UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Identificação de proteínas associadas ao envelhecimento usando aprendizado de máquina

Denilson Antonio Marques Junior

São Carlos – SP

Identificação de proteínas associadas ao envelhecimento usando aprendizado de máquina

Denilson Antonio Marques Junior

Orientador: **Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho**

Monografia final de conclusão de curso apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como requisito parcial para obtenção do título de Engenheiro de Computação.
Área de Concentração: Inteligência Computacional, Aprendizado de Máquina

USP – São Carlos
Novembro de 2017

Junior, Denilson Antonio Marques
Identificação de proteínas associadas ao envelhecimento usando aprendizado de máquina / Denilson Antonio Marques Junior. - São Carlos - SP, 2017.
49 p.; 29,7 cm.

Orientador: André Carlos Ponce de Leon Ferreira de Carvalho.
Monografia (Graduação) - Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos - SP, 2017.

1. Aprendizado de Máquina. 2. Classificação. 3. Bioinformática. I. Carvalho, André Carlos Ponce de Leon Ferreira de. II. Instituto de Ciências Matemáticas e de Computação (ICMC/USP). III. Título.

RESUMO

MARQUES JUNIOR, D. A.. **Identificação de proteínas associadas ao envelhecimento usando aprendizado de máquina** . 2017. 49 f. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Há estudos na literatura que empregam aprendizado de máquina para prever a classificação de genes em ligados ou não ao envelhecimento. O tipo de informação utilizada nesse problema para se criar atributos varia, mas dois exemplos são informações de interações entre proteínas e papéis biológicos dos genes. Neste projeto, investiga-se a combinação desses dois tipos de informação para se treinar classificadores. Ademais, também estuda-se como a variação do algoritmo de aprendizado, técnica de seleção de atributos e número de atributos selecionados influencia do desempenho final dos classificadores treinados.

Palavras-chave: Aprendizado de Máquina, Classificação, Bioinformática.

ABSTRACT

MARQUES JUNIOR, D. A.. **Identificação de proteínas associadas ao envelhecimento usando aprendizado de máquina** . 2017. 49 f. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Some studies in the literature use machine learning to predict whether genes are or not linked to the aging process. The type of information used to create features varies, but two examples are protein protein interaction information and information about the biological role of each gene. This project investigates the combination of these two types of information in the task of training classifiers. Furthermore, it investigates how the variation of the learning algorithm, attribute selection technique and number of selected attributes affects the final performance of the trained classifiers.

Key-words: Machine Learning, Classification, Bioinformatics.

LISTA DE ILUSTRAÇÕES

Figura 1 – Incidência de doenças como função da idade em humanos. Adaptado de (MAGALHÃES, 2011). 19

Figura 2 – Estimativa do erro de generalização em função do número de amostras utilizadas no conjunto de treino. Foi utilizado o gerador de datasets correspondente a função *make_hastie_10_2* da biblioteca *sklearn* para gerar tanto o conjunto de treino quanto o de teste. A curva mostra a média do erro num conjunto de teste de 20.000 amostras para um classificador *random forest*. Embora ruidosa, a curva tem uma tendência clara: mais exemplos implica menor erro de generalização. 25

Figura 3 – Média e variância das acurácias amostradas pela validação cruzada de 10 vias, para modelos que utilizaram RelieF no *dataset* aging_go. 37

Figura 4 – Média e desvio padrão das acurácias amostradas pela validação cruzada de 10 vias, para modelos que utilizaram os 1000 atributos sugeridos pelo algoritmo ReliefF. 37

Figura 5 – Relevância dos atributos selecionados segundo XGBoost, para o *dataset* aging_go. 38

Figura 6 – Relevância dos atributos selecionados segundo *Random Forest*, para o *dataset* aging_go. 39

Figura 7 – Relevância dos atributos selecionados segundo XGBoost, para o *dataset* aging_complex. 39

Figura 8 – Relevância dos atributos selecionados segundo *Random Forest*, para o *dataset* aging_complex. 40

Figura 9 – Relevância dos atributos selecionados segundo XGBoost, para o *dataset* aging_complex. 41

LISTA DE TABELAS

Tabela 1 – Nomes dos atributos de medidas topológicas 34
Tabela 2 – Resultados dos experimentos realizados da combinação entre algoritmos de AM, métodos de seleção de SA, número de atributos selecionados e *dataset*. Células vazias indicam que um experimento não foi realizado. Por exemplo, não faz sentido usar algum método de SA para selecionar todos os atributos. Cada célula contém a média e o erro padrão da acurácia resultante do experimento usando validação cruzada de 10 vias. Cores claras indicam médias altas, enquanto cores escuram indicam médias baixas. 42

LISTA DE ABREVIATURAS E SIGLAS

AM ..... Aprendizado de Máquina
CART .... *Classification and Regression Trees*
CR ..... *Caloric Restriction*
GBM ..... *Gradient Boosted Trees*
GO ..... *Gene Ontology*
PPI ..... *Protein protein interaction*
RF ..... *Random Forest*
SA ..... Seleção de Atributos
SVM ..... *Support Vector Machine*
XGBoost . *Extreme Gradient Boosting*

SUMÁRIO

1 INTRODUÇÃO 15
1.1 Motivação e Contextualização 15
1.2 Objetivos 16
1.3 Organização 16

2 MÉTODOS, TÉCNICAS E CONCEITOS 19
2.1 Envelhecimento 19
2.2 Aprendizado de Máquina 20
2.2.1 Aprendizado supervisionado *21*
2.2.2 Algoritmos de aprendizado *21*
2.2.3 Seleção de atributos *23*
2.2.4 Validação de modelos *25*
2.3 Redes Complexas 27
2.4 Trabalhos similares 28

3 DESENVOLVIMENTO 31
3.1 Atividades Realizadas 31
3.1.1 Coleta dos Dados *31*
3.1.2 Preparação dos dados *32*
3.1.3 Preparação dos experimentos *34*
3.2 Resultados experimentais 36
3.2.1 Variações de métodos *36*
3.2.2 Análise de atributos *38*
3.2.3 Dificuldades e limitações *41*

4 CONCLUSÃO 45
4.1 Contribuições 45
4.2 Considerações sobre o Curso de Graduação 46

REFERÊNCIAS 47

Capítulo 1

INTRODUÇÃO

1.1 Motivação e Contextualização

Nas últimas décadas, o aumento na expectiva média de vida de humanos tem aumentado dramaticamente, devido aos avanços na área de nutrição e na área médica, o que tem possibilitado a identificação e subsequente cura de doenças com eficácia cada vez maior. No entanto, isso também significa que mais pessoas chegam à senescência, além de passarem mais tempo nessa fase da vida. Como consequência, uma maior pressão é colocada sobre o Estado devido a programas de segurança social como a aposentadoria, e devido a custos médicos, visto que nessa fase da vida há uma maior incidência de doenças.

Com o avanço na tecnologia, cada vez mais teremos o poder de tomar em nossas mãos as rédeas de nossos destinos e prolongar a vida humana. Contanto, considerando os problemas citados anteriormente, surge a dúvida: devemos? Uma das maneiras de evitar os problemas é não limitar-se a estender a expectativa de vida, mas também o prolongar período saudável da vida de uma pessoa, isto é, efetivamente adiar o envelhecimento. De fato, recentemente têm surgido pesquisas com esse foco (GOLDMAN *et al.*, 2013). Uma maneira de elucidar melhor a diferença entre o aumento da longevidade e o adiamento do envelhecimento é por meio do mito de Tithonus. Segundo a lenda grega, Zeus teria concedido a Tithonus vida eterna, mas não juventude eterna. Com o tempo, Tithonus teria ficado cada vez mais fraco e debilitado conforme envelhecia, eventualmente se tornando incapaz de se mover.

Estimativas sugerem que o valor econômico de adicionar apenas 2,2 anos saudáveis à vida de humanos seria de 7,1 trilhões de dólares ao longo de 50 anos, apenas nos Estados Unidos (GOLDMAN *et al.*, 2013). A razão econômica para se adiar o envelhecimento, portanto, existe. Ademais, nota-se que o fascínio do Homem pela vida eterna não é recente. Há muitas lendas e mitos na história que revelam isso, como o próprio mito de Tithonus citado, além da pedra filosofal, e da fonte da juventude. As razões para se prolongar a vida saudável vão além da razão econômica, e tocam nos desejos individuais das pessoas de viverem mais do que a natureza lhes permitiria na ausência de intervenções artificiais.

Há estudos que sugerem que há razões genéticas para o envelhecimento (MAGALHÃES, 2011). Logo, é interessante determinar qual a relação entre genes e envelhecimento, e mais especificamente ainda determinar quais genes estão de fato ligados ao envelhecimento. Uma das maneiras de fazer isso é por meio de experimentos em laboratórios, mas esse método é

acompanhado da desvantagem de ser caro e lento (FABRIS; MAGALHÃES; FREITAS, 2017). Experimentos computacionais são comparativamente mais baratos, e podem ser empregados para afunilar a busca por genes ligados ao envelhecimento. Este é um dos objetivos deste projeto, mais especificamente por meio da utilização de aprendizado de máquina, que permite a extração automática de conhecimento.

Por fim, encontra-se na literatura tanto estudos que utilizam informações sobre os papéis biológicos de cada genes para prever sua ligação ou não a envelhecimento (FREITAS; VASIEVA; MAGALHÃES, 2011), quanto estudos que utilizam informações sobre as interações em proteínas para prever essa mesma variável (LI; ZHANG; GUO, 2010). Isso levanta a questão de se a combinação desses dois tipos de informação poderia eventualmente levar a modelos com capacidade preditiva maior do que modelos que utilizam apenas um tipo ou outro de informação. Este projeto também tem como objetivo investigar isso.

1.2 Objetivos

O objetivo do projeto é investigar a aplicação de aprendizado de máquina à tarefa de classificação de genes entre genes ligados ao processo de envelhecimento ou não. Por meio da escolha de alguns algoritmos de aprendizado que permitem interpretação do modelo gerado, espera-se também chegar a conclusões acerca de quais são os fatores realmente importantes na tarefa de classificação proposta, em particular. Os objetivos específicos são:

- Revisão bibliográfica sobre aprendizado de máquina;
- Estudar os efeitos da variação de métodos de seleção de atributos, número de atributos selecionados e algoritmo de aprendizado no poder preditivo do modelo elaborado;
- Avaliar a relevância dos tipos de atributos considerados na tarefa de classificação de genes;
- Comparação dos resultados obtidos com resultados encontrados na literatura;
- Combinar informações de diversas fontes sobre os genes humanos em *datasets*, que serão utilizados no projeto.

