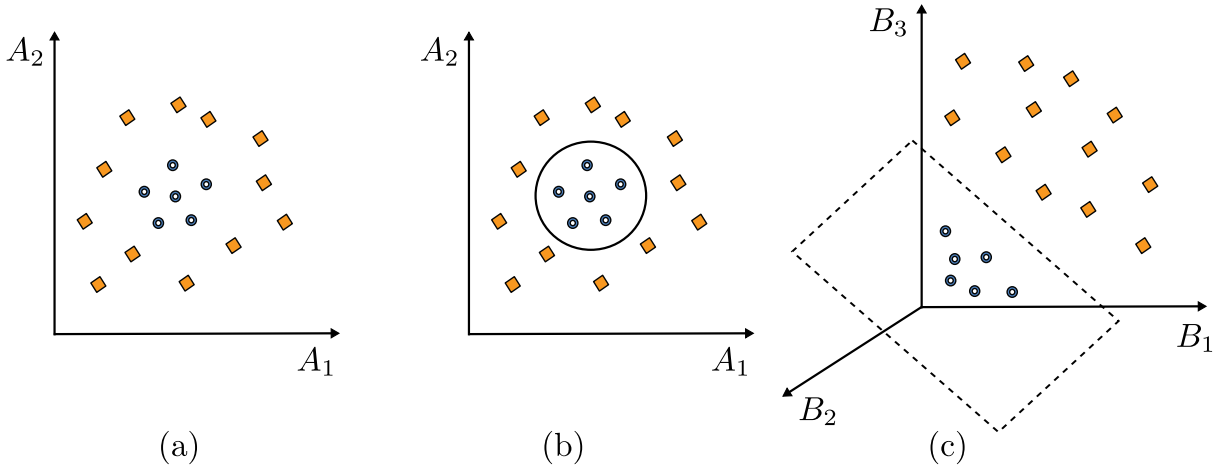


Fonte: adaptado de (HAYKIN, 2009)

As SVMs lidam com problemas não lineares realizando um mapeamento da forma $\Phi : \mathbf{A} \rightarrow \mathbf{B}$ no qual $\mathbf{A}$ é o espaço de características original do problema e $\mathbf{B}$ é o espaço de

destino do mapeamento, que deve ter dimensionalidade maior que a de **A** (veja Figura 4.3). As funções que realizam este tipo de mapeamento são denominadas funções Kernel. Uma escolha apropriada de função Kernel Φ faz com que o conjunto de treinamento **Q** mapeado do espaço de características **A** para **B** seja separável por uma SVM linear (teorema de Cover (HAYKIN, 2009)). Os tipos de funções Kernel mais utilizadas na prática são as polinomiais, gaussianas (*Radial Basis Functions* - RBfs) e as sigmoidais.

Figura 4.3 – Mapeamento de um conjunto de dados não linearmente em um linearmente separável: (a) Conjunto de dados não linearmente separável; (b) Fronteira não linear entre as classes no espaço original; (c) Fronteira linear entre as classes no espaço transformado.



Fonte: adaptado de (HAYKIN, 2009)

Existem basicamente duas abordagens de SVM multiclasse: a de decomposição do problema multiclasse em vários subproblemas binários e a de reformulação do algoritmo de treinamento das SVM em versões multiclasse. Em geral, esse último procedimento leva a algoritmos computacionalmente complexos (HSU; LEE; ZHANG, 2001). Por esse motivo, a estratégia decomposicional é empregada mais frequentemente. Uma revisão a respeito da obtenção de previsões multiclasse com SVM pode ser encontrada em (LORENA, 2006).

4.5 Classificador dos Vizinhos mais Próximos (*k-Nearest Neighbor*)

Os classificadores apresentados até o momento são caracterizados pelo fato de construírem um modelo de classificação utilizando os dados de treinamento. Normalmente, a construção de modelo demanda um custo computacional considerável, enquanto que a classificação de novos objetos é feita de forma rápida. Tais classificadores são chamados de classificadores apressados (*eager classifiers*). Ao contrário dos classificadores apresentados até então, os classificadores preguiçosos (*lazy classifiers*) não constroem modelos de classificação na fase de treinamento. Os objetos não rotulados são classificados com base na em

estatísticas dos padrões de treinamento que mais se assemelham a eles. Como não é construído um modelo, para cada objeto a ser classificado é analisado todo conjunto de treinamento. Obviamente, este processo é computacionalmente dispendioso, especialmente para conjuntos de treinamento com um elevado número de atributos e de instâncias. O exemplo mais popular de classificador preguiçoso é o *k-Nearest Neighbor* (kNN) (DUDA; HART; STORK, 2001).

O classificador kNN funciona da seguinte forma. Suponha um conjunto **Q** de amostras de treinamento. Cada elemento de **Q** é uma tupla $(\mathbf{x}, c)$, onde **x** é um objeto (vetor de características) m dimensional e c é o seu rótulo. Seja **y** um novo objeto não rotulado. Com o objetivo de classificar **y** calcula-se a distância de **y** a todos os objetos de treinamento **Q**. O rótulo de **y** é dado pela classe que ocorre com maior frequência nos k objetos mais próximos de **y**.

Antes do processo de classificação, os valores de características são, normalmente, normalizados para que valores em diferentes escalas não produza *bias* (tendência) no cálculo de distância (DUDA; HART; STORK, 2001). As métricas de normalização mais utilizadas são *standardization* (também conhecida como *z-score*), dada pela Equação 4.10 e *normalization* dada pela Equação 4.11, onde v_i é o valor a ser normalizado, $\mu(\mathbf{v})$ e $\sigma(\mathbf{v})$ correspondem à média e ao desvio padrão dos valores em **v**, sendo **v** um vetor de valores de uma determinada característica das instâncias do conjunto de dados.

$$z_i = \frac{v_i - \mu(\mathbf{v})}{\sigma(\mathbf{v})} \tag{4.10}$$

$$n_i = \frac{v_i - \min(\mathbf{v})}{\max(\mathbf{v}) - \min(\mathbf{v})} \tag{4.11}$$

4.6 *Multilayer Perceptron*

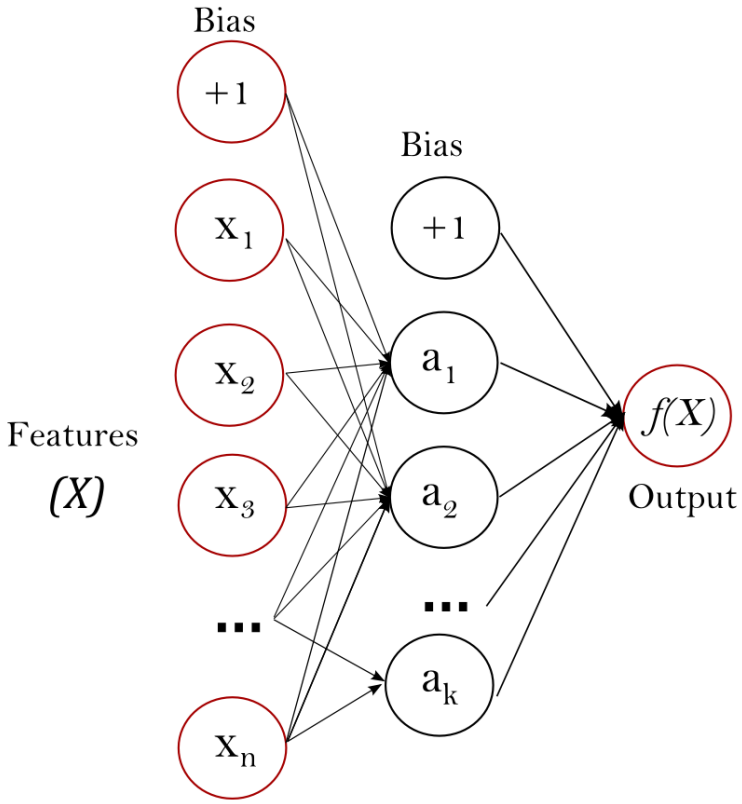
Multilayer Perceptron (MLP) trata-se de uma rede neural que se assemelha com a *Perceptron* ((HAYKIN, 2001)), mas possui múltiplas camadas de neurônios em alimentação direta. As camadas desta rede neural são ligadas entre si por sinapses contendo pesos.

As redes MLP apresentam desempenho computacional maior quando comparado com redes sem camadas intermediárias e podem lidar com dados que são não linearmente separáveis. O algoritmo de aprendizado supervisionado MLP aprende uma função $f(\cdot) : R^m \rightarrow R^o$ treinando em um conjunto de dados, onde m é o número de dimensões para entrada e o é o número de dimensões para saída. Dado um conjunto de características $\mathbf{x} = x_1, x_2, ..., x_m$ e um alvo **y**, uma rede neural pode aprender uma aproximação de função não linear para classificação ou regressão.

A Figura 4.4 mostra uma MLP de camada oculta com saída escalar (NIEMINEN, 2016). A camada mais a esquerda, conhecida como camada de entrada, consiste em um conjunto

de neurônios $x_i|x_1, x_2, \ldots, x_m$ representando as características de entrada. Cada neurônio na camada oculta transforma os valores da camada anterior com uma soma linear ponderada $w_1x_1 + w_2x_2 + \cdots + w_mx_m$, seguido por uma função de ativação não-linear $g : R \rightarrow R$, como a função tangente hiperbólica. A camada de saída recebe os valores da última camada oculta e transforma-os em valores de saída.

Figura 4.4 – Rede MLP com uma camada oculta.