1.3 Organização

O texto que descreve o trabalho realizado está organizado da seguinte maneira:

- **Capítulo 2**: Apresenta-se aqui alguns dos principais conceitos de aprendizado de máquina que serão relevantes para o desenvolvimento do trabalho. Há também breves discussões sobre redes complexas e o entendimento atual sobre o processo de envelhecimento;

- **Capítulo 3**: Aqui são descritos os passos tomados na coleta e preparação dos dados, bem como os experimentos realizados e seus subsequentes resultados;
- **Capítulo 4**: Neste capítulo são reiterados os principais pontos; apresentados durante o projeto e apresenta-se uma breve discussão sobre o curso da Engenharia da Computação.

Capítulo 2

MÉTODOS, TÉCNICAS E CONCEITOS

2.1 Envelhecimento

Ao se discutir sobre envelhecimento, é importante diferenciar o adiamento do envelhecimento do aumento da longevidade. Magalhães (2011) define envelhecimento como a deterioração progressiva das funções fisiológicas, acompanhado de um aumento de mortalidade e vulnerabilidade com a idade. Isso significa que adiar o envelhecimento como um todo leva a um aumento na longevidade, mas um aumento na longevidade - por meio do tratamento de doenças individualmente, por exemplo - não implica adiar o envelhecimento. Na Figura 1 podemos observar como a incidência de doenças varia ao longo da vida de uma pessoa.

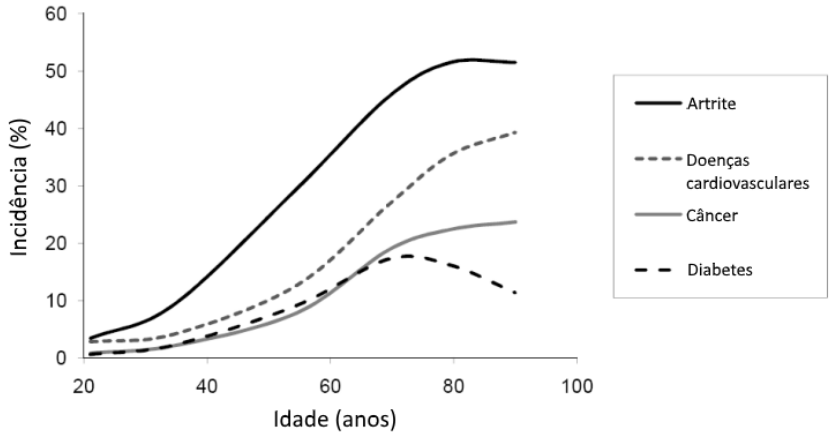

Figura 1 – Incidência de doenças como função da idade em humanos. Adaptado de (MAGALHÃES, 2011).

Há muito tempo tenta-se explicar quais são as razões do envelhecimento. De fato, Aristóteles escreveu o tratado "Longevidade e a Brevidade da Vida", no qual trata desse exato assunto. Infelizmente, apesar dos esforços feitos e das várias teorias propostas, ainda hoje os mecanismos exatos do envelhecimento permanecem um mistério (KENYON, 2010). Mesmo dada essa nossa falta de entendimento, pelo menos em algumas espécies certas intervenções no sentido de prolongar a vida e adiar o envelhecimento têm demonstrado sucesso. A primeira intervenção bem sucedida foi feita durante a época da crise de 1929, quando temia-se que a fome diminuisse a longevidade da população. Experimentos em ratos demonstraram que ao se reduzir

as calorias consumidas mas mantendo os níveis de nutrientes na dieta, a longevidade dos animais aumentava (KENYON, 2010). Mais tarde, demonstrou-se também que esse tipo de dieta também adia o envelhecimento. Essa intervenção na dieta é conhecida como restrição calórica (CR, do inglês *caloric restriction*).

Apesar de suas vantagens, CR também tem efeitos negativos: ratos se demonstraram mais suscetíveis a infecções, além de terem ficado menores que outros ratos que não passaram por CR. Em humanos, não há estudos de longo-termo sobre os efeitos de CR, mas sabe-se que essa intervenção tem pelo menos como efeitos colaterais disfunção sexual e estresse mental. Considerando isso, ainda não se pode afirmar nada sobre a efetividade de CR em humanos. Além disso, até hoje não se conhece nenhuma maneira comprovadamente capaz de adiar, mesmo que só um pouco, o envelhecimento em humanos (MAGALHÃES, 2011).

Sobre as causas do envelhecimento, múltiplas teorias já foram propostas. Aqui a discussão será restrita a apenas algumas poucas, apenas para propósito de ilustração, e recomenda-se ao leitor interessado a discussão feita por Magalhães (2011). Uma das teorias mais conhecidas é a teoria dos radicais livres, que propõe que o envelhecimento é causado pelo dano de radicais livres. Radicais livres são moléculas altamente reativas produzidas pela mitocôndria e que danificam todo tipo de componente celular (MAGALHÃES, 2011). No entanto, há evidências que contradizem essa teoria: ratos modificados para produzirem quantidades menores de enzimas antioxidantes, apesar de realmente apresentarem mais dano oxidativo com a idade, não apresentaram sintomas de envelhecimento acelerado (MAGALHÃES, 2011). Outra teoria é de que o envelhecimento é consequência do dano acumulado no DNA. Há também evidências contra essa teoria também, pois prolongar a vida de ratos em laboratórios por meio da otimização ou melhora nos mecanismos de reparo de DNA se mostrou difícil (MAGALHÃES, 2011).

Curiosamente, entretanto, a síndrome de Werner é causada pela mutação de um único gene específico (WRN) ligado ao reparo de DNA. Portadores dessa síndrome apresentam prematuramente muitas características típicas da senescência, como cabelos brancos, pele envelhecida, diabetes, osteoporose, etc. Há outras síndromes que são resultado de mutações nos genes e que causam envelhecimento prematuro, como a síndrome de Cockayne e a síndrome de Hutchinson-Gilford (MAGALHÃES, 2011). Por meio dessas síndromes e da descoberta que mutações em únicos genes podem levar a aumentos na longevidade (MAGALHÃES, 2011), fica claro que o processo de envelhecimento tem um componente genético.

2.2 Aprendizado de Máquina

Várias definições para aprendizado de máquina podem ser encontradas na literatura. A título de exemplo, aprendizado de máquina é, segundo Mitchell (1997), o campo que estuda a questão de como construir programas de computador capazes de melhorar automaticamente por meio de experiência (tradução livre). Em outras palavras, o objetivo é ser capaz de, a partir de

dados, extrair alguma forma conhecimento. Faceli *et al.* (2011) divide as tarefas de aprendizado, ou extração de conhecimento, em duas modalidades: o aprendizado supervisionado e o não supervisionado.

No aprendizado supervisionado, tem-se um conjunto de treino, que é um conjunto de exemplos que possuem atributos e uma saída, e a partir desse conjunto elabora-se um modelo, também chamado na literatura de função ou hipótese (FACELI *et al.*, 2011), capaz de prever a saída de novos exemplos. No aprendizado não supervisionado, por outro lado, no conjunto de dados não há atributos de saída para cada exemplo - daí o nome: não supervisionado - e portanto o que se faz é agrupar exemplos de acordo com sua similaridade, encontrar uma descrição simples e compacta para o conjunto de dados ou então encontrar padrões frequentes de associações entre os atributos do conjunto (FACELI *et al.*, 2011).

É interessante levantar a questão de porque usar máquinas ao invés de humanos nessas classes de problemas, que já humanos se demonstram notavelmente capazes de aprender e aplicar conhecimento obtido com experiências prévias. É possível encontrar uma breve discussão sobre isso em (FACELI *et al.*, 2011), no qual os autores argumentam que por vezes uma tarefa necessita ser realizada num volume demasiadamente grande ou então a quantidade de informações necessárias para realizá-la torna-a demasiadamente complexa, inviabilizando sua execução por humanos.

2.2.1 Aprendizado supervisionado

Como definido anteriormente, no aprendizado supervisionado elabora-se um modelo capaz de predizer a saída a partir de atributos de um exemplo, a partir apenas de exemplos que o algoritmo já viu anteriormente e que possuiam tanto atributos quanto a saída correta. No entanto, dependendo do tipo de saída, categórica ou contínua, que se espera no problema, classifica-se problemas supervisionado em duas categorias: classificação e regressão.

Consideremos que $y = f(\mathbf{x})$, onde $\mathbf{x}$ é o conjunto de atributos de um exemplo, organizados num vetor, y é a saída do exemplo e $f(\mathbf{x})$ é nosso modelo. Em problemas de classificação, a saída pode apenas assumir valores de um conjunto, ou seja, $y = f(\mathbf{x}) \in \{c_1, c_2, ..., c_N\}$ (FACELI *et al.*, 2011). Para problemas de regressão, por outro lado, a saída pode assumir qualquer valor de um conjunto infinito e ordenado de valores, ou seja, $y = f(\mathbf{x}) \in \mathbb{R}$.

2.2.2 Algoritmos de aprendizado

Neste estudo utiliza-se quatro classificadores diferentes: um algoritmo de indução de árvores de classificação e regressão (algoritmo CART, do inglês *classification and regression trees*), dois algoritmos que constroem um comitê (*ensemble*) desses algoritmos, RF, do inglês *random forests* e XGBoost, do inglês *extreme gradient boosting*, e um algoritmo baseado em aprendizado estatístico, máquinas de vetores de suporte (SVC, do inglês *support vector*

classifiers). Para os algoritmos de aprendizado de máquina CART, RF e SVC, foram utilizadas suas implementações existentes na biblioteca (PEDREGOSA *et al.*, 2011), na linguagem *Python*. No caso do XGBoost, foi utilizada a biblioteca de *Python* homônima, disponível no *website* da biblioteca. A seguir apresentamos uma breve discussão sobre o funcionamento de cada um dos algoritmos, além de algumas de suas vantagens e desvantagens.

No algoritmo CART, a ideia fundamental é particionar recursivamente o espaço de atributos preditivos, de tal maneira que cada partição resultante corresponda a uma saída (HASTIE; TIBSHIRANI; FRIEDMAN, 2013). Em problemas de regressão, em cada partição toma-se como saída a média da saída dos exemplos de treinamento que se encontram naquela região, enquanto em problema de classificação assume-se que a saída de cada partição é a classe que corresponde a maioria dos exemplos do treinamento que nela se encontram. Uma maneira mais natural de se enunciar CART é a construção de uma árvore binária (há métodos baseados em árvore aceitam árvores n-árias, como o C4.5, vide (PRESS, 2009)), de tal maneira que em cada nó há uma regra que diz se os exemplos que por ele passam irão para o filho esquerdo ou para o filho direito. Todos os exemplos começam a classificação no nó raiz da árvore, e os nós folhas correspondem às partições previamente mencionadas.

Visto que a busca pela árvore que separa os dados de maneira ótima é um problema NP-completo, a construção da árvore no CART é feita de maneira gulosa: inicia-se com o nó raiz, e então repete-se o processo de dividir cada nó folha em dois novos usando o critério que maximiza alguma medida da qualidade da divisão, como por exemplo o índice Gini, entropia cruzada (tradução livre de *cross-entropy*) ou erro de classificação (supondo que a classe correta é a majoritária no nó, e o erro é a porcentagem de amostras com classe diferente da correta) (HASTIE; TIBSHIRANI; FRIEDMAN, 2013). Esse processo para quando não há mais nós para dividir, por exemplo quando todos os nós folha já estão homogêneos em termos de classe, quando não há nenhum nó com uma quantidade de observações maior do que um valor determinado, ou então quando já foi atingida uma profundidade máxima pré-determinada.

Uma das maiores vantagens do CART é sua interpretabilidade. O conceito de classificação usando uma árvore de regras é bastante intuitivo, e, contanto que a árvore gerada não seja demasiadamente grande, permite a humanos que entendam o conhecimento extraído do conjunto de dados pelo CART. Há diversas desvantagens no uso de CART, mas uma das mais interessantes de serem mencionadas é que as árvores produzidas pelo algoritmo são modelos com alta variância (HASTIE; TIBSHIRANI; FRIEDMAN, 2013) - uma discussão sobre variância e viés de modelos pode ser encontrada tanto em (HASTIE; TIBSHIRANI; FRIEDMAN, 2013) quanto em (FACELI *et al.*, 2011).

Random forest é um algoritmo que emprega um conjunto de árvores para prever saídas. No caso de uma *random forest* utilizada como classificador, cada árvore que compõe o modelo prevê a classe do exemplo sendo classificado, e a saída do modelo como um todo é então a classe mais votada pelas árvores individuais. Cada árvore é construída usando apenas um subconjunto

dos exemplos do conjunto de treino, e cada divisão de um nó folha em dois novos nós folhas durante o processo de construção da árvore considera apenas um subconjunto dos atributos existentes. Esse algoritmo está fundamentado na ideia de reduzir a variância do modelo final em relação a cada árvore individual por meio da extração da média das previsões individuais das árvores.

O algoritmo XGBoost é baseado em *boosted trees* (GBM) (CHEN; GUESTRIN, 2016). O GBM é parecido com o *random forest* no sentido de usar um comitê de árvores para dar sua previsão final, mas difere na construção do comitê. Nesse algoritmo, adiciona-se novas árvores iterativamente ao comitê, e cada nova árvore é construída de forma a minimizar a função objetivo definida (CHEN; GUESTRIN, 2016), e portanto sua construção é dependente de todas as árvores adicionadas anteriormente, ao contrário de *random forest*. Segundo Chen e Guestrin (2016), XGBoost, em termos de elaboração de um modelo, diferente de GBM principalmente na forma como propõe regras para dividir nós enquanto constrói novas árvores. Além disso, a implementação do XGBoost é modificada em relação ao GBM a fim de melhorar a utilização dos recursos computacionais disponíveis para o algoritmo.

Por fim, o algoritmo SVM consiste, segundo Vapnik (1995), em essencialmente achar o melhor hiperplano que separa as classes, embora de uma maneira diferente da regressão logística. No entanto, nem sempre é possível separar linearmente as classes do conjunto de dados. Por essa razão, em conjunto com esse algoritmo frequentemente se utiliza a técnica de mapear as amostras do espaço de atributos original para outro espaço através da utilização de *kernels*, e então achar o hiperplano ótimo separador nesse segundo espaço. SVMs se destacam entre os algoritmos de AM por ter uma boa capacidade de generalização e por sua robustez a alta dimensionalidade, mas tem a desvantagem de gerar modelos pouco interpretáveis (FACELI *et al.*, 2011).

2.2.3 Seleção de atributos

No passado, poucos domínios explorados usavam mais de 40 atributos (GUYON; ELISSEEFF, 2003). No entanto, hoje em dia há domínios no qual se podem encontrar problemas que usam dezenas de milhares de atributos, muitas vezes para poucas amostras. Exemplos disso incluem este estudo, no qual foram considerados mais de 10.000 atributos para cada instância, problemas de classificação de texto e a área da bioinformática (GUYON; ELISSEEFF, 2003), na qual este estudo se insere. Esse fato levanta a questão de se é realmente necessária essa quantidade de atributos para o sucesso de um algoritmo de aprendizado, ou se a eliminação de atributos traria vantagens.

De fato, segundo Guyon e Elisseeff (2003), há muitos benefícios em potencial na execução de seleção de atributos: facilitar a visualização e entendimento dos dados, reduzir a necessidade de espaço em disco para se armazenar os dados, reduzir os tempos de treinamento e utilização e - nas palavras dos autores - desafiar a maldição da alta dimensionalidade a fim de melhorar a capacidade de previsão dos algoritmos de aprendizado.

Quanto ao método em si de realizar a seleção, os autores separam as possíveis abordagens em dois tipos possíveis: construir e achar o subconjunto de atributos que são de fato úteis para construir um bom modelo, e ordenar os atributos de acordo com alguma métrica de sua relevância para o problema e então selecionar os melhores atributos. Enquanto achar o subconjunto de atributos úteis pode excluir atributos redundantes, embora úteis, selecionar apenas os atributos mais relevantes pode ser sub-ótimo, especialmente se as variáveis escolhidas são redundantes (GUYON; ELISSEEFF, 2003).

Métodos de ranqueamento de variáveis de acordo com sua relevância têm a vantagem de ser computacionalmente escaláveis, já que demandam apenas uma pontuação para cada variável e sua sucessiva ordenação (GUYON; ELISSEEFF, 2003). Outra vantagem está no fato desses métodos serem robustos a *overfitting*, visto que, embora introduzam viés, diminuem consideravelmente mais a variância (HASTIE; TIBSHIRANI; FRIEDMAN, 2013). Esses métodos que avaliam atributos individualmente têm, no entanto, a desvantagem de serem incapazes de capturar efeitos resultantes de interações entre atributos. De fato, segundo Guyon e Elisseeff (2003), um atributo que por si só é inútil pode resultar em considerável aumento no desempenho do modelo gerado contanto que considerado juntamente com outros atributos, e dois atributos que individualmente são inúteis podem ser úteis se utilizados em conjunto.

Neste estudo, foram utilizados dois algoritmos de seleção de atributos: ReliefF e o índice Gini. Analisando o código da biblioteca *skfeature*, nota-se que a implementação utilizada de seleção de atributos pelo índice Gini dá a seguinte importância $gini_i$ ao i-ésimo atributo:

$$gini_i = \min_{z \in Z} \left[P(W_{>z}) \left(1 - \sum_j P(C_j|W_{>z})^2 \right) + P(W_{<z}) \left(1 - \sum_j P(C_j|W_{<z})^2 \right) \right] \tag{2.1}$$

Onde Z é o conjunto de todos os pontos de divisão possíveis para o atributo i, $W_{>z}$ é o conjunto de todas as amostras cujo valor do atributo i é maior que z e $W_{<z} = \overline{W_{>z}}$ e C_j é a j-ésima classe. Naturalmente, a função P denota probabilidade. Quanto mais heterogêneos os conjuntos resultantes de um ponto de divisão, maior serão as probabilidades condicionais na equação e menor será o termo entre colchetes. Logo, a relevância de um atributo segundo essa fórmula será inversamente proporcional à heterogeneidade resultante da divisão no melhor ponto de divisão possível. Quanto menor $gini_i$, maior a importância do atributo.

Quanto ao ReliefF, uma de suas vantagens é não assumir independência entre as variáveis, ao contrário de muitas outras heurísticas da relevância de atributos. Esse método também é eficiente, sensível à informação contextual e pode corretamente estimar a qualidade de atributos em problemas com forte interdependência entre os atributos (ROBNIK-ŠIKONJA; KONONENKO, 2003). A ideia fundamental do algoritmo consiste de selecionar aleatoriamente uma amostra D_i do conjunto de dados, e então encontrar as k amostras mais perto de D_i que pertencem à mesma classe de D_i, onde k é um inteiro determinado pelo usuário, e para cada classe C diferente da

classe de D_i, as k amostras mais perto de D_i que pertencem à classe C. Essas amostras encontradas são ditas, respectivamente, acertos mais próximos (do inglês *nearest hits*) e erros mais próximos (do inglês *nearest misses*). Penaliza-se então atributos nos quais há grandes distâncias entre D_i e os acertos mais próximos, e aumenta-se a importância de atributos nos quais há grandes distâncias entre D_i e os erros mais próximos (ROBNIK-ŠIKONJA; KONONENKO, 2003). Esse processo é repetido um número de vezes definido pelo usuário a fim de obter importâncias mais fiéis às características do conjunto de dados.

2.2.4 Validação de modelos

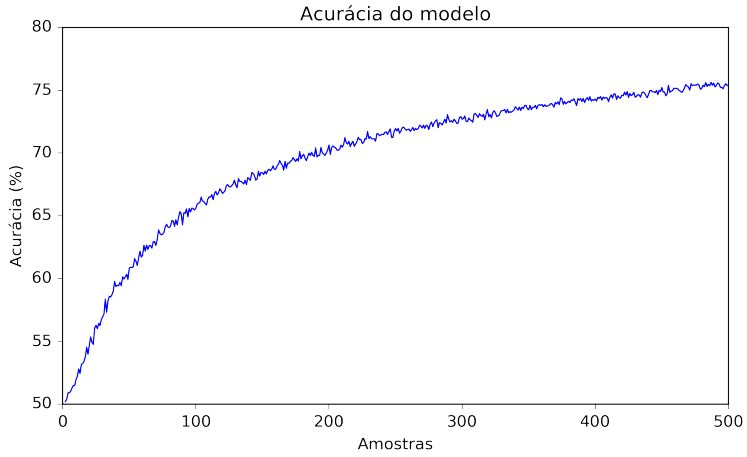

Figura 2 – Estimativa do erro de generalização em função do número de amostras utilizadas no conjunto de treino. Foi utilizado o gerador de datasets correspondente a função *make_hastie_10_2* da biblioteca *sklearn* para gerar tanto o conjunto de treino quanto o de teste. A curva mostra a média do erro num conjunto de teste de 20.000 amostras para um classificador *random forest*. Embora ruidosa, a curva tem uma tendência clara: mais exemplos implica menor erro de generalização.

Um modelo gerado por um algoritmo de aprendizado supervisionado é apenas tão útil quanto sua capacidade de generalização, isto é, de prever corretamente saídas para exemplos nunca antes vistos. Assim, é desejável medir essa capacidade de generalização. Para realizar isso, é possível separar o conjunto de dados que se tem em dois conjuntos disjuntos, um conjunto de treino, com o qual o algoritmo de aprendizado será treinado, e o conjunto de teste, com o qual o modelo será avaliado. Aqui, há um dilema: quanto mais se aumenta o conjunto de treino, melhor tende a ser o modelo gerado, conforme pode ser visto na Figura 2, mas como os conjuntos de teste e treino são disjuntos, aumentar o conjunto de treino implica diminuir o conjunto de teste. Diminuir o conjunto de teste significa que a estimativa do erro de generalização tenderá a ser mais distante do erro de generalização real.