Fonte: adaptado de Nieminen (2016)

A rede neural MLP utiliza geralmente o algoritmo de *backpropagation* para realizar o aprendizado de máquina, sendo o mais comum na literatura. Este algoritmo foi introduzido por Rumelhart, Hinton e Williams (1986) e oferece a base para um grande número de mecanismos de aprendizagem para funções de custos diferenciáveis. O método de *backpropagation* calcula derivadas parciais (gradiente) do erro em relação a cada peso e, em seguida, atualiza os pesos na direção oposta ao gradiente, multiplicado por uma pequena taxa de aprendizado. O cálculo para os ajustes nos pesos sinápticos do algoritmo *backpropagation* pode ser visto na equação 4.12.

$$w_{ji}^{(l)}(n+1) = w_{ji}^{(l)}(n) + \alpha(w_{ji}^{(l)}(n-1)) + \eta\delta_j^{(l)(n)} y_i(l-1)(n), \tag{4.12}$$

onde η é o parâmetro de taxa de aprendizado, α é uma constante de momento, $y_i^{(l-1)}$ corresponde ao valor de saída do neurônio da camada anterior, n corresponde ao n-ésimo exem-

plo de treinamento e $\delta_j^{(l)}(n)$ corresponde ao cálculo de gradiente de erro para o neurônio, sendo dado por:

$$\delta_j^{(l)}(n) = \begin{cases} e_j^{(L)}(n)\varphi'_j(v_j^{(L)}(n)), \text{para um neurônio j na camada de saída} \\ \varphi'_j(v_j^{(L)}(n)\sum_k \delta_k^{(l+1)}(n) w_{kj}^{(l-1)}(n), \text{para um neurônio j em uma camada escondida } l \end{cases} \tag{4.13}$$

onde $\varphi'_j(.)$ corresponde a diferenciação da função de ativação com respeito ao argumento e $e_j^{(L)}$ corresponde ao cálculo de erro no neurônio j na camada de saída sendo dado por $e_j^{(L)}(n) = d_j^{(L)}(n) - y_j^{(L)}(n)$, onde $d_j^{(L)}(n)$ é o valor desejado e $y_j^{(L)}(n)$ é a saída real do neurônio.

Uma rede neural MLP é treinada pela alimentação repetida por padrões de treinamento para ajuste dos pesos, de forma a minimizar o erro da saída da rede.

4.7 Classificador *Radial-Basis Function Network*

Radial-Basis Function Networks (redes RBF) também são espécies de redes neurais de múltiplas camadas, contudo usam uma abordagem complementante diferente de treinamento. Especificamente, elas resolvem o problema de classificar padrões não linearmente separáveis por proceder de uma maneira híbrida que envolve dois estágios:

- O primeiro estágio transforma um dado conjunto de padrões não linearmente separáveis em um novo conjunto que, sob certas condições, a probabilidade dos padrões transformados se tornarem linearmente separáveis é alta; a justificativa matemática para esta transformação é atribuída ao artigo de (COVER, 1965).
- O segundo estágio completa a solução para o problema de classificação prescrito usando estimação dos mínimos quadrados.

A estrutura padrão de uma rede RBF consiste de três camadas:

- A camada de entrada é dada por nós fontes (unidades sensórias) que conecta a rede ao seu ambiente;
- A segunda camada consiste de neurônios escondidos e aplica uma transformação não linear do espaço de entrada para o espaço escondido, também chamado de espaço de características. Para maioria das aplicações a dimensionalidade da única camada escondida da rede é alta, ou seja, esta camada contém um número grande de neurônios. Está camada é treinada de uma maneira não supervisionada usando o primeiro estágio do procedimento de aprendizagem híbrido.

- A camada de saída é linear, projetada para fornecer uma resposta da rede para o padrão de ativação aplicado na camada de entrada; esta camada é treinada de maneira supervisionada usando o segundo estágio do procedimento híbrido.

Uma rede RBF funciona com base no princípio do teorema de Cover sobre separabilidade de padrões: "Um problema complexo de classificação de padrões, convertido não-linearmente em um espaço de alta dimensão, é mais provável de ser linearmente separável do que em um espaço de baixa dimensão, desde que o espaço não seja densamente povoado". A transformação não-linear do espaço de entrada da rede para o espaço escondido e a alta dimensionalidade do espaço escondido satisfazem os dois únicos ingredientes do teorema de Cover, os quais são:

1. Uma formulação não linear de função definida como $\phi_i(\mathbf{x})$, onde $\mathbf{x}$ é o vetor de entrada e $i = 1, ..., m_1$, sendo m_1 o número de neurônio da camada escondida, que também corresponde a dimensionalidade do espaço escondido.

2. uma alta dimensionalidade do espaço escondido (de características) comparado com o espaço de entrada, onde a dimensionalidade do espaço escondido é dada por m_1, isto é, o número de neurônios escondidos.

Um ponto importante que emerge do teorema de Cover sobre a separabilidade de padrões é que na resolução de problemas de classificação não linearmente separáveis há usualmente benefícios práticos em mapear o espaço de entrada em um novo espaço de dimensão alta o suficiente. Basicamente o mapeamento não linear é usado para transformar um problema de classificação não linearmente separável em um linearmente separável com alta probabilidade.

Considere então uma rede de alimentação para frente com uma camada de entrada, uma única camada escondida e uma camada de saída contendo um único neurônio. Escolhe-se propositalmente, sem perda de generalidade, um único neurônio de saída para simplificar a exposição. A rede é projetada para executar um mapeamento não-linear no espaço do espaço de entrada para o espaço escondido, seguido de um mapeamento linear do espaço escondido para o espaço de saída. Seja m_0 a dimensão do espaço de entrada. Então, de um modo global, a rede representa um mapeamento de um espaço m_0-dimensional para um espaço unidimensional, escrito como:

$$s : \mathbb{R}^{m_0} \rightarrow \mathbb{R}^1 \tag{4.14}$$

Pode-se imaginar o mapeamento s como uma hipersuperfície (grafo) $\Gamma \subset \mathbb{R}^{m_0+1}$, do mesmo modo que se pode pensar no mapeamento elementar $s : \mathbb{R}^1 \rightarrow \mathbb{R}^1$, onde $s(x) = x^2$,

como uma parábola desenhada no espaço $\mathbb{R}^2$. A superfície Γ é um gráfico multidimensional da saída como uma função da entrada. Em uma situação prática a superfície Γ é desconhecida e os dados de treinamento são usualmente contaminados com ruído. A fase de treinamento e a fase de generalização do processo de aprendizagem podem ser vistas respectivamente como segue:

- A fase de treinamento constitui na otimização de um procedimento de ajuste para a superfície Γ, baseado nos dados de treinamento apresentados para a rede na forma de exemplos de entrada-saída.

- A fase de generalização é sinônimo de interpolação entre os dados de treinamento com a interpolação sendo executada ao longo da superfície restrita gerada pelo procedimento de ajuste como a aproximação ótima para a superfície verdadeira Γ.

Assim, somos levados para a teoria de interpolação multivariável em um espaço de alta dimensionalidade (DAVIS, 1963). O problema de interpolação, em seu modo estrito, pode ser declarado da seguinte maneira: dado um conjunto de N diferentes pontos $\{\mathbf{x}_i \in \mathbb{R}^{m_0}, i = 1,\ldots,N\}$ e um conjunto correspondente de N números reais $\{d_i \in \mathbb{R}^1, i = 1,\ldots,N\}$, encontrar uma função $F : \mathbb{R}^N \rightarrow \mathbb{R}^1$ que satisfaça a condição de interpolação:

$$F(\mathbf{x}_i) = d_i, \qquad i = 1,\ldots,N \tag{4.15}$$

Para a interpolação estrita como especificado, a superfície de interpolação (isto é, a função F) é restrita a passar através de todos os pontos de dados de treinamento. A técnica RBF consiste em escolher uma função F que tem a forma:

$$F(\mathbf{x}) = \sum_{i=1}^{N} w_i \varphi(||\mathbf{x} - \mathbf{x}_i) \tag{4.16}$$

Onde $\{\varphi(||\mathbf{x} - \mathbf{x}_i||, i = 1,\ldots,N\}$ é na Equação 4.16, um conjunto de N funções arbitrárias (geralmente não-lineares), conhecidas como funções de base radial, e $||.||$ denota uma norma que é usualmente a distância Euclidiana (POWELL, 1988). Os pontos de dados conhecidos $\mathbf{x}_i \in \mathbb{R}^{m_0}, i = 1,\ldots,N$ são tomados como os centros das funções de base radial.

Muita da teoria desenvolvida sobre redes RBF é suportada pela função Gaussiana, um importante membro das classe de funções de base radial. A função Gaussiana pode também ser vista como um kernel – daí a denominação do procedimento em dois estágios baseado em função Gaussiana como método de kernel.

Inserindo as condições de interpolação da Equação 4.15 na Equação 4.16, obtem-se um conjunto de equações lineares simultâneas para os coeficientes desconhecidos (pesos)

dadas por:

$$\begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1N} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_{N1} & \varphi_{N2} & \cdots & \varphi_{NN} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} \tag{4.17}$$

onde

$$\varphi_{ij} = \varphi(||\mathbf{x}_i - \mathbf{x}_j||), i, j = 1, \ldots, N$$

Seja

$$\mathbf{d} = [d_1, d_2, \ldots, d_n]^T,$$

$$\mathbf{w} = [w_1, w_2, \ldots, w_n]^T.$$

Os vetores $N \times 1$ **d** e **w** representam o vetor resposta desejada e o vetor de pesos lineares, respectivamente, onde N é número de amostras de treinamento. Seja Φ uma matriz $N \times N$ com elementos φ_{ij}:

$$\Phi = \{\varphi_{ij}\}_{i=j=1}^{N} \tag{4.18}$$

Esta matriz é chamada de matriz de interpolação. Pode-se re-escrever a Equação 4.17 de forma mais compacta como:

$$\Phi\mathbf{w} = \mathbf{x} \tag{4.19}$$

Assumindo que Φ é não-singular, e, consequentemente, que a a matriz inversa Φ^{-1} existe, pode-se resolver a Equação 4.20 para o vetor de pesos **w**, obtendo:

$$\mathbf{w} = \Phi^{-1}\mathbf{x} \tag{4.20}$$

Com base nas Equações 4.15 até 4.20 é possível agora imaginar uma rede RBF na forma de uma estrutura em camadas, como ilustrado na Figura 4.5. Especificamente, observa-se três camadas:

1. **Camada de entrada**, que consiste em m_0 nós fontes, onde m_0 é a dimensionalidade do vetor de entrada **x**.