Uma maneira melhor de estimar o erro de generalização é fazer uso de validação cruzada. A execução do método de validação cruzada *K-fold cross validation* consiste de separar os exemplos existentes em K conjuntos disjuntos $d_1, d_2, ..., d_K$ e realizar o seguinte procedimento K vezes: na i-ésima vez, treina-se um classificador com o conjunto de exemplos $D - d_i$, onde D é o

conjunto inteiro de exemplos, e então testa-se o erro do modelo gerado no conjunto d_i. Toma-se o erro de generalização como a média das K estimativas de erro obtidas.

A partir Figura 2 podemos tirar conclusões interessantes acerca da escolha de um k para uso no k-fold. Numa situação real de emprego de aprendizado de máquina, treinaria-se o modelo com o conjunto de amostras inteiro que se tem, e o k-fold seria usado apenas para estimar o erro desse conjunto. No entanto, estamos usando modelos treinados com uma fração $(k-1)/k$ dos dados que se tem para estimar o erro do modelo treinado com o conjunto inteiro. Logo, quanto menor o k utilizado, mais distante tenderá a estar o erro estimado pelo k-fold do erro real, conforme pode ser visto pela figura em questão. A figura ainda mostra que, para um número pequeno número de amostras, essa discrepância é mais dramática do que para um grande número de amostras.

Quanto a forma de quantificar a qualidade de um modelo, existem várias abordagens possíveis. A maneira mais simples é por meio da acurácia, que é apenas a porcentagem de acertos de predições sobre um conjunto. Essa abordagem, porém, tende a ser ruim em conjuntos desbalanceados. Consideremos uma situação na qual o conjunto possui 10 exemplos negativos e 990 exemplos positivos, e um modelo gerado que sempre prevê a classe de um exemplo como sendo positiva. Observaríamos uma impressionante acurácia de $0,99$, mas isso não revela nada sobre o problema do modelo não ser capaz de acertar sequer um exemplo da classe negativa. Existem outras métricas, no entanto, que visam corrigir justamente esse problema, como a especificidade, a sensibilidade, o *gmean*. Seja TP a taxa de verdadeiros positivos de um modelo sobre um conjunto, FP a taxa de falsos positivos, TN a taxa de verdadeiros negativos e FN a taxa de falsos negativos. A partir dessas quatro medidas, podemos construir duas outras mais simples, sensibilidade e especificidade, dadas pelas equações a seguir:

$$Especificidade = \frac{TN}{TN+FP} \tag{2.2}$$

$$Sensibilidade = \frac{TP}{TP+FN} \tag{2.3}$$

Podemos ainda combinar essas duas medidas em uma única, se observarmos que uma alta taxa de acerto nas duas classes resultará em um valor alto para as duas medidas. Uma maneira de fazer isso é com a medida *gmean*, que é a média geométrica da especificidade e da sensibilidade. Outra maneira parte da observação que alguns classificadores são construídos a partir de regressores por meio do estabelecimento de um limiar na saída, abaixo do qual um exemplo pertencerá a uma das classes, ou à outra caso contrário. A curva sensibilidade versus especificidade formada pela variação do limiar é chamada *receiver operating characteristic* (ROC), e a área sob essa curva é frequentemente usada como medida da qualidade de um modelo, e é chamada *area under curve* (AUC).

2.3 Redes Complexas

Podemos aplicar a modelagem de grafos em diversos problemas da vida real, como na rede de amizades entre pessoas, onde nós são pessoas e arestas são amizades, na *World-Wide web*, considerando *websites* como nós e *links* como arestas, na rede de interações entre proteínas, onde nós são proteínas e arestas são interações. Essa universalidade dos grafos levanta a questão de se as redes encontradas no mundo real são fundamentalmente aleatórias ou não. Essa questão está ligada ao conceito de redes complexas, que podem ser definidas como grafos de larga escala que possuem padrões de conexão não triviais (SILVA; ZHAO, 2016).

Outro questão importante de ser levantada é se a maneira como uma rede se organiza, ou seja, sua topologia, tem ligação com as propriedades do sistema que ela modela. De fato, em vários domínios essa relação é encontrada. A vulnerabilidade de uma rede de distribuição de energia elétrica é inerente a topologia da rede (ALBERT; ALBERT; NAKARADO, 2004). Em pelo menos alguns organismos a importância de uma proteína à sobrevivência está associada a sua posição na rede de interação de proteínas (JEONG *et al.*, 2001). Outros exemplos desse tipo de análise podem ser encontrados na literatura (ALBERT; BARABÁSI, 2002).

A fim de estudar a topologia de uma rede, é necessário que seja possível caracterizá-la, por exemplo quantificando certas características da rede. A seguir, fornecemos uma descrição simples de cada uma das medidas que utilizamos neste estudo: k-*coreness*, coeficiente de agrupamento, *betweenness centrality*, assortatividade local, grau médio dos vizinhos e grau.

O grau $deg(i)$ de um nó i é simplesmente seu número de vizinhos, sendo vizinhos apenas os nós com os quais o nó em questão tem conexão direta. O coeficiente de agrupamento de um nó i é a razão entre o número de triângulos $T(i)$ no grafo em que pelo menos um de seus vértices é i e a quantidade de triângulos em que o nó i estaria envolvido se todos os seus vizinhos fossem conectados entre si (ALBERT; BARABÁSI, 2002). Matematicamente, o coeficiente de agrupamento C_i de um nó i é dado por:

$$C_i = \frac{2T(i)}{deg(i)\,(deg(i)-1)} \tag{2.4}$$

O *k-core* de um grafo é o máximo subgrafo cujos nós ainda têm grau maior ou igual a k (BATAGELJ; ZAVERSNIK, 2003). Uma medida derivada desta é o *k-coreness*, definida para cada nó do grafo. O *k-coreness* de um nó i corresponde ao máximo X tal que i pertence a $X-core$ mas não a nenhum $Y-core$, tal que $Y > X$ (LÜ *et al.*, 2016). Segundo Wuchty e Almaas (2005), *k-coreness* pode ser usado como uma medida da centralidade e grau de um nó. A assortatividade de um grafo é apenas o coeficiente de correlação de Pearson entre o grau de pares de nós ligados. Uma alta assortatividade positiva revela que nós tendem a se conectar com outros nós de graus parecidos, enquanto uma assoratividade negativa alta revela que nós de alto grau tendem, em média, a se conectar com nós de baixo grau, e vice-versa. A assortatividade local A_i de um nó i corresponde à sua contribuição individual para a assortatividade da rede

(PIRAVEENAN; PROKOPENKO; ZOMAYA, 2008), e pode ser dada pela equação:

$$A_i = \frac{(j+1)(j\bar{k} - \mu_q^2)}{2N\sigma_q^2} \tag{2.5}$$

Onde j é o grau faltante (tradução livre) do nó i - definido como o grau do nó em questão menos 1 (SOLÉ; VALVERDE, 2004) -, $\bar{k}$ é a média do grau faltante dos nós vizinhos do nó i, N é o número de nós no grafo inteiro e μ_q e σ_q são, respectivamente, a média e desvio padrão do grau dos nós do grafo inteiro. Por fim, a medida *betweenness centrality*, também conhecida como *shortest-path betweenness*, B_i de um nó i é a fração dos menores caminhos no grafo que passam por i, e é dada por:

$$B_i = \sum_{s,t \in V} \frac{\sigma(s,t|i)}{\sigma(s,t)} \tag{2.6}$$

Sendo V o conjunto de nós do grafo, $\sigma(s,t)$ a quantidade de menores caminhos que existem entre os nós s e t, e $\sigma(s,t|i)$ a quantidade de menores caminhos entre os nós s e t que passam pelo nó i (BRANDES, 2008). Essa medida pode ser interpretada como o controle que um nó tem sobre as conexões dois-a-dois entre outros nós, assumindo que a importância de cada conexão é igualmente dividida entre cada caminho mais curto (BRANDES, 2008).

2.4 Trabalhos similares

A ideia de aplicar aprendizado de máquina a fim de determinar as causas do envelhecimento e fatores ligados a longevidade não é nova. Como exemplos de aplicação de regressão na literatura, podemos citar estudos que tentam predizer a idade cronológica de cada indivíduo (HORVATH, 2013; HANNUM *et al.*, 2013) e um outro estudo que tenta predizer a taxa de envelhecimento (NAKAMURA; MIYAO, 2007). Como exemplos de aplicação de classificação, há estudos que criam modelos para predizer se proteínas são ou não ligadas ao processo de envelhecimento (FREITAS; VASIEVA; MAGALHÃES, 2011; LI; ZHANG; GUO, 2010), se genes são pró ou anti-longevidade (WAN; FREITAS; MAGALHÃES, 2015) ou se determinados genes estão ou não relacionados com mudanças na longevidade esperada (LI; DONG; GUO, 2010).

Uma análise mais aprofundada e completa acerca da aplicação de AM em pesquisas sobre envelhecimento pode ser encontrada em (FABRIS; MAGALHÃES; FREITAS, 2017). Os autores dividem os atributos utilizados nesses estudos em três categorias distintas: atributos que codificam informações sobre as funções biológicas de cada gene e proteína, atributos derivados de interações entre proteínas e a rede por elas formadas, e atributos a nível de organismo, que incluem medidas feitas em indivíduos da população. Nosso estudo, em particular, junta os dois primeiros tipos de atributos mencionados.

Os algoritmos de aprendizado empregados também variam bastante. Há estudos que focam em apenas gerar modelos caixa-preta para predizer informações de envelhecimento (LI; ZHANG; GUO, 2010). Outros pesquisadores estão mais interessados em utilizar aprendizado de máquina como um meio de extrair conhecimento interpretável, e por isso preferem algoritmos que geram modelos interpretáveis (FREITAS; VASIEVA; MAGALHÃES, 2011; FABRIS; MAGALHÃES; FREITAS, 2017), como métodos baseados em árvores.

De maneira geral, esse tipo de pesquisa tem corroborado com fatos biológicos já sabidos e formulado hipóteses relacionadas ao processo de envelhecimento, mas uma das principais fraquezas tem sido a falta de verificação experimental do conhecimento extraído (LI; ZHANG; GUO, 2010).