2. **Camada escondida**, que consiste de neurônio de computação na quantidade N de amostras de treinamento. Cada neurônio desta camada é descrito matematicamente

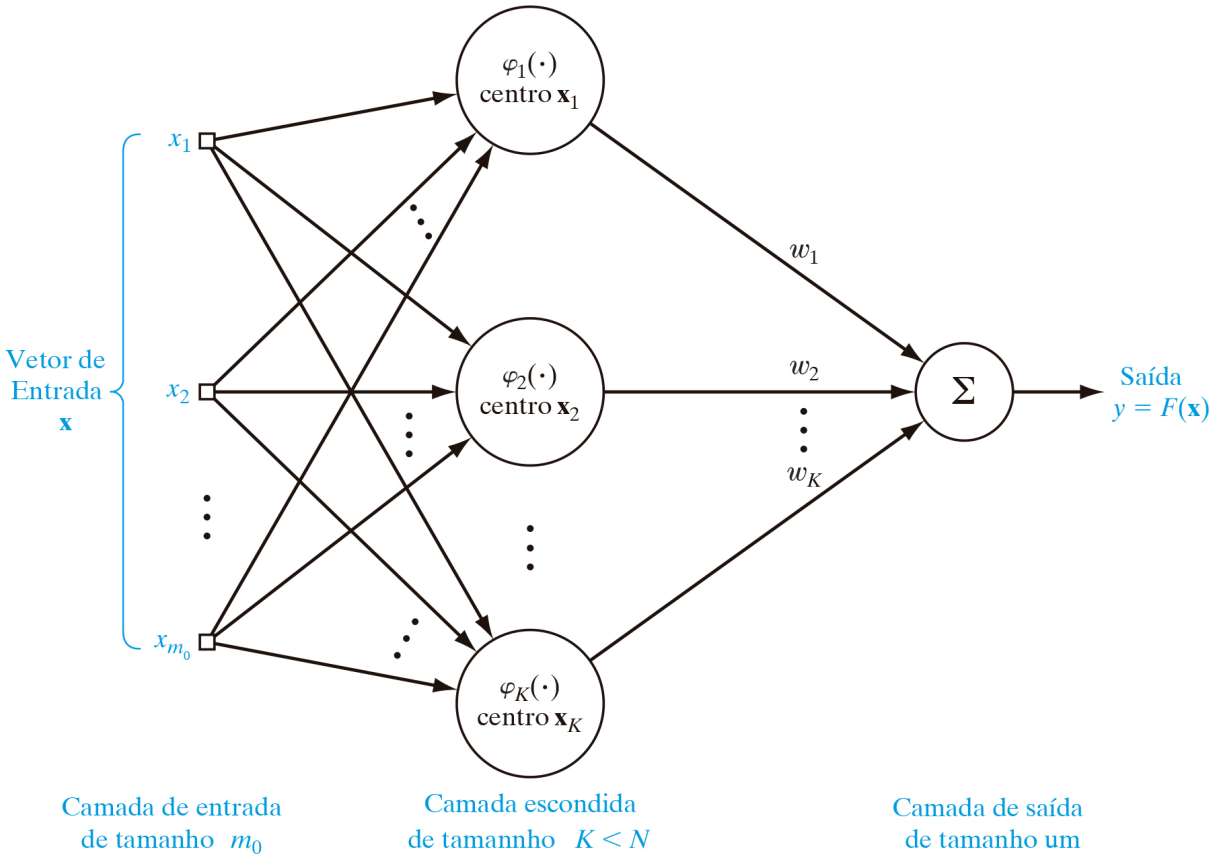
por uma função de base radial:

$$\varphi_j(\mathbf{x}) = \varphi(||\mathbf{x}-\mathbf{x}_j), j = 1,\ldots,N$$

O j-ésimo ponto de dado de entrada define o centro da função de base radial, e o vetor $\mathbf{x}$ é o sinal (padrão) aplicado na camada de entrada. Assim, diferentemente de uma rede perceptron de múltiplas camadas, as conexões ligando os nós fontes aos neurônios escondidos são conexões diretas sem pesos.

3. **Camada de saída**, que na estrutura RBF da Figura 4.5 consiste de um único neurônio computacional. Claramente, não há restrições quanto ao número de neurônios da camada de saída, contudo normalmente o número de neurônios na camada de saída é muito menor que o número de neurônios na camada escondida.

Figura 4.5 – Estrutura usual de uma rede RBF.



Fonte: Autor.

A função de base radial é normalmente definida com base em uma função Gaussiana:

$$\varphi_j(\mathbf{x}) = \varphi(\mathbf{x}-\mathbf{x}_j) = \exp(-\frac{1}{2\sigma_j^2}||\mathbf{x}-\mathbf{x}_j||^2), j = 1,\ldots,N \tag{4.21}$$

onde σ_j é uma medida da amplitude da j-ésima função Gaussiana com centro em $\mathbf{x}_j$. Tipicamente, mas nem sempre, todas os neurônios escondidos são atribuídos com a mesma amplitude σ. Em situações deste tipo, o parâmetro que distingue uma neurônio escondido de outro é o centro $\mathbf{x}_j$.

4.8 Técnicas de amostragem de dados

Buscando a obtenção de estimativas mais consistentes com relação ao desempenho de um classificador e diminuir o *bias* em relação aos dados de treinamento pode-se utilizar técnicas de amostragem na validação de modelos, tais como, a amostragem aleatória (*random sampling*) e os métodos de validação cruzada: *k-fold cross-validation* e *leave-one-out.*

***Random sampling*:** consiste em dividir aleatoriamente o conjunto de dados em subconjuntos disjuntos. Por exemplo: 70% das amostras para treinamento e 30% para teste. Este processo pode ser repetido várias vezes buscando uma melhor estimativa média de desempenho de um modelo.

***k-fold cross-validation*:** consiste em dividir o conjunto de dados em k partições mutuamente exclusivas e experimentar o modelo k vezes, utilizando $k-1$ partições para treinamento e uma partição para teste. A taxa de erro é dada pela média dos erros de teste obtidos nas k repetições. Quando a proporção de objetos por classe do conjunto completo é mantida nas partições, este procedimento recebe o nome de *stratified k-fold cross validation.*

***Leave-one-out*:** o modelo é executado N vezes, considerando um conjunto de N amostras. Em cada iteração, $N-1$ amostras são utilizadas para treinamento do modelo e uma amostra é utilizada para teste. A taxa de erro é obtida dividindo o número de erros obtidos nos N testes por N.

4.9 Balanceamento de classes

O desbalanceamento de classes ocorre quando há uma grande desproporção na quantidade de itens para classes distintas , por exemplo, em uma base de doadores contendo 500 indivíduos, onde 50 doaram e 450 não doaram. O desbalanceamento de classes influencia negativamente na validação de um classificador, com base nos dados anteriores, se um classificador predizer que nenhum doador doou, ele teria 90% de acurácia.

Diversas técnicas podem ser utilizadas para tornar as classes balanceadas. A forma mais direta é a reamostragem dos dados eliminando exemplos a classe majoritária (*under-sampling*) ou replicando exemplos da classe minoritária (*up-sampling*).

A sobreamostragem (*up-sampling*) pode replicar exemplos preexistentes ou gera dados sintéticos. No primeiro caso a seleção dos dados pode ser aleatória ou direcionada e um dos métodos mais utilizados para seleção aleatória é o *Synthetic Minority Over-sampling Technique* (SMOTE) (BOWYER *et al.*, 2011).

O SMOTE trata-se de uma abordagem que superestima a classe minoritária criando exemplos sintéticos diferente da uma sobre-amostragem com replicação. Este método foi inspirado por uma técnica que obteve sucesso no reconhecimento de caracteres manuscritos, onde foram criados dados adicionais de treinamento executando operações em dados reais, como rotação e distorção. Já o SMOTE gera exemplos sintéticos de uma maneira menos específica operando no espaço de característica ao invés do "espaço de dados" (BOWYER *et al.*, 2011).

A superestimação da classe minoritária acontece tomando cada amostra e introduzindo os exemplos sintéticos nos segmentos de linha juntando os vizinhos mais próximos da classe. Por padrão o SMOTE utiliza 5 vizinhos mais próximos podendo variar de acordo com a taxa de *up-sampling*. Por exemplo, se taxa de *up-sampling* for 200%, apenas dois vizinhos dos cinco vizinhos mais próximos são escolhidos e uma amostra é gerada na direção de cada um. As amostras sintéticas são geradas obtendo a diferença entre o vetor de características (amostra) em consideração e seu vizinho mais próximo. Depois, multiplicando essa diferença por um número aleatório entre 0 e 1, e adicionando o resultado obtido ao vetor de características em consideração. Isso faz com que a seleção de um ponto aleatório seja feito longo do segmento de linha entre duas amostras específicas. Esta abordagem efetivamente obriga a região de decisão da classe minoritária tornar-se mais generalizada (BOWYER *et al.*, 2011).

4.10 Métodos de estimativa da probabilidade de classe

Para que seja possível um classificador retornar não só o rótulo (classificação) de uma amostra mas também a probabilidade desta pertence em pertencer a classe prevista, pode-se utilizar métodos como o *Platt Scaling* baseado na *Logistic Regression* descrito a seguir.

Logistic Regression (LR) trata-se de um método estatístico para analise de dados com variáveis de resultados muti-classes. Este método busca encontrar o modelo mais adequando e a economia na relação entre o resultado (variável dependente ou resposta) e um conjunto de variáveis independentes (preditoras ou explicativa). LR é muito utilizado em conjunto com classificadores como o SVM para melhorar estimativa de parâmetros do modelo diminuindo erro de previsões (PERME; BLAS; TURK, 2004).

Em (PLATT, 2000) o autor cria um método que ajusta um modelo de LR para gerar *scores* para um classificador SVM, fazendo com que a saída deste classificador seja uma distribuição de probabilidade sobre as classes.

Para exemplificar, considere um problema com o objetivo de determinar se uma dada entrada $\mathbf{x}$ pertence as classes $+1$ ou -1, sendo resolvido por uma função de valor real f, onde a rotulação de classe é prevista por $y = sign(f(\mathbf{x}))$. Normalmente um classificador apenas realiza a rotulação (classificação) mas pode ser necessário obter uma probabilidade $P(y = 1|x)$, dando além da classe predita, o grau de certeza sobre a predição. Alguns modelos de classificação não fornecem informações de probabilidade ou suas estimativas são pouco confiáveis. Porém, aplicando Platt Scaling é possível produzir estimativas de probabilidade através da equação:

$$P(y = 1|x) = \frac{1}{1 + \exp(Af(x) + B)} \tag{4.22}$$

sendo esta uma transformação logística dos valores obtidos pelo classificador $f(x)$, onde A e B são dois parâmetros escalares que são aprendidos pelo algoritmo. Com isso as predições são feitas de acordo com $y = P(y = 1|x) > \frac{1}{2}$; se $B \neq 0$, as estimativas de probabilidade contêm uma correção em relação à antiga função de decisão $y = sign(f(x))$.