Capítulo 3

DESENVOLVIMENTO

Este capítulo descreve como foram rrealizados os experimentos relacionados a este trabalho. Para isso, apresenta-se inicialmente de onde foram obtidos usados neste estudo, e como foram processados de maneira a obter-se os *datasets* que mais tarde foram utilizados. A seguir, descreve-se os experimentos realizados, e analisa-se os resultados obtidos.

3.1 Atividades Realizadas

3.1.1 Coleta dos Dados

Neste estudo foram usados dados de múltiplos bancos de dados biológicos disponíveis, de forma a descrever diversos aspectos sobre cada gene. Utilizamos dados de funções biológicas de cada gene, obtido do *website* do projeto *Gene Ontology*, de interações entre proteínas, obtidas do banco de dados I2D, e da relação de cada gene com o processo de envelhecimento. Todos os dados obtidos estão disponíveis ao público e podem ser obtidos gratuitamente.

O projeto *Gene Ontology* (GO), de onde foram obtidos os dados das funções biológicas de cada gene, tem o objetivo de produzir um vocabulário controlado, estruturado e definido precisamente, a fim de descrever os papéis dos genes e de seus respectivos produtos em qualquer organismo (ASHBURNER *et al.*, 2000). Com esse fim, o projeto mantém três ontologias[1] independentes: uma para processos biológicos, outra para funções moleculares e ainda outra para componentes celulares. Uma explicação bem detalhada de cada uma dessas categorias, bem como a razão para essa divisão em três ontologias, pode ser encontrada em (ASHBURNER *et al.*, 2000). Além de disponibilizar esses ontologias, o projeto também provê o que chama de anotações, que é apenas um mapeamento dos genes de diversas espécies - humanos e organismos modelo - para termos da ontologia, assim como uma lista das evidências na literatura que suporta cada um dos mapeamentos (CONSORTIUM *et al.*, 2017). Podem ser encontrados no *website* do projeto, www.geneontology.org, todos os dados aqui mencionados. No momento da escrita deste estudo, o projeto já reunia mais de 40.000 conceitos biológicos, e as anotações são baseadas nos resultados de mais de 100.000 estudos científicos revisados por pares. É possível encontrar mais de 400.000 anotações apenas para genes humanos.

[1] Uma ontologia compreende um conjunto de termos (ou funções biológicas) bem definidos e com relações bem definidas entre si (ASHBURNER *et al.*, 2000).

O conjunto de genes ligados a envelhecimento foi obtido do banco de dados GenAge, que é parte de uma iniciativa maior chamada *Human Ageing Genomic Resources* (HAGR). Essa iniciativa tem o propósito de manter uma coleção de bancos de dados e ferramentas voltadas para o estudo de aspectos biológicos e genéticos do fenômeno do envelhecimento. O GenAge, ainda, é divido em três seções: o conjunto de genes diretamente associados ao envelhecimento em humanos e os melhores genes candidatos a serem ligados ao processo de envelhecimento, baseado evidências em outros organismos modelo, o conjunto de genes cuja expressão é comumente alterado durante o envelhecimento em humanos e o conjunto de genes ligados a longevidade em organismos modelo (TACUTU *et al.*, 2012). Os critérios específicos que levam a inclusão de um gene no conjunto de genes humanos ligados ao envelhecimento podem ser encontrados no mesmo artigo que descreve a iniciativa, mas é importante ressaltar que genes ligados apenas indiretamente ao envelhecimento e evidências em organismos que não humanos também são aceitos, segundo o *website* devido ao impacto de falsos negativos ser maior do que o impacto de falsos positivos. Foi usado em particular o primeiro conjunto citado, que compreende 307 genes humanos. Os dados estão disponíveis do *website* do projeto.

Por fim, foram utilizados os dados a respeito de interações entre proteínas do projeto OPHID, que contém dados de interações entre proteínas (PPI, do inglês *protein protein interaction*) em múltiplos organismos, inclusive humanos. Uma das características desse banco de dados que o diferencia dos demais que se propõe a catalogar PPIs é que contém não apenas interações que já foram encontradas na literatura, mas também previsões de interações entre proteínas humanas baseadas em interações encontradas em organismos modelo (BROWN; JURISICA, 2005). O projeto, no entanto, não realiza curação manual das interações para as quais há evidências diretas na literatura; ao invés disso, reúne esse tipo de interações que outros bancos de dados que realizam curação manual.

3.1.2 Preparação dos dados

Juntar os dados das três bases de dados utilizadas foi particularmente dificultado pela existência de múltiplos identificadores para genes e proteínas. O GenAge utiliza identificadores *Entrez Gene* para genes ligados a envelhecimento, enquanto a base GO e as proteínas do I2D utilizam identificadores *UniProtKB accession number* (AC). Utilizamos o arquivo de mapeamento entre identificadores que pode ser encontrado no *website* do projeto UniProt. No entanto, nota-se que o mapeamento entre identificadores *Entrez* e AC não é bijetivo. Há multiplos *Entrez Gene* que mapeiam para um único AC, e há múltiplos AC que mapeiam para um mesmo *Entrez Gene*. É esperado que haja múltiplos AC que mapeiam para um mesmo *Entrez Gene*, visto que AC identifica proteínas e *Entrez Gene* identifica genes, e um único gene pode codificar mais de uma proteína (MATLIN; CLARK; SMITH, 2005). Todavia, não é esperado que multiplos *Entrez Gene* mapeiem para um único AC. A fim de eliminar ambiguidades, foram removidas todas proteínas que mapeavam para mais de um *Entrez Gene* e todos os genes que mapeavam

para pelo menos uma proteína com essa característica, a fim de eliminar ambiguidades no estudo.

Neste estudo em particular estamos interessados em humanos, portanto os dados GO e I2D foram filtrados para a espécie de interesse. A base de dados do I2D lista interações entre proteínas para 18.265 genes humanos diferentes, enquanto GO provê anotações para 19.473 genes humanos. Ao se eliminar os genes para o qual o mapeamento entre identificadores não era bijetivo, restaram 17.687 genes.

Foi criado um atributo binário *age_related* que indica se um gene é ou não relacionado ao processo de envelhecimento. Todos os genes inclusos na base GenAge foram colocados na classe positiva. No entanto, dos 307 genes originais, havia mapeamentos bijetivos para apenas 289 destes genes, e portanto eliminou-se os genes restantes a fim de eliminar ambiguidades do estudo. Para a classe negativa, foram selecionados aleatoriamente outros 289 genes do universo de genes humanos. Vale ressaltar que não necessariamente esses genes na classe negativa não estão ligados ao processo de envelhecimento humano, visto que essa ausência de informação pode ocorrer apenas porque ainda não se realizaram estudos sobre um determinado gene. Ainda assim, se supusermos que boa parte dos genes de fato ligados ao envelhecimento já estão no GenAge, então teremos uma baixa probabilidade de selecionar genes relacionados ao envelhecimento para a classe negativa.

Essa análise é feita propositalmente de maneira especulativa e qualitativa: não encontramos na literatura estimativas de quantos genes há cuja relação com o processo de envelhecimento é desconhecida, portanto não podemos fazer uma análise quantitativa. No entanto, essa abordagem para adoção de exemplos na classe negativa já foi utilizada anteriormente na literatura (FREITAS; VASIEVA; MAGALHÃES, 2011; LI; ZHANG; GUO, 2010), inclusive em outros domínios (LÓPEZ-BIGAS; OUZOUNIS, 2004).

Foram criados dois tipos de atributos baseados nos dados de interação entre proteínas que obtivemos. O primeiro tipo compreende as medidas de redes complexas mencionadas no capítulo anterior: foram capturados, para cada nó, seu grau, grau médio de seus vizinhos, assortatividade local, coeficiente de agrupamento, *betweenness centrality* e k-*coreness*. Os nomes desses atributos conforme dados no *dataset* podem ser encontrados na tabela 1. Para extrair essas medidas, utilizou-se a biblioteca NetworkX (HAGBERG; SCHULT; SWART, 2008). O segundo tipo de atributo visa fornecer a informação de cada gene interage com outro gene específico. Um exemplo desse tipo de atributo pode ser *inter_WRN*, que assume o valor 1 para genes que interagem com o gene WRN e o valor 0 para genes que não interagem. Foram criados todos os atributos desse tipo possíveis. No total, há 10.124 atributos que capturam informações sobre interações entre proteínas, que é diferente do número de genes pois há genes no conjunto de genes humanos que não interagem com nenhum dos genes que se encontram no *dataset*.

Quanto aos papéis biológicos de cada gene, os termos das ontologias são organizados num grafo. As relações entre os termos podem ser *A é um B*, *A é parte de B* ou *A regula B*. Há também especializações dessa última relação, *A regula positivamente B* ou *A regula negativamente B*. Uma

Tabela 1 – Nomes dos atributos de medidas topológicas

Medida topológica	Nome do atributo
Grau	intercount
Grau médio dos vizinhos	and
Assortatividade local	assortativity
Coeficiente de agrupamento	clustering
Betweenness centrality	betweenness
k-*coreness*	coreness

descrição detalhada sobre as relações e sobre conclusões que podem ser tiradas dessas relações podem ser encontradas na documentação do projeto GO[2]. Por simplicidade, foi considerado apenas o primeiro tipo de relação na criação de nossos atributos, *A é um B*. As anotações de genes obtidas só se referem a suas funções mais específicas, ou seja, se um gene é associado com uma função A e A é um B, então, embora o gene realmente cumpra a função B, no arquivo a associação entre o gene e a função B não existe. Ademais, essa relação é transitiva. Portanto, o primeiro passo para a criação dos atributos de papéis biológicos foi determinar todas as relações *é um* levando em conta as considerações feitas.

Finalmente, foi criado um atributo para cada termo encontrado na ontologia. Cada atributo desses leva no nome o identificador do termo usado pelo GO, por exemplo, *GO:0006281*. Se um gene possui uma função biológica, direta ou indiretamente por meio de relações *é um*, e sua associação com uma função não é acompanhada pelo qualificador *NOT* na base de dados, consideramos que o gene possui a função e recebe, portanto, o valor 1 no atributo, ou o valor 0 caso contrário. Novamente, estritamente falando aqui não se pode inferir por meio da ausência de uma associação entre gene e função a sua não existência. No entanto, essa consideração é feita também é feita em outros artigos na literatura (LI; ZHANG; GUO, 2010; FREITAS; VASIEVA; MAGALHÃES, 2011), e portanto julga-se ser razoável essa suposição.

Com os dados aqui discutidos, foram construídos três *datasets*: um apenas com os atributos relacionados a interações entre proteínas (aging_complex), outro apenas com as funções biológicas de cada gene (aging_go), e um terceiro com ambos os tipos de atributos (aging_both). Em todos os *datasets*, o atributo a se prever é a ligação ou não do gene com o processo de envelhecimento.

3.1.3 Preparação dos experimentos

Os experimentos deste estudo foram executados em Python, utilizando as bibliotecas XGBoost (CHEN; GUESTRIN, 2016) e sklearn (PEDREGOSA *et al.*, 2011) para os algoritmos de aprendizado e skfeature (LI *et al.*, 2016) para a seleção de atributos. No caso de nosso problema, há muitos parâmetros a serem variados: os *datasets* a serem utilizados, os algoritmos de seleção de atributos, o número de atributos selecionados, os algoritmos de aprendizado

[2] http://www.geneontology.org/page/ontology-relations

utilizados e seus respectivos hiperparâmetros. Testar todas essas combinações aumenta nossas chances de encontrar um modelo bom, mas também dificulta a visualização e interpretação dos resultados. Portanto, optou-se por um meio termo: testamos todas as combinações dos parâmetros apresentados, exceto pelos hiperparâmetros de cada algoritmo de aprendizado descritos mais a seguir, para os quais foram utilizados os valores padrão das bibliotecas utilizadas.

Há 10.124 atributos na base de dados com apenas atributos relacionados a interações entre proteínas *aging_complex*, 8.974 atributos na base de dados com apenas atributos dos papéis biológicos de cada gene *aging_go* e 19.098 atributos na base de dados que contém os atributos de ambas as outras duas *aging_both*. Essa altíssima dimensionalidade de nossos dados torna interessante a aplicação de métodos de seleção de atributos. Foram executados experimentos que não utilizaram nenhum método de seleção de atributos, e outros que utilizaram o índice Gini ou ReliefF. No caso de experimentos que utilizaram SA, por desejar-se descobrir quantos atributos são necessários para se fazer uma boa predição, foram selecionados os melhores 25, 50, 100, 250 e 1000 atributos, conforme julgado pelos algoritmos.

Vale notar que tomou-se o cuidado de executar os algoritmos de SA apenas utilizando o conjunto de treino, e então a seleção feita apenas com base nesses dados é então aplicada ao conjunto de testes. Segundo Hastie, Tibshirani e Friedman (2013), essa é maneira correta de se fazer seleção de atributos em aprendizado supervisionado, pois caso o algoritmo de SA seja aplicado ao conjunto inteiro, as métricas de acurácia do modelo extraidas mais tarde sobre o conjunto de testes pode sobrestimar a acurácia real.

Os algoritmos de aprendizado escolhidos foram CART, *Random Forest*, XGBoost, SVM linear e SVM com kernel RBF. CART elabora um modelo de árvore, cuja interpretação pode ser feita por uma simples análise da árvore gerada. Para *Random Forest* e XGBoost, que usam um conjunto de árvores para gerar suas previsões finais, sua interpretação é um pouco mais difícil, mas há maneiras de atribuir importância às variáveis, baseado nos atributos escolhidos para a divisão de cada nó.

Todos os experimentos foram executados para as três bases de dados possíveis, com o propósito de aferir quanto interações entre proteínas e papéis biológicos de genes conseguem predizer, individualmente, a ligação de um gene com o envelhecimento, e qual o desempenho dessas duas fontes de informações se combinadas. Em cada experimento foi utilizada validação cruzada de 10 vias para estimar o erro dos modelos gerados, utilizando a acurácia como métrica, por ser uma medida fácil de interpretar. Em particular, foi utilizada a validação cruzada estratificada, que mantém as proporções entre as duas classes das amostras em cada partição da validação cruzada. Portanto, como temos uma distribuição exata de 50% das amostras em cada classe, cada partição também seguirá essa proporção. Dessa maneira fica garantido que tanto no conjunto de treino quanto no de testes haverá uma boa quantia de amostras de ambas as classes de genes. Todos os atributos numéricos foram normalizados para ter média 0 e desvio padrão 1.

3.2 Resultados experimentais

Nesta seção, apresenta-se e analisa-se os resultados obtidos dos experimentos realizados. Na primeira subseção, discute-se o efeito da escolha de diferentes algoritmos de aprendizado, métodos de SA e número de atributos selecionados. Na segunda, compara-se os modelos gerados a partir de cada uma das bases de dados que geramos. Por fim, na terceira compara-se a relevância dada aos atributos pelos algoritmos *Random Forest* e XGBoost.

3.2.1 Variações de métodos

A acurácia de todos os modelos encontrados pode ser encontrada na Tabela 2. Um comportamento que pode ser claramente observado é que CART teve sistematicamente o pior desempenho entre todos os classificadores utilizados. O melhor desempenho foi obtido pelo classificador SVC radial, com 84% de acurácia utilizando apenas os 500 melhores atributos da base *aging_both*, conforme indicado pelo índice Gini. Nota-se também que a variação da acurácia obtida não foi tão grande na tabela, pois há apenas 12% de diferença entre a pior e a melhor acurácia. Ressalta-se que o pior classificador ainda obteve uma acurácia significantemente acima de 50%, e portanto considera-se CART ruim apenas relativamente aos outros classificadores.

Observa-se que, conforme pode ser visto na Tabela 2 e em parte nas Figuras 3 e 4, os algoritmos SVC radial e XGBoost tendem a obter desempenhos médios equiparáveis, e também tendem a ser os melhores algoritmos de aprendizado entre o conjunto de algoritmos utilizados. Os algoritmos *RandomForest* e SVC linear tendem a ter um desempenho levemente inferior aos melhores algoritmos, mas ainda são melhores que o CART.

A implementação do CART que foi utilizada, em particular, não utiliza nenhum método de poda da árvore após seu crescimento, o que facilita a ocorrência de *overfitting*. Esse fato sozinho, no entanto, não explica o mau desempenho do algoritmo, pois, mesmo em situações em que selecionamos apenas 25 atributos, CART teve um desempenho baixo. Poderíamos argumentar que o CART teve um desempenho baixo quando selecionamos apenas 25 atributos porque os métodos de seleção de atributos falharam em selecionar os melhores atributos. Entretanto, podemos facilmente descartar essa hipótese, pois outros algoritmos conseguiram obter uma boa acurácia mesmo com poucos atributos. Logo, isso sugere que CART é ruim para gerar hipóteses para distinguir genes ligados ao envelhecimento dos não ligados.

Surpreendentemente, não se observa nenhuma tendência óbvia do desempenho dos classificadores a melhorar ou piorar dependendo do número de atributos. Isso deixa claro que ambos os algoritmos de SA conseguiram com sucesso selecionar bons atributos em termos de relevância, além de que não é necessária uma grande quantidade de atributos para se obter uma boa acurácia. Será analisada mais adiante a hipótese dos classificadores terem escolhido os mesmos atributos independente da quantidade de atributos selecionados. Na Figura 3 fixamos o método de SA para ReliefF e a base de dados para *aging_go* para ilustrar melhor o efeito

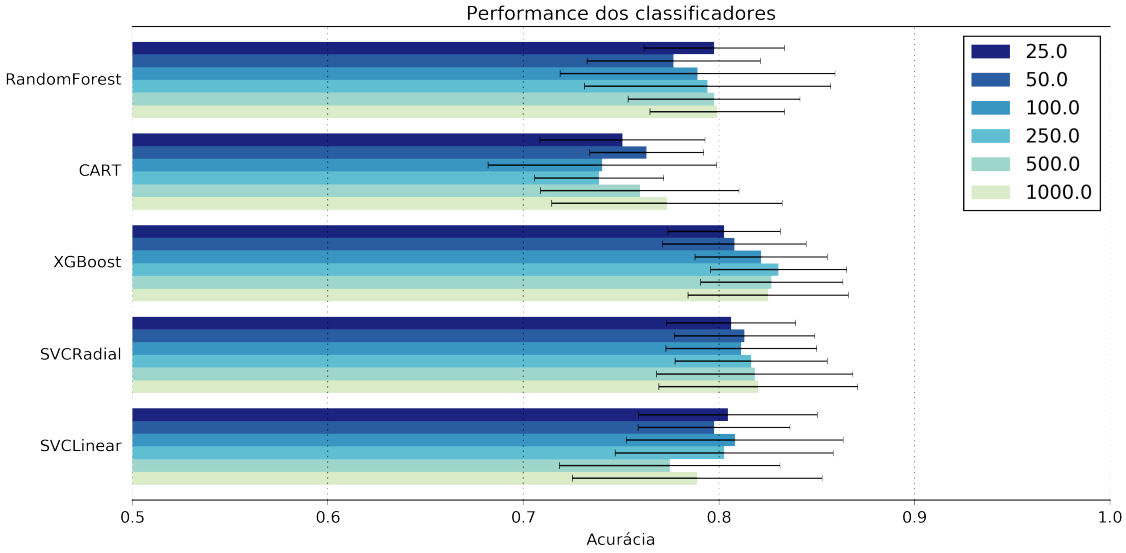

Figura 3 – Média e variância das acurácias amostradas pela validação cruzada de 10 vias, para modelos que utilizaram RelieF no *dataset* aging_go.

da variação do número de atributos escolhidos. A figura apresentada coloca em perspectiva a variância que há na acurácia estimada dos modelos gerados em cada *fold*. Podemos ver que é relativamente alta para todos os casos vistos, o que dificulta a comparação entre modelos.

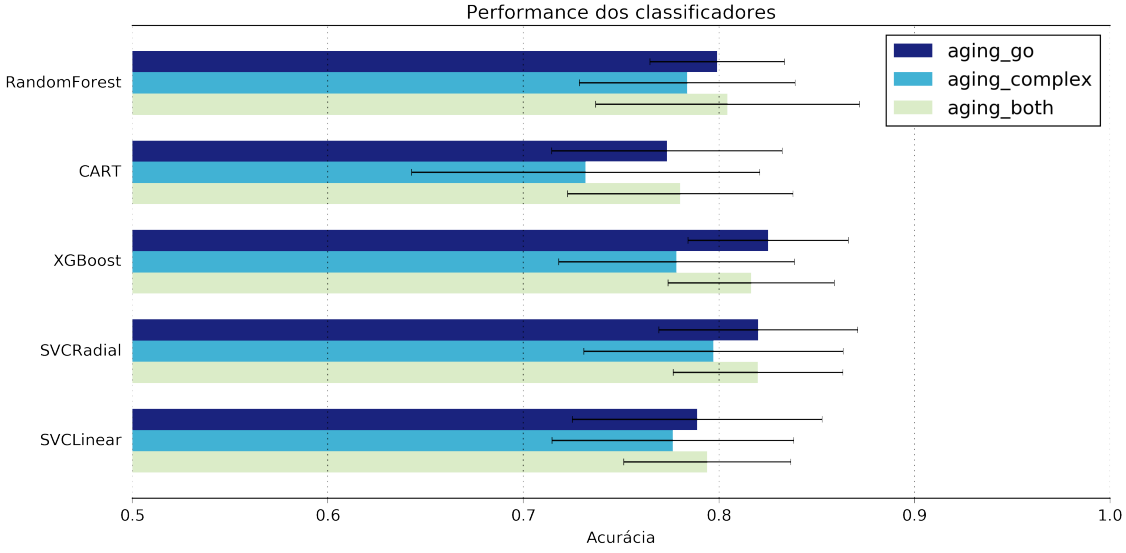

Figura 4 – Média e desvio padrão das acurácias amostradas pela validação cruzada de 10 vias, para modelos que utilizaram os 1000 atributos sugeridos pelo algoritmo ReliefF.