Os parâmetros A e B são estimados usando um método de máxima verossimilhança no mesmo conjunto de treinamento usado pelo classificador original f. Para evitar um *overfitting* a este conjunto, é possível usar um conjunto de calibração estendido ou validação cruzada (PLATT, 2000).

Capítulo 5

DESENVOLVIMENTO E EXPERIMENTAÇÃO

5.1 Considerações Iniciais

Por questões de custo das UHs e principalmente pela necessidade em obter doadores de sangue de forma rápida em situações de emergência, esta pesquisa focou em técnicas computacionais para a predição de doadores que irão doar sangue em resposta a uma dada solicitação. Desta forma, considerando uma base de dados de treinamento com características de pessoas com registro em UHs, pode-se ciar um mecanismo de predição que prediz os doadores mais prováveis a doar sangue em resposta a uma requisição. Basicamente temos um problema de decisão binário, que consiste em predizer se uma dada pessoa irá doar sangue ou não, em resposta a uma dada solicitação. Este problema pode obviamente ser modelado como um problema de classificação. Assim, inicialmente a recomendação de doadores foi tratada como um problema de classificação, onde se sugere, em resposta a uma solicitação de doação, os doadores cuja predição é que eles doarão. A seguir é descrita a base de dados utilizada, enunciado formalmente o problema, enumerado os classificadores experimentados, juntamente com suas configurações e descritos os resultados obtidos.

5.2 Base de dados

Os métodos computacionais desenvolvidos e testados nesta pesquisa são validados através da base de dados *Blood Transfusion Service Center Data Set* (YEH; YANG; TING, 2009), que possui dados de doadores reais com 748 indivíduos, coletados aleatoriamente do Centro de Serviço de Transfusão de Sangue da cidade de Hsin-Chu em Taiwan, no ano de 2007. Tal base de dados é publica e atualmente encontra-se disponível para download no Repositório de Aprendizado de Máquina da Universidade da California/Irvine, em <https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center>. Os atributos da base de dados

são:

- Recência (R) – quantidade de meses desde a última doação;
- Frequência (F) – número total de doações;
- Quantidade (M) – total doado em ml;
- Uma variável binária que representa se o indivíduo doou sangue em março 2007 que foi uma campanha de doação realizada (1 significa doou, 0 significa não doou).

Os três primeiros atributos da base de dados se baseiam no método *Recency*, *Frequency* e *Monetary* (RFM), utilizado em marketing direto para segmentar clientes (FADER; HARDIE; LEE, 2005).

Nas experimentações deste trabalho assim como de outros trabalhos da literatura que utilizam esta base, tal como Santhanam e Sundaram (2010), utilizou-se as seguintes denominações para a análise dos métodos:

- Novo Doador Voluntário (*New Volunteer Donor* – NVD): doador de sangue voluntário não remunerado, que nunca doou sangue antes;
- Doador Voluntário Irregular (*Irregular Volunteer Donor* – IVD): doador de sangue voluntário não remunerado que doou sangue no passado, mas não cumpre os critérios de um doador regular;
- Doador Voluntário Regular (*Regular Volunteer Donor* – RVD): é um doador de sangue voluntário não remunerado que doa sangue regularmente, conforme uma frequência pré-definida.

Para a definição de doadores voluntários regulares (RVD) foi utilizado os critérios sugeridos em (YEH; YANG; TING, 2009), aos quais são descritos na Tabela 5.1.

Tabela 5.1 – Classificação RVD

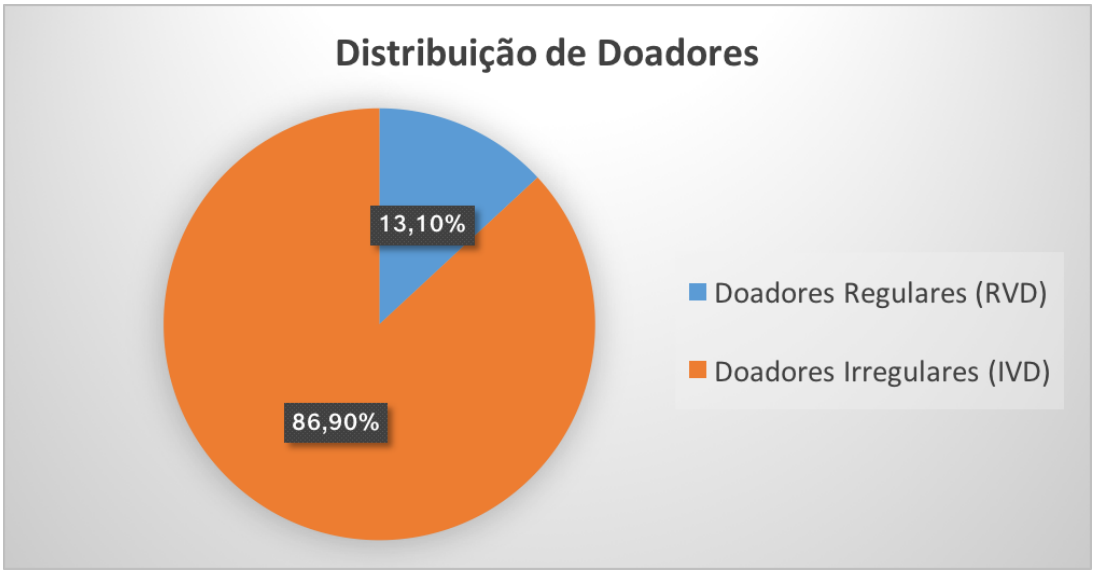
Atributo	Condição	Valor
Recência	<=	6 meses
Frequência	>=	4 meses
Quantidade	>=	2000 ml
Tempo	>	24 meses

Fonte: Santhanam e Sundaram (2010)

Em uma análise da base de dados segundo os critérios definidos, percebe-se que os doadores RVDs correspondem a 13.10% dos indivíduos do banco de dados, e foram responsáveis por 28,65% das doações de março de 2007. Realizando a comparação com os IVDs

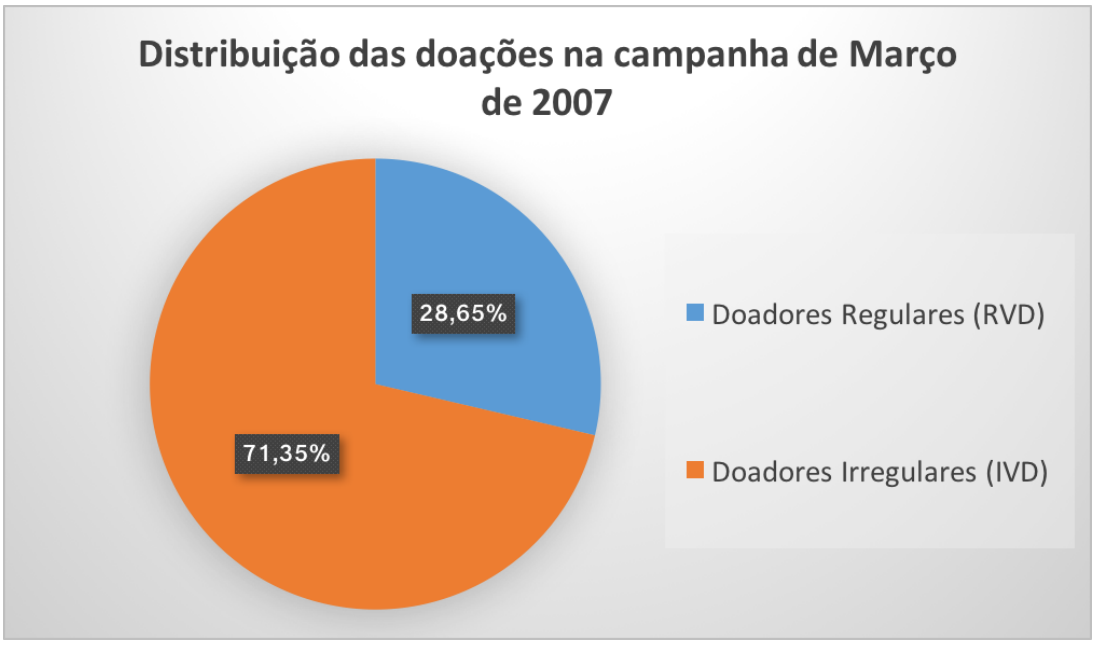
os números obtidos são favoráveis aos RVDs como pode ser visto nas Figuras 5.1 e 5.2, pois os RVDs tem uma taxa de conversão a doação 33% maior que a dos IVDs. Além disso, existem ainda outras vantagens na recomendação de RVDs. Segundo a Organização Mundial de Saúde (OMS) World Health Organization (2010) os RVDs normalmente apresentam maior qualidade de sangue pela possível maior conscientização, altruísmo, responsabilidade social e maior quantidade de triagens clínicas realizadas no processo de cada doação.

Figura 5.1 – Distribuição de doadores.



Fonte: Autor.

Figura 5.2 – Distribuição das doações na campanha de Março de 2007.



Fonte: Autor.

5.2.1 Balanceamento da Base de dados

A análise da base de dados mostrou que a divisão entre os indivíduos que doaram e não doaram no evento de março geram duas classes desbalanceadas entre si, sendo os indivíduos que não doaram 76% e que doaram 24%, dificultando a análise de desempenho dos métodos computacionais. Devido ao desbalanceamento, se um classificador predisser que nenhum individuo irá doar terá 76% de precisão. Para amenizar este aspecto problemático da base de dados este trabalho aplica o método SMOTE (BOWYER *et al.*, 2011), sendo um dos mais comumente encontrados na literatura.

5.3 Metodologia de avaliação de modelos de classificação

Para a tarefa de classificação, foi considerada uma base de dados de treinamento onde se tem o perfil do doador (características) e a informação se ele doou ou não na data da campanha (classe). Sendo que a amostra de treinamento pode ser representada por uma tupla $(\mathbf{x}, c)$, onde $\mathbf{x}$ é um vetor contendo as características (perfil) do doador e c, também denominado de classe, indica se o doador doou sangue na data da campanha. Dado um doador com um determinado perfil $\mathbf{y}$, o problema de classificação consiste em predizer se este indivíduo doará ou não sangue em uma data anunciada; tais amostras em que deseja-se predizer a saída (classe) são denominadas amostras de teste. Com a finalidade de estimar a precisão dos algoritmos de classificação a base de dados foi dividida em treinamento e teste usando a técnica de validação cruzada *random sampling* usando 70% das amostras para treinamento e 30% para teste. Os resultados são avaliados através das medidas *recall*, *precision* e *accuracy*, com base em 10000 execuções de cada modelo de classificação considerando um particionamento do tipo *random sampling* aleatório. A seguir são descritas as medidas de *recall*, *precisiom* e *accuracy*.

- *Recall* (*sensitivity*): proporção dos indivíduos que doaram sangue (falsos negativos e verdadeiros positivos) que foram preditos como doaram;
- *Precision*: proporção dos indivíduos que doaram sangue que foram preditos como doaram (verdadeiros positivos) que estão entre total de preditos como doaram (verdadeiros positivos e falsos positivos);
- *Accuracy*: corresponde a medida percentual de acertos entre a predição e o resultado real.

Matematicamente, as medidas de *sensitivity* (*recall*), *precision* e *accuracy* são dadas pelas Equações 5.1, 5.2 e 5.3, respectivamente, sendo:

- *True positive* (TP) : a quantidade de indivíduos preditos como “doou” que realmente doaram;

- *False Positive* (FP) : a quantidade de indivíduos preditos como “doou”, mas que não doaram;

- *True negative* (TN) : a quantidade de indivíduos preditos como “não-doou” que, de fato, não doaram;

- *False Negative* (FN) : a quantidade de indivíduos preditos como “não-doou”, mas que doaram;

$$sensitivity\ (recall) = \frac{TP}{TP+FN} \tag{5.1}$$

$$precision = \frac{TP}{TP+FP} \tag{5.2}$$

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{5.3}$$

5.4 Resultados dos algoritmos classificadores

Com a finalidade de encontrar o classificador mais preciso para gerar recomendações de doadores de sangue, analisou-se o desempenho de predição dos seguintes classificadores: Naive Bayes (NB), *Support Vector Machine* (SVM), *Multi-Layer Perceptron* (MLP), *Radial Basis Function Network* (RBFNet), *k-Nearest Neighbor* (kNN), e as árvores de decisão C4.5 e CART. Devido ao classificador kNN ser sensível a quantidade de vizinhos mais próximos considerados, foram experimentados vários valores diferentes para k, sendo este número ímpares de 1 à 33. Os experimentos desta etapa foram executados na ferramenta Weka usando os parâmetros *default*[1] para os demais classificadores. Os resultados de classificação obtidos, mensurados pelas medidas de *recall* (*sensitivity*), *precision* e *accuracy* são reportados na Tabela 5.2. Pode-se perceber através da Tabela 5.2 que conforme as medidas *accuracy*, *recall* e *precision* o classificador que alcançou os melhores resultados foi o Cart. Porém, o fato da base ter uma distribuição de classes desbalanceada, onde 76% das amostras pertencem a classe "não-doou", um classificador que classifica todos os indivíduos como "não-doou" teria 76% de *accuracy*, sendo uma taxa comparável a do melhor classificador.

[1] Configuração padrão

Tabela 5.2 – Resultados obtidos pelos classificadores experimentados.

Classificador	Accuracy		Recall		Precision	
	Original	Balanceada	Original	Balanceada	Original	Balanceada
Naive Bayes	75,23%	62,19%	92,62%	81,49%	78,67%	59,36%
SVM	75,42%	67,25%	96,11%	76,92%	77,21%	64,53%
MLP	77,81%	66,14%	94,76%	73,21%	79,90%	64,23%
RBFNetwork	78,33%	68,55%	93,35%	67,82%	81,14%	69,73%
kNN (k=1)	69,21%	70,12%	79,70%	69,20%	79,87%	70,63%
kNN (k=3)	75,17%	71,13%	89,16%	72,23%	80,42%	70,79%
kNN (k=5)	76,60%	71,37%	91,26%	72,42%	80,61%	71,05%
kNN (k=7)	77,01%	71,19%	92,31%	71,55%	80,43%	71,16%
kNN (k=9)	77,60%	71,48%	93,03%	70,81%	80,60%	71,89%
kNN (k=11)	77,94%	71,60%	93,42%	70,29%	80,70%	72,30%
kNN (k=13)	78,09%	71,53%	93,56%	70,11%	80,76%	72,27%
kNN (k=15)	78,25%	71,33%	93,72%	69,96%	80,82%	72,04%
kNN (k=17)	78,54%	71,15%	94,04%	69,73%	80,92%	71,88%
kNN (k=19)	78,73%	71,09%	94,30%	69,56%	80,95%	71,88%
kNN (k=21)	78,72%	71,12%	94,46%	69,48%	80,86%	71,97%
kNN (k=23)	78,59%	71,10%	94,56%	69,41%	80,69%	71,98%
kNN (k=25)	78,44%	70,90%	94,70%	69,17%	80,48%	71,78%
kNN (k=27)	78,35%	70,59%	94,93%	68,86%	80,28%	71,45%
kNN (k=29)	78,28%	70,28%	95,23%	68,57%	80,07%	71,13%
kNN (k=31)	78,20%	70,04%	95,55%	68,27%	79,83%	70,91%
kNN (k=33)	78,06%	69,90%	95,85%	68,01%	79,57%	70,82%
C4.5	77,55%	71,78%	90,31%	68,54%	82,14%	73,91%
CART	77,46%	73,80%	91,94%	74,17%	81,16%	73,84%

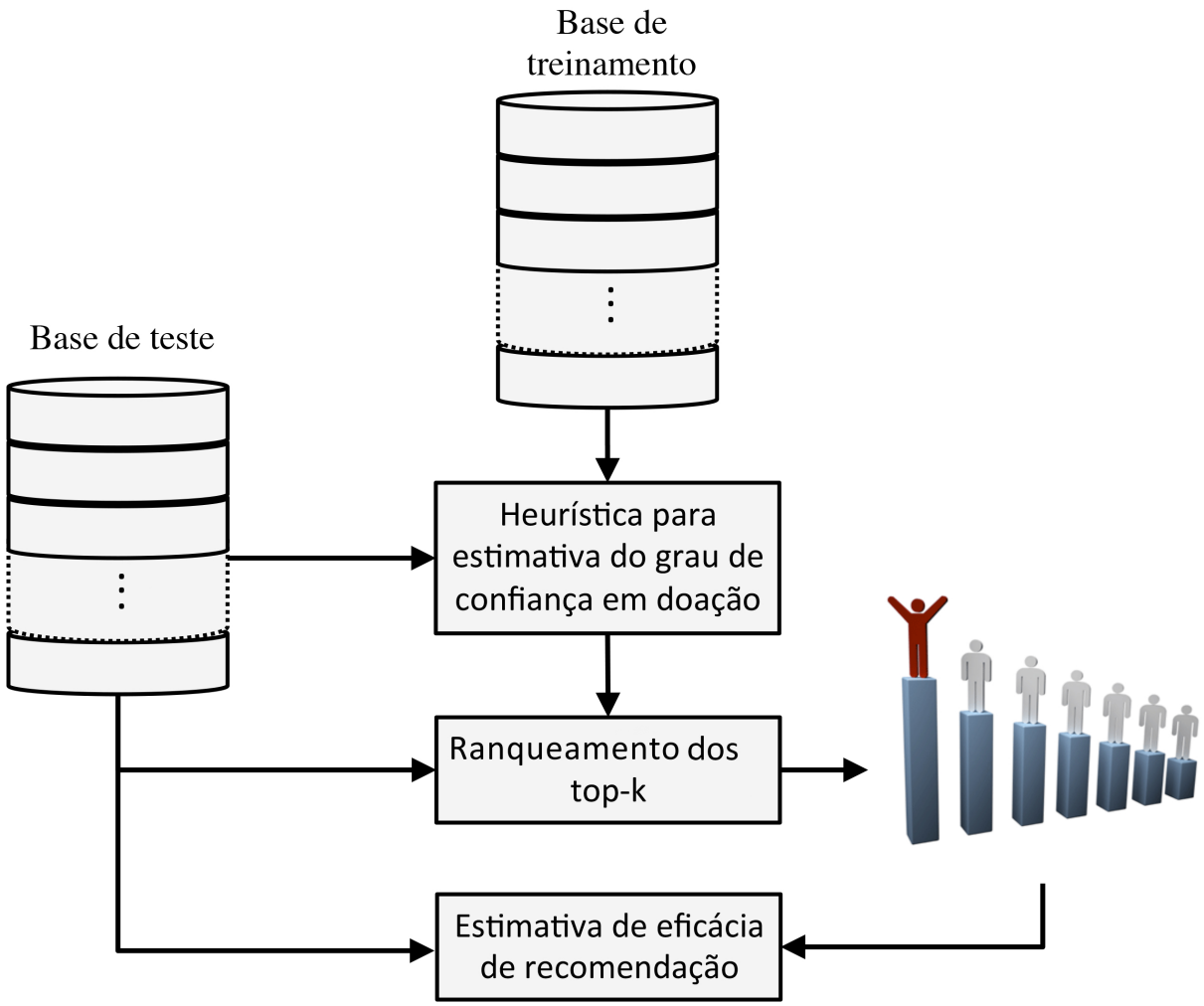
Fonte: Autor.

Como os testes computacionais desta primeira fase de experimentos não resultaram em taxas ideais de classificação optou-se pelo desenvolvimento de uma abordagem para a recomendação dos Top-k, a qual é descrita e analisada nas seções seguintes.