Na Figura 4 observa-se a acurácia dos classificadores em função do algoritmo de aprendizado utilizado e do *dataset*, fixando-se os atributos selecionados em 1000 e pelo algoritmo ReliefF. Novamente, aqui o desvio padrão dos dados é muito grande, o que dificulta comparações entre desempenhos, mas parece haver uma tendência do desempenho tanto em aging_go quanto em aging_both ser melhor do que o desempenho em aging_complex. Entre aging_both e aging_go não há nenhuma tendência clara. Utilizamos o teste estatístico Wilcoxon *signed-ranks*

para comparar os desempenhos nas condições mencionadas. Os únicos padrões encontrados com 95% de confiança foram que, para XGBoost, a acurácia em aging_both é maior que em aging_complex (valor-p=0.02469), e a acurácia em aging_go é maior que em aging_complex (valor-p=0.01459).

Ressalta-se que a ausência de uma diferença de desempenho estatísticamente significante não significa que essa diferença não exista de fato. Uma possível explicação para o desempenho em aging_both não ter sido melhor que em aging_go seria que os algoritmos de aprendizado simplesmente não foram capazes de utilizar as novas informações para produzir modelos melhores.

3.2.2 Análise de atributos

Neste, três dos classificadores selecionados permitem a interpretação dos modelos gerados: CART, *Random Forest* e XGBoost. A interpretação do CART é direta, visto que é uma única árvore. A interpretação dos dois outros métodos baseados em comitê, no entanto, é mais difícil visto que suas previsões finais são baseadas na combinação das previsões individuais de cada árvore. Uma das maneiras possíveis de medir a importância dos atributos é pelo número de amostras que passam por nós que utilizam o atributo em questão durante o treino. Essa será a métrica utilizada.

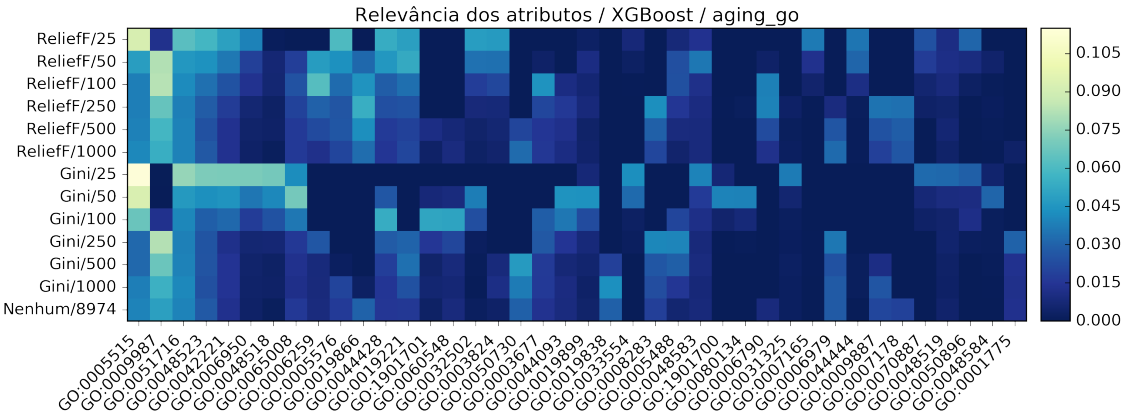

Figura 5 – Relevância dos atributos selecionados segundo XGBoost, para o *dataset* aging_go.

Nas Figuras 5 e 6, estão apresentados os atributos mais relevantes no *dataset* aging_go e suas respectivas métricas médias de relevância, conforme descrito anteriormente, num formato de *heatmap*. O valor apresentado é a média das relevâncias obtidas ao longo da execução da validação cruzada. Valores mais altos - e portanto cores mais claras - indicam uma relevância maior. Vê-se que para este *dataset*, os dois algoritmos de aprendizado concordam, pelo menos em partes, sobre os atributos mais relevantes: entre os 5 atributos mais relevantes segundo os dois algoritmos há 3 atributos em comum. Seja N o número de atributos selecionados em cada

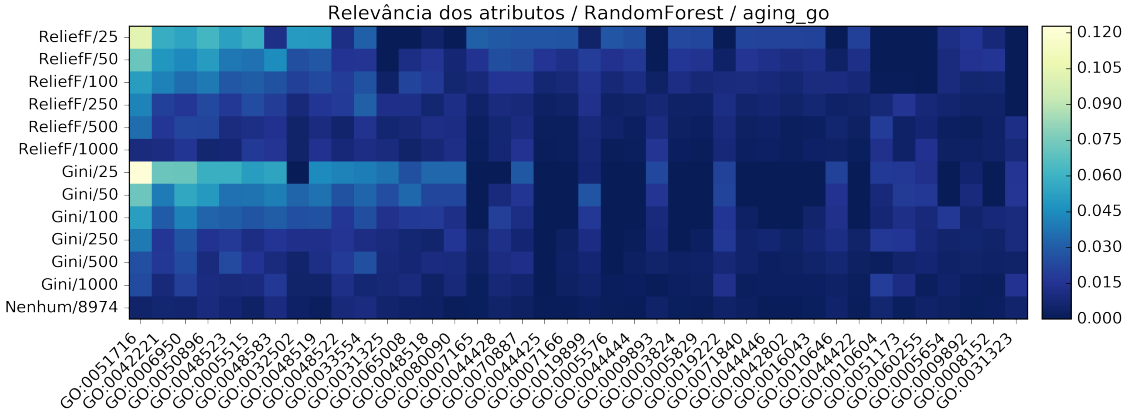

Figura 6 – Relevância dos atributos selecionados segundo *Random Forest*, para o *dataset* aging_go.

um dos casos. Uma das tendências que podem ser observadas é que, conforme se aumenta N, a importância de atributos dos atributos que já tinham alguma importância antes tende a diminuir. Isso é bastante interessante pois, como vimos anteriormente, não há tendências claras de mudança de desempenho dos classificadores treinados, por mais que haja mudança nas relevâncias relativas de cada atributo. Ademais, em alguns casos a importância passa abruptamente de zero para algum valor maior que zero, como é o caso do atributo GO:0009987 no XGBoost ou o GO:0032502 no *Random Forest*. Isso poderia ser explicado pelo fato do atributo não ter sido selecionado para valores menores de N, ou então nessa mesma situação surgiu algum outro atributo com o qual o em questão tem alguma interdependência.

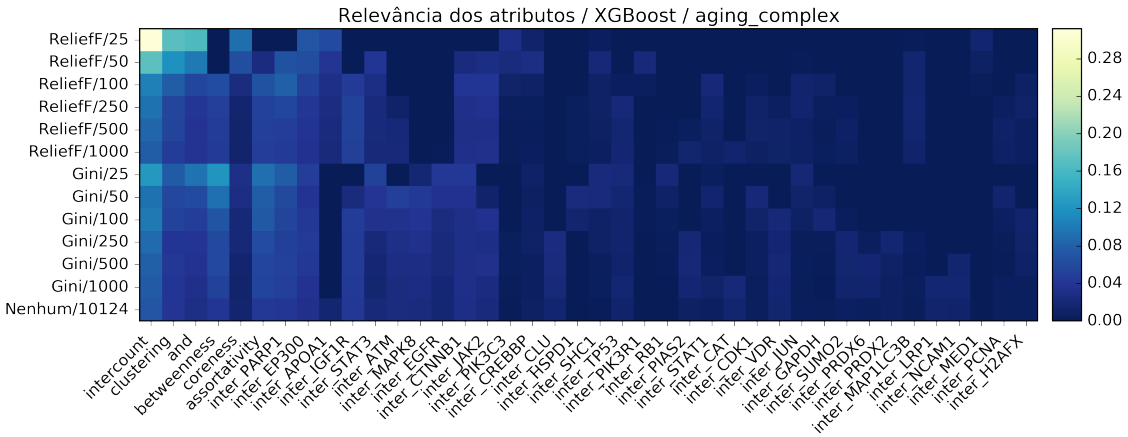

Figura 7 – Relevância dos atributos selecionados segundo XGBoost, para o *dataset* aging_complex.

Nas Figuras 7 e 8 estão apresentados a relevância dos atributos segundo, respectivamente, o XGBoost e o *Random Forest*. O atributo "and"corresponde ao grau médio dos vizinhos (*average neighbor degree*). Aqui, nota-se que a diferença nas relevâncias atribuidas pelos dois algoritmos varia muito. O algoritmo *Random Forest* dá notavelmente altas importâncias aos atributos de medidas de redes complexas, e quase nenhuma a atributos de interações binárias (atributos

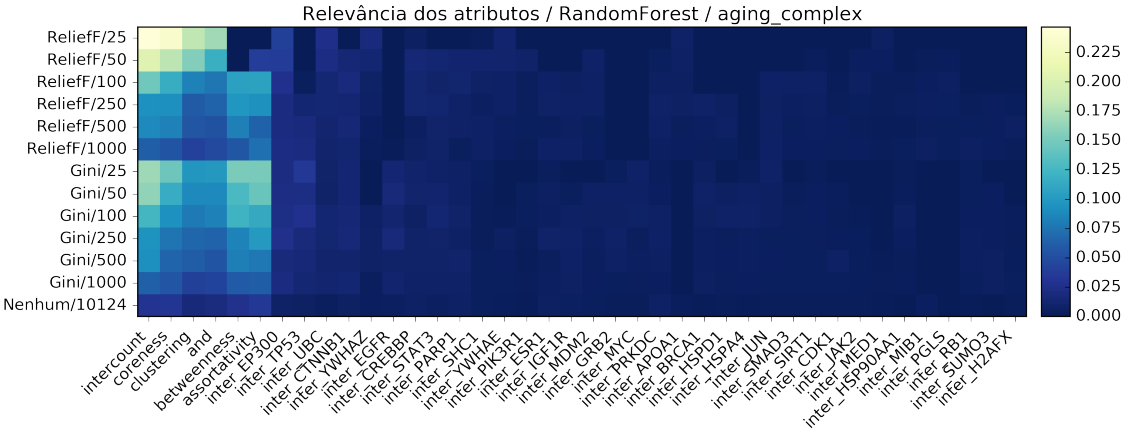

Figura 8 – Relevância dos atributos selecionados segundo *Random Forest*, para o *dataset* aging_complex.

que tem o prefixo "inter_"). Uma maneira de explicar isso seria pelo fato de que o critério de seleção de atributo para divisão de nós nas árvores é o índice Gini. Esse critério tende a favorecer atributos com um grande número de valores diferentes, como é o caso de atributos com valores contínuos. Considerando esse fato, é questionável utilizar a relevância dos atributos conforme definido pelo *Random Forest* para se tirar conclusões acerca da relevância real das variáveis. Ao contrário do *Random Forest*, vemos que embora o XGBoost também tenha atribuido muita importância aos atributos de redes complexas, este atribuiu pelo menos alguma importância a atributos de interações binárias, como por exemplo interações com os genes PARP1 e EP300.