5.5 Recomendação dos Top-k

A abordagem de recomendação dos top-k proposta visa ordenar os itens (potenciais doadores de sangue) conforme um valor de confiança de que estes pertençam à classe doar. A Figura 5.3 ilustra de forma geral a abordagem proposta. Em resumo considera-se uma base de treinamento que é utilizada como referência para a heurística que infere o grau de confiança de dado indivíduo contido no conjunto de teste em pertencer a classe doar. Para estimar tais valores de confiança são propostas várias diferentes heurísticas baseadas em vários critérios. A seguir são definidas tais heurísticas.

Figura 5.3 – Fluxograma da abordagem de recomendação dos top-k proposta.



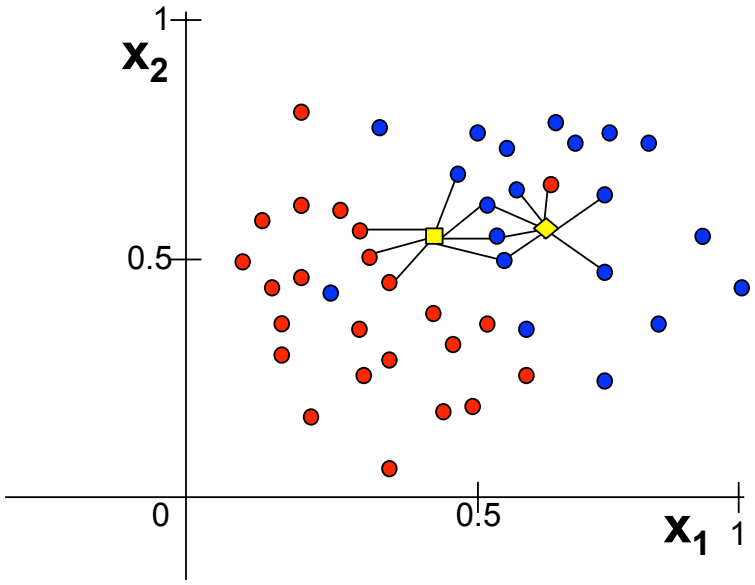
Fonte: Autor.

5.5.1 Heurística baseada nos vizinhos mais próximos

A heurística é dada pela taxa entre os vizinhos mais próximos que pertencem a classe doar. Assim, dada uma amostra de teste, em que devemos decidir sobre sua recomendação, e em caso positivo sobre sua posição na lista dos top-k, calculamos a distância das características desta para as características dos indivíduos do conjunto de treinamento. Em seguida, são verificadas as classes dos indivíduos mais próximos pertencentes ao conjunto de treinamento e calculado a taxa destes indivíduos mais próximos que pertencem a classe doar. A Figura 5.4 ilustra o cálculo da heurística dos vizinhos mais próximos, onde consideramos os pontos representados por círculo como sendo as instância de treinamento, para as quais são conhecidas suas classes. Suponhamos que os círculos azuis representa os indivíduos que doaram sangue e os círculos vermelhos representam os indivíduos que não doaram sangue. Agora considere a execução de uma consulta que retorne os rótulos dos k vizinhos mais próximos. Se considerarmos k=7 e o ponto de consulta como o sendo aquele marcado pelo

losangolo amarelo, temos que seis entre os sete vizinhos mais próximos pertence a classe "doaram sangue". Assim, temos um grau de confiança de 6/7 que este indivíduo em questão doará sangue. Repetindo o mesmo processo para o ponto marcado pelo quadrado amarelo, temos um grau de confiança menor, de 4/7 que este doará sangue. Assim, ordenamos os indivíduos pertencentes ao conjunto de teste em ordem crescente conforme os valores de confiança retornados heurística baseada nos vizinhos mais próximos, para a composição da lista contendo os top-k itens (potenciais doadores, no nosso caso) recomendados.

Figura 5.4 – Ilustação da heurística baseada nos vizinhos mais próximos.



Fonte: Autor.

5.5.2 Heurística baseada no Teorema de Bayes

O Teorema de Bayes pode ser utilizado para calcular probabilidades condicionais. Assim, este pode ser utilizado para decidir a probabilidade de um dado indivíduo doar sangue dadas as suas características. Dada uma amostra (indivíduo), podemos calcular a probabilidade deste doar com base na Equação 5.4.

$$P(\text{'doar'}|\mathbf{x}) = \frac{P(\mathbf{x}|\text{'doar'})P(\text{'doar'})}{P(\mathbf{x})}, \tag{5.4}$$

Na Equação 5.4, $P(\text{'doar'})$ trata-se da probabilidade *a priori* da classe 'doar', $P(\mathbf{x})$ é a probabilidade *a priori* de $\mathbf{x}$ e $P(\mathbf{x}|\text{'doar'})$ é a probabilidade *a posteriori* de $\mathbf{x}$ condicionada a classe 'doar'. As probabilidades $P(\text{'doar'})$, $P(\mathbf{x})$ e $P(\mathbf{x}|\text{'doar'})$ são estimadas a partir das instâncias de treinamento.

Para calcular $P(\mathbf{x}|\text{'doar'})$ na usamos a hipótese de independência entre os atributos. Desta forma considerando que $\mathbf{x}$ é definido por m características $(x_1, x_2, \ldots, x_m)$ usa-se a fór-

mula seguinte.

$$P(\mathbf{x}|\text{'doar'}) = \prod_{i=1}^{n} P(x_i|\text{'doar'}), \tag{5.5}$$

sendo $P(x_i|\text{'doar'})$ estimada a partir dos exemplos de treinamento pela seguinte fórmula:

$$P(x_i|\text{doar}) = \frac{1}{\sigma_{doar}\sqrt{2\pi}} e^{-\frac{(x_k-\mu_{doar})^2}{2\sigma^2_{doar}}}, \tag{5.6}$$

onde μ_{doar} e σ_{doar} são, respectivamente, a média e o desvio padrão dos valores da característica de índice i para os exemplos da classe $doar$.

Neste caso, as amostras $\mathbf{x}$ pertencentes ao conjunto de teste são ordenadas em ordem crescente conforme os valores de $P(\mathbf{x}|\text{'doar'})$, para compor a lista dos top-k recomendados.

5.5.3 Heurística baseada no hiperplano de separação das classes

Nesta heurística é utilizado o classificador SVM citado na seção 4.4 que faz um mapeamento não-linear no espaço de dados e calcula o hiperplano de separação das classes. Com base nos hiperplanos de separação, utilizamos a técnica *Platt Scalling* (4.10) para encontrar a probabilidade de cada amostra do conjunto de teste pertencer a classe doar. A técnica *Platt Scalling* executa uma regressão logística com base no hiperlano de separação do SVM para ajustar os parâmetros de uma função de probabilidade gaussiana, que é usada para retorna as probabilidades de classe para cada amostra (doador candidato) do conjunto de teste. O ajuste de parâmetros desta função de probabilidade é feito considerando as amostras do conjunto de treinamento. No nosso caso específico, a função gaussiana de probabilidade que dever ser ajustada é dada por:

$$P(y = \text{'doar'}|x) = \frac{1}{1+\exp(Af(x)+B)} \tag{5.7}$$

sendo esta uma transformação dos valores obtidos a partir do valor da função do classificador aplicada à amostra de teste, representado por $f(x)$, onde A e B são dois parâmetros escalares que são aprendidos pelo algoritmo. Os parâmetros A e B são estimados usando um método de máxima verossimilhança no mesmo conjunto de treinamento usado pelo classificador original f. Depois de feito o ajuste dos parâmetros A e B com base no conjunto de treinamento, usamos diretamente a Eq. 5.7 para obter os valores de probabilidade de cada amostra de teste pertencer a classe doar. Neste trabalho usamos a biblioteca SKLEARN do Python que fazem a otimização dos parâmetros A e B conforme o conjunto de treinamento e o hiperplano de separação obtido pelo SVM.

5.6 Resultados de recomendação dos top-k

Nos testes realizados os algoritmos foram processados 10.000 vezes retornando a média das probabilidades para cada item afim de reduzir a variação das taxas. Os algoritmos são avaliados na base de dados já mencionada na seção 5.2 com 70% dos itens para treino e 30% para teste.

A escolha dos algoritmos classificadores para o experimento top-k itens se deu pela compatibilidade ao método de analise de confiança do individo pertencer a uma classe, descrito nas heurísticas nas seções anteriores e o fato de serem os classificadores os que tiveram melhor média (entre *accuracy, recall e precision*) nos experimentos realizados.

Os resultados que podem ser vistos na Tabela 5.3 que demonstra as taxas de *precision* para as recomendação de top-1 até top-10. Como pode ser observado a heurística baseada em vetores de suporte, denominada hSVM em referência à *Support Vector Machine*, obteve os melhores resultados de *precision*, sendo estes superiores a 84,53% de *precision* para os top-10 recomendados e chegando a 99,90% para o top-1, seguido pela heurística baseada em vizinhos mais próximos, denominada hNN em referência a *nearest neighbor*, por fim, a baseada no Teorema de *Bayes* denominada hTB.

Tabela 5.3 – Resultados por classificadores *Top-k Recomendation* ($k = 10$).

	hSVM	hNN (k=17)	hTB
Top-1	99,90%	93,96%	92,83%
Top-2	99,45%	93,18%	92,53%
Top-3	98,87%	92,12%	91,74%
Top-4	98,08%	91,35%	90,39%
Top-5	96,52%	90,46%	88,69%
Top-6	94,60%	89,76%	86,81%
Top-7	92,10%	89,07%	85,09%
Top-8	89,70%	88,47%	83,70%
Top-9	87,14%	87,99%	82,98%
Top-10	84,53%	87,55%	82,88%

Fonte: Autor.

5.7 Análise e discussão

Conforme a análise dos resultados percebe-se que a recomendação dos top-k resultaram em alto índice de *precision*, com valores acima de 96% para os top-5 e 84,53% para os

top-10. Tal resultado tem bastante relevância, principalmente em situações de emergência onde se deseja recomendar indivíduos com alta probabilidade de doar sangue. Também a análise de classificadores para predição de doação de sangue, apesar de não ter obtidos resultados satisfatórios, pode ser vista como uma contribuição para a literatura pois nenhum trabalho encontrado na literatura compara, para o contexto de predição de doação de sangue, a gama de classificadores analisados nesta pesquisa.

Capítulo 6

CONCLUSÕES

6.1 Considerações Iniciais

Esta pesquisa focou na experimentação e desenvolvimento de métodos da mineração de dados para predizer doadores de sangue com maior taxa de conversão a doação. Foram estudados os métodos de sistema de recomendação clássicos e experimentados uma ampla gama de algoritmos classificadores em uma base de dados real. Ao todo foram experimentados sete classificadores, porém estes não apresentaram resultados satisfatórios para a aplicação. Dado o resultado insatisfatório dos métodos de classificação, foi proposta uma abordagem de recomendação top-k que se baseia em heurísticas para calcular uma espécie de grau de confiança de que um dado indivíduo irá doar sangue. Tal abordagem proposta, na qual foram analisadas três heurísticas, resultaram em altos índices de precisão, principalmente para os indivíduos próximos ao topo da lista de recomendados.

6.2 Principais Contribuições

Este trabalho contribui para a comunidade acadêmica por experimentar uma gama de algoritmos classificadores para a predição de doação de sangue e principalmente pela proposição da abordagem de recomendação top-k, na qual foi analisada três heurísticas. Além da contribuição ao estado da arte de métodos de predição de doação de sangue, a otimização da busca por doadores focado na maior taxa de conversão a doação pode ser altamente efetiva nas aplicações práticas, sendo de grande auxilio as unidade hemoterápicas na manutenção do estoque sanguíneo e, em situações críticas onde é necessário obter doações de sangue urgentemente.

6.3 Propostas de Trabalhos Futuros

Como o método estabelecido nesta pesquisa tem grande potencial em sua implementação no mundo real, pois consegue garantir alto nível de probabilidade para indicação de doadores, uma próxima etapa consiste de uma investigação sobre plataformas e recursos tecnológicos para colocar a pesquisa em prática. Primeiramente é necessário investigar as melhores plataformas para serem implementadas junto a um hemocentro, incluindo tecnologias de Banco de Dados, Hardwares, aplicativos cliente, e painéis ou terminais de acesso e análise de informações do sistema. Além disso, um aplicativo mobile que auxilia no contato do hemocentro ao doador e também possibilita o usuário solicitar uma doação aos outros usuários com maior propensão a doar pode ser de grande contribuição para fomentar a doação de sangue. Também, podem ser adicionadas outras informações (características de indivíduos) que podem maximizar o desempenho de classificação ou da recomendação de doadores, como distância do hemocentro, idade, sexo, tipo sanguíneo, dentre outros, as quais não estão presentes na base de dados utilizada. Uma possível implementação prática poderia ainda contribuir com a literatura caso colete informações anonimas dos doadores a fim de gerar uma base de dados real com características mais aproximadas a realidade de nosso país ou de uma região específica, e disponibilize esta para a comunidade científica.

REFERÊNCIAS

AGUIAR, K. *et al.* Prevalência de inaptidão sorológica dos doadores de sangue no hemocentro regional de montes claros, minas gerais. *Revista de Pesquisa: Cuidado é Fundamental Online*, v. 8, n. 3, p. 4864–4871, 2016. ISSN 2175-5361. Disponível em: <http://www.seer.unirio.br/index.php/cuidadofundamental/article/view/5305>. Citado na página 23.

BALABANOVIć, M.; SHOHAM, Y. Fab: Content-based, collaborative recommendation. *Commun. ACM*, ACM, New York, NY, USA, v. 40, n. 3, p. 66–72, mar. 1997. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/245108.245124>. Citado na página 39.

BOSTANDJIEV, S.; O'DONOVAN, J.; HÖLLERER, T. Tasteweights: a visual interactive hybrid recommender system. In: ACM. *Proceedings of the sixth ACM conference on Recommender systems*. [S.l.], 2012. p. 35–42. Citado na página 39.

BOWYER, K. W. *et al.* SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011. Disponível em: <http://arxiv.org/abs/1106.1813>. Citado 2 vezes nas páginas 59 e 64.

BRASIL. *Ministério da Saúde, Portaria nº 127, de 8 de Dezembro de 1995. Institui o Programa Nacional de Inspeção em Unidades Hemoterápicas - PNIUH*. Brasília/DF: [s.n.], 1995. Disponível em: <http://bvsms.saude.gov.br/bvs/saudelegis/anvisa/1995/prt0127_08_12_1995.html>. Citado na página 29.

____. *Lei 10.205 de 21 de março de 2001. Regulamenta o § 4º do art. 199 da Constituição Federal*. 2001. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/LEIS_2001/L10205.htm>. Citado na página 29.

____. *Ministério da Saúde, Portaria nº 1.353, de 13 de Junho de 2011. Regulamento Técnico de Procedimentos Hemoterápicos*. Brasília/DF: [s.n.], 2011. Disponível em: <http://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt1353_13_06_2011.html>. Citado na página 33.

____. *Ministério da Saúde, PORTARIA Nº 2.712, DE 12 DE NOVEMBRO DE 2013*. Brasília/DF: [s.n.], 2013. Disponível em: <http://www.hemominas.mg.gov.br/images/doacao_sangue/portaria_2712_de_12_novembro_2013.pdf>. Citado 2 vezes nas páginas 32 e 33.

BREIMAN, L. *et al. Classification and regression trees*. [S.l.]: CRC press, 1984. Citado 2 vezes nas páginas 47 e 48.

BURKE, R. The adaptive web. In: BRUSILOVSKY, P.; KOBSA, A.; NEJDL, W. (Ed.). Berlin, Heidelberg: Springer-Verlag, 2007. cap. Hybrid Web Recommender Systems, p. 377–408. ISBN

978-3-540-72078-2. Disponível em: <http://dl.acm.org/citation.cfm?id=1768197.1768211>. Citado na página 36.

CHOMUTARE, T.; ARSAND, E.; HARTVIGSEN, G. Mobile peer support in diabetes. *Studies in health technology and informatics*, v. 169, p. 48–52, 2010. Citado na página 41.

COLSAN. *COLSAN - Associação Beneficente de Coleta de Sangue - Para quem eu posso doar sangue?* 2016. Disponível em: <http://www.colsan.org.br/site/doador/para-quem-eu-posso-doar-sangue.html>. Citado na página 29.

CONGDON, P. *Bayesian Statistical Modelling*. [S.l.]: Wiley Series in Probability and Statistics, 2006. Citado na página 43.

COVER, T. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14, p. 326–334, 1965. Citado na página 53.

DARWICHE, M. *et al.* Prediction of blood transfusion donation. In: *2010 Fourth International Conference on Research Challenges in Information Science (RCIS)*. [S.l.: s.n.], 2010. p. 51–56. ISSN 2151-1349. Citado na página 24.

DAVIS, P. *Interpolation and Approximation*. 1st. ed. New York, NY, USA: Blaisdell, 1963. Citado na página 55.

DESHPANDE, M.; KARYPIS, G. Item-based top-n recommendation algorithms. *ACM Trans. Inf. Syst.*, ACM, New York, NY, USA, v. 22, n. 1, p. 143–177, jan. 2004. ISSN 1046-8188. Disponível em: <http://doi.acm.org/10.1145/963770.963776>. Citado na página 24.

DUDA, R.; HART, P.; STORK, D. *Pattern Classification and Scene Analysis*. Second edition. [S.l.]: John Wiley & Sons, Inc., 2001. Citado 2 vezes nas páginas 43 e 51.

FADER, P. S.; HARDIE, B. G.; LEE, K. L. Rfm and clv: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, American Marketing Association, v. 42, n. 4, p. 415–430, 2005. Citado na página 62.

FELFERING, A. *et al. Recommender Systems: An Introduction*. [S.l.]: Cambridge University Press, 2011. Citado 2 vezes nas páginas 38 e 39.

GARRAUD, O.; TISSOT, J.-D. Blood donation and/or donated blood acceptance: The different stakeholders' ethical considerations. *Ethics, Medicine and Public Health*, v. 2, n. 2, p. 213 – 219, 2016. ISSN 2352-5525. Disponível em: <http://www.sciencedirect.com/science/article/pii/S235255251630024X>. Citado 2 vezes nas páginas 28 e 31.

GILLESPIE, T. W.; HILLYER, C. D. Blood donors and factors impacting the blood donation decision. *Transfusion Medicine Reviews*, Elsevier, v. 16, n. 2, p. 115–130, 2002. Citado 2 vezes nas páginas 23 e 33.

GOLDBERG, D. *et al.* Using collaborative filtering to weave an information tapestry. *Commun. ACM*, ACM, New York, NY, USA, v. 35, n. 12, p. 61–70, dez. 1992. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/138859.138867>. Citado na página 35.

GUIDDI, P. *et al.* New donors, loyal donors, and regular donors: Which motivations sustain blood donation? *Transfusion and Apheresis Science*, Elsevier, v. 52, n. 3, p. 339–344, 2015. ISSN 1473-0502. Disponível em: <http://dx.doi.org/10.1016/j.transci.2015.02.018>. Citado na página 23.

HAND, D. J.; MANNILA, H.; SMYTH, P. *Principles of data mining*. [S.l.]: MIT press, 2001. Citado na página 47.

HAYKIN, S. *Redes Neurais - 2ed.* BOOKMAN COMPANHIA ED, 2001. ISBN 9788573077186. Disponível em: <https://books.google.com.br/books?id=lBp0X5qfyjUC>. Citado na página 51.

____. *Neural Networks and Learning Machines*. 3rd. ed. [S.l.]: Prentice Hall, 2009. Citado 3 vezes nas páginas 43, 49 e 50.

HERLOCKER, J. L.; KONSTAN, J. A.; RIEDL, J. Explaining collaborative filtering recommendations. In: *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM, 2000. (CSCW '00), p. 241–250. ISBN 1-58113-222-0. Disponível em: <http://doi.acm.org/10.1145/358916.358995>. Citado na página 36.

HSU, W.; LEE, M.; ZHANG, J. Image mining: trends and developments. *Journal of Intelligent Information Systems*, p. 7–23, 2001. Citado na página 50.

JANNACH, D. *et al. Recommender Systems: An Introduction*. 1st. ed. New York, NY, USA: Cambridge University Press, 2010. ISBN 0521493366, 9780521493369. Citado na página 39.

KARYPIS, G. Evaluation of item-based top-n recommendation algorithms. In: *Proceedings of the Tenth International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2001. (CIKM '01), p. 247–254. ISBN 1-58113-436-3. Disponível em: <http://doi.acm.org/10.1145/502585.502627>. Citado na página 24.

KAUR, G.; GUPTA, N. E-health: A new perspective on global health. *Journal of Evolution and Technology*, v. 15, n. 1, p. 23–35, 2006. Citado na página 40.

KAY, M.; SANTOS, J.; TAKANE, M. mhealth: New horizons for health through mobile technologies. *World Health Organization*, v. 64, n. 7, p. 66–71, 2011. Citado na página 40.

KIM, J.-H. *et al.* Design of diet recommendation system for healthcare service based on user information. In: IEEE. *2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*. [S.l.], 2009. p. 516–518. Citado na página 40.

LEE, J. *et al.* Improving the accuracy of top-n recommendation using a preference model. *Information Sciences*, v. 348, p. 290 – 304, 2016. ISSN 0020-0255. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0020025516300524>. Citado na página 24.

LORENA, A. C. *Investigação de estratégias para a geração de máquinas de vetores de suporte multiclasses*. Tese (Tese de Doutorado) — Instituto de Ciências Matemáticas e de Computação (ICMC/USP), 2006. Citado na página 50.

LYLE, H. F. I.; SMITH, E. A.; SULLIVAN, R. J. Blood donations as costly signals of donor quality. *Journal of Evolutionary Psychology*, v. 7, n. 4, p. 263–286, 2009. Disponível em: <http://dx.doi.org/10.1556/JEP.7.2009.4.1>. Citado na página 33.

MELVILLE, P.; MOONEY, R. J.; NAGARAJAN, R. Content-boosted collaborative filtering for improved recommendations. In: *Aaai/iaai*. [S.l.: s.n.], 2002. p. 187–192. Citado na página 39.

Ministério da Saúde. *Hemocentros iniciam estratégia especial para os Jogos Olímpicos*. 2016. Disponível em: <http://portalsaude.saude.gov.br/index.php/cidadao/principal/agencia-saude/24981-hemocentros-iniciam-estrategia-especial-para-os-jogos-olimpicos>. Citado na página 34.

MITRA, S.; ACHARYA, T. *Data Mining: Multimedia, Soft Computing and Bioinformatics*. [S.l.]: John Wiley & Sons, 2003. Citado na página 49.

MOGHADDASI, H. *et al.* E-health: a global approach with extensive semantic variation. *Journal of medical systems*, Springer, v. 36, n. 5, p. 3173–3176, 2012. Citado na página 40.

NIEMINEN, P. Multilayer perceptron training with multiobjective memetic optimization. *Jyväskylä studies in computing 247.*, 2016. Citado 2 vezes nas páginas 51 e 52.

OLIVEIRA, G. C. d. *Plasma humano : componentes e derivados : conservação e utilização terapêutica em ambiente hospitalar*. Dissertação (Mestrado) — Instituto Superior de Ciências da Saúde Egas Moniz, Março 2016. Citado na página 27.

OUHBI, S. *et al.* Free blood donation mobile applications. *Journal of medical systems*, Springer, v. 39, n. 5, p. 1–20, 2015. Citado na página 41.

PERME, M.; BLAS, M.; TURK, S. Comparison of logistic regression and linear discriminant analysis: A simulation study. v. 1, p. 143–161, 01 2004. Citado na página 59.

PINHO, A. *et al.* Triagem clínica de doadores de sangue. *Brasília: Ministério da Saúde, Coordenação Nacional de Doenças Sexualmente Transmissíveis e AIDS*, 2001. Citado 2 vezes nas páginas 31 e 32.

PLATT, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. v. 10, 06 2000. Citado 2 vezes nas páginas 59 e 60.

PORTAL BRASIL. *Hemocentros adotam estratégia para os Jogos Olímpicos*. 2016. Disponível em: <http://www.brasil.gov.br/saude/2016/08/hemocentros-adotam-estrategia-para-os-jogos-olimpicos>. Citado na página 33.

POWELL, M. Radial basis function approximations to polynomials. In: *Numerical Analysis 1987 Proceedings*. Dundee, UK: [s.n.], 1988. p. 223–241. Citado na página 55.

PRADHAN, S.; GAY, V.; NEPAL, S. Improving the matching process of dental care recommendation systems by using subjective criteria for both patients and dentists. *PACIS 2014 Proceedings. Paper 296.*, AISEL, 2014. Citado na página 40.

QUINLAN, J. R. Induction of decision trees. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 1, n. 1, p. 81–106, mar. 1986. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A:1022643204877>. Citado na página 46.

____. *C4.5: Programs for machine learning*. San Francisco, USA: Morgan Kaufmann, 1993. Citado 2 vezes nas páginas 43 e 46.

RAZOUK, F. H.; REICHE, E. M. V. Caracterização, produção e indicação clínica dos principais hemocomponentes. *Revista Brasileira de Hematologia e Hemoterapia*, scielo, v. 26, p. 126 – 134, 00 2004. ISSN 1516-8484. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-84842004000200011&nrm=iso>. Citado na página 31.

RESNICK, P.; VARIAN, H. R. Recommender systems. *Commun. ACM*, ACM, New York, NY, USA, v. 40, n. 3, p. 56–58, mar. 1997. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/245108.245121>. Citado na página 35.

RICCI, F.; ROKACH, L.; SHAPIRA, B. *Introduction to recommender systems handbook*. [S.l.]: Springer, 2011. Citado na página 39.

RINGWALD, J.; ZIMMERMANN, R.; ECKSTEIN, R. Keys to open the door for blood donors to return. *Transfusion Medicine Reviews*, Elsevier, v. 24, n. 4, p. 295–304, 2010. Citado na página 34.

RODRIGUES, R. S. M.; REIBNITZ, K. S. Estratégias de captação de doadores de sangue: uma revisão integrativa da literatura. *Texto & Contexto - Enfermagem*, scielo, v. 20, p. 384 – 391, 06 2011. ISSN 0104-0707. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-07072011000200022&nrm=iso>. Citado 2 vezes nas páginas 27 e 29.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *nature*, Nature Publishing Group, v. 323, n. 6088, p. 533, 1986. Citado na página 52.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3rd. ed. [S.l.]: Prentice Hall, 2009. Citado na página 38.

SANTHANAM, T.; SUNDARAM, S. Application of cart algorithm in blood donors classification. *Journal of computer Science*, Citeseer, v. 6, n. 5, p. 548, 2010. Citado 2 vezes nas páginas 24 e 62.

SCHAFER, J. B.; KONSTAN, J. A.; RIEDL, J. E-commerce recommendation applications. *Data Min. Knowl. Discov.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 5, n. 1-2, p. 115–153, jan. 2001. ISSN 1384-5810. Disponível em: <http://dx.doi.org/10.1023/A:1009804230409>. Citado na página 37.

SCHÖLKOPF, B.; SMOLA, A. J. *Learning with Kernels*. [S.l.]: MIT Press, 2002. Citado na página 43.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, ACM, v. 34, n. 1, p. 1–47, 2002. Citado na página 37.

SHANNON, C. A mathematical theory of communication. *The Bell Systems Technical Journal*, v. 27, n. 1, p. 379–423, 1948. Citado na página 45.

SHARDANAND, U.; MAES, P. Social information filtering: Algorithms for automating “word of mouth”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1995. (CHI '95), p. 210–217. ISBN 0-201-84705-1. Disponível em: <http://dx.doi.org/10.1145/223904.223931>. Citado na página 35.

SHIH, Y.-Y.; LIU, D.-R. Hybrid recommendation approaches: collaborative filtering via valuable content information. In: IEEE. *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*. [S.l.], 2005. p. 217b–217b. Citado na página 39.

SILVA, G. E. d. M.; VALADARES, G. V. Conhecendo os meandros da doação de sangue: Implicações para a atuação do enfermeiro na hemoterapia. *Revista Brasileira de Enfermagem*, scielo, v. 68, p. 32 – 39, 02 2015. ISSN 0034-7167. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-71672015000100032&nrm=iso>. Citado 2 vezes nas páginas 23 e 33.

SU, X.; KHOSHGOFTAAR, T. M. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, Hindawi Publishing Corp., v. 2009, p. 4, 2009. Citado na página 37.

SUN, T.; LU, S. F.; JIN, G. Z. Solving shortage in a priceless market: Insights from blood donation. *Journal of Health Economics*, v. 48, p. 149 – 165, 2016. ISSN 0167-6296. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167629616300224>. Citado na página 27.

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition*. New York, USA: Academic Press, 1999. Citado na página 48.

VANE, L. A.; GANEM, E. M. Doação homóloga versus autóloga e substitutos da hemoglobina. *Medicina Perioperatória*, Sociedade de Anestesiologia do Estado do Rio de Janeiro, p. 292–306, 2006. Citado na página 32.

VAPNIK, V. N. *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8. Citado na página 49.

WANG, Y.; WANG, X.-J. A new approach to feature selection in text classification. In: IEEE. *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*. [S.l.], 2005. v. 6, p. 3814–3819. Citado na página 37.

WING, C.; YANG, H. Fityou: Integrating health profiles to real-time contextual suggestion. In: *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2014. (SIGIR '14), p. 1263–1264. ISBN 978-1-4503-2257-7. Disponível em: <http://doi.acm.org/10.1145/2600428.2611185>. Citado na página 40.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2005. Citado na página 44.

World Health Organization. Towards 100% voluntary blood donation: a global framework for action. Geneva: World Health Organization, 2010. Disponível em: <http://www.who.int/iris/handle/10665/44359>. Citado 3 vezes nas páginas 23, 27 e 63.

YEH, I.-C.; YANG, K.-J.; TING, T.-M. Knowledge discovery on rfm model using bernoulli sequence. *Expert Systems with Applications*, Elsevier, v. 36, n. 3, p. 5866–5871, 2009. Citado 2 vezes nas páginas 61 e 62.

Z., M.; M.R., J.; M., a. M. A comparison of seroprevalence of blood-borne infections among regular, sporadic, and first-time blood donors in isfahan. *The Scientific Journal of Iranian Blood Transfusion Organization*, v. 2, n. 7, 2006. Disponível em: <http://bloodjournal.ir/article-1-48-en.html>. Citado na página 23.