Uma das tendências que não foi encontrada, ao contrário do que se esperava, foi que nenhum dos algoritmos atribuiu importância significativa à interação binária com as proteínas WRN e XRCC5. Um estudo que investigou também a aplicação de aprendizado de máquina ao problema de classificação de genes em ligados ou não ao envelhecimento (FREITAS; VASIEVA; MAGALHÃES, 2011) encontrou que as interações com as proteínas XRCC5 e WRN eram importantes. No caso da proteína WRN, os autores discutiram a possibilidade da interação com WRN ter sido considerada importante devido a um viés no *dataset* que utilizaram, pois a proteína WRN e as proteínas com as quais interage tendem a ser mais estudadas no contexto de envelhecimento que outros tipos de proteínas. Para a outra proteína, XRCC5, no entanto, os autores consideraram isso um resultado válido. Há 3 principais diferenças entre nossos estudos que podem ter causado essa diferença de relevância atribuída às interações com essas proteinas: (i) o estudo em questão, embora também tenha usado a mesma variável de saída, se restringiu apenas ao conjunto de genes que são parte do mecanismo de reparo de DNA em humanos, (ii) o estudo em questão utiliza outra métrica de relevância das variáveis, mais especificamente o número de vezes que um atributo aparece no raiz das árvores geradas, (iii) nossas bases de dados de interações entre proteínas são fundamentalmente diferentes, pois em nosso projeto foram utilizadas não apenas interações de alta confiança mas também interações previstas, enquanto no estudo em questão utilizou-se apenas interações de alta confiança.

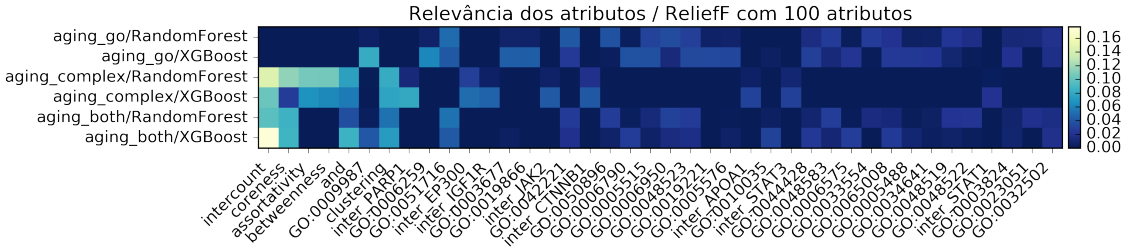

Figura 9 – Relevância dos atributos selecionados segundo XGBoost, para o *dataset* aging_complex.

Para analisar a diferença entre os atributos escolhidos nas diferentes bases de dados, optou-se por fixar o algoritmo de SA para ReliefF e N em 100. Visto que atributos tendem a reter, pelo menos parcialmente, sua importância conforme N aumenta, julga-se que nossa análise será generalizável para valores de N maiores. Na Figura 9, observa-se que quando juntamos os dois tipos de atributos, os classificadores tendem a atribuir grande importância aos atributos de medidas de redes complexas, com exceção da assortatividade local (*assortativity*) e k-*coreness* (*coreness*), que perdem completamente sua importância. Todos os atributos de interações binárias também passam a ser notavelmente irrelevantes. O número de interações de uma proteína (intercount), se mostrou particularmente bom para predizer a classe dos genes. Isso concorda com os resultados de (FREITAS; VASIEVA; MAGALHÃES, 2011; LI; ZHANG; GUO, 2010). A figura também sugere que *coreness* tem alta relevância. De fato, Li, Zhang e Guo (2010) estuda a relevância dessa medida nesse problema e chega à conclusão de que proteínas ligadas ao envelhecimento tendem a ter um grau de *coreness* mais alto, e portanto a serem mais centrais no interactoma[3] humano.

Outro resultado que obtivemos que concorda com Freitas, Vasieva e Magalhães (2011) foi a relevância de resposta ao estímulo em nosso problema de classificação. Ao atributo GO:0051716, definido como "resposta celular a estímulo", foi atribuída grande relevância pelos algoritmos de aprendizado.

3.2.3 Dificuldades e limitações

Não foi realizado nenhum ajuste no hiperparâmetro dos algoritmos de aprendizado que utilizamos. Esse ajuste poderia possivelmente melhorar o desempenho destes, mas optou-se por não o fazer por simplicidade.

Outra dificuldade encontrada foi que, de certa forma, nossos atributos e a classe a ser prevista são ambos ruidosos. Não encontrei nenhum banco de dados que contivesse um conjunto de genes não ligados ao envelhecimento para usar como classe negativa. O fato de escolhermos

[3] Nome que se dá à rede formada pelas interações entre proteínas

Tabela 2 – Resultados dos experimentos realizados da combinação entre algoritmos de AM, métodos de seleção de SA, número de atributos selecionados e *dataset*. Células vazias indicam que um experimento não foi realizado. Por exemplo, não faz sentido usar algum método de SA para selecionar todos os atributos. Cada célula contém a média e o erro padrão da acurácia resultante do experimento usando validação cruzada de 10 vias. Cores claras indicam médias altas, enquanto cores escuram indicam médias baixas.

Dataset	Algoritmo AM	Método SA	Número de atributos selecionados						
			25	50	100	250	500	1000	Todos
aging_go	RandomForest	ReliefF	0.80 ± 0.04	0.78 ± 0.04	0.79 ± 0.07	0.79 ± 0.06	0.80 ± 0.04	0.80 ± 0.03	-
aging_go	RandomForest	Gini	0.76 ± 0.03	0.79 ± 0.05	0.76 ± 0.05	0.80 ± 0.05	0.80 ± 0.04	0.80 ± 0.07	-
aging_go	RandomForest	Nenhum	-	-	-	-	-	-	0.81 ± 0.08
aging_go	CART	ReliefF	0.75 ± 0.04	0.76 ± 0.03	0.74 ± 0.06	0.74 ± 0.03	0.76 ± 0.05	0.77 ± 0.06	-
aging_go	CART	Gini	0.74 ± 0.05	0.76 ± 0.04	0.73 ± 0.04	0.77 ± 0.05	0.74 ± 0.07	0.77 ± 0.07	-
aging_go	CART	Nenhum	-	-	-	-	-	-	0.78 ± 0.05
aging_go	XGBoost	ReliefF	0.80 ± 0.03	0.81 ± 0.04	0.82 ± 0.03	0.83 ± 0.03	0.83 ± 0.04	0.83 ± 0.04	-
aging_go	XGBoost	Gini	0.78 ± 0.04	0.80 ± 0.03	0.80 ± 0.04	0.83 ± 0.03	0.82 ± 0.04	0.83 ± 0.04	-
aging_go	XGBoost	Nenhum	-	-	-	-	-	-	0.83 ± 0.03
aging_go	SVCRadial	ReliefF	0.81 ± 0.03	0.81 ± 0.04	0.81 ± 0.04	0.82 ± 0.04	0.82 ± 0.05	0.82 ± 0.05	-
aging_go	SVCRadial	Gini	0.79 ± 0.03	0.80 ± 0.04	0.82 ± 0.04	0.83 ± 0.03	0.82 ± 0.05	0.82 ± 0.05	-
aging_go	SVCRadial	Nenhum	-	-	-	-	-	-	0.83 ± 0.04
aging_go	SVCLinear	ReliefF	0.80 ± 0.05	0.80 ± 0.04	0.81 ± 0.06	0.80 ± 0.06	0.77 ± 0.06	0.79 ± 0.06	-
aging_go	SVCLinear	Gini	0.78 ± 0.03	0.79 ± 0.05	0.80 ± 0.04	0.78 ± 0.05	0.79 ± 0.04	0.78 ± 0.06	-
aging_go	SVCLinear	Nenhum	-	-	-	-	-	-	0.80 ± 0.05
aging_complex	RandomForest	ReliefF	0.75 ± 0.06	0.75 ± 0.05	0.76 ± 0.05	0.79 ± 0.03	0.77 ± 0.04	0.78 ± 0.06	-
aging_complex	RandomForest	Gini	0.74 ± 0.05	0.78 ± 0.05	0.76 ± 0.06	0.76 ± 0.05	0.76 ± 0.04	0.77 ± 0.05	-
aging_complex	RandomForest	Nenhum	-	-	-	-	-	-	0.78 ± 0.05
aging_complex	CART	ReliefF	0.72 ± 0.06	0.72 ± 0.07	0.74 ± 0.06	0.75 ± 0.07	0.73 ± 0.08	0.73 ± 0.09	-
aging_complex	CART	Gini	0.74 ± 0.07	0.76 ± 0.06	0.76 ± 0.05	0.75 ± 0.06	0.73 ± 0.04	0.73 ± 0.07	-
aging_complex	CART	Nenhum	-	-	-	-	-	-	0.73 ± 0.06
aging_complex	XGBoost	ReliefF	0.77 ± 0.06	0.76 ± 0.06	0.78 ± 0.07	0.78 ± 0.07	0.78 ± 0.06	0.78 ± 0.06	-
aging_complex	XGBoost	Gini	0.78 ± 0.07	0.80 ± 0.07	0.80 ± 0.06	0.79 ± 0.06	0.79 ± 0.06	0.79 ± 0.05	-
aging_complex	XGBoost	Nenhum	-	-	-	-	-	-	0.80 ± 0.07
aging_complex	SVCRadial	ReliefF	0.78 ± 0.03	0.79 ± 0.04	0.79 ± 0.05	0.79 ± 0.06	0.79 ± 0.06	0.80 ± 0.07	-
aging_complex	SVCRadial	Gini	0.79 ± 0.05	0.79 ± 0.06	0.79 ± 0.06	0.79 ± 0.06	0.79 ± 0.06	0.78 ± 0.07	-
aging_complex	SVCRadial	Nenhum	-	-	-	-	-	-	0.79 ± 0.05
aging_complex	SVCLinear	ReliefF	0.78 ± 0.04	0.77 ± 0.05	0.77 ± 0.05	0.77 ± 0.06	0.77 ± 0.05	0.78 ± 0.06	-
aging_complex	SVCLinear	Gini	0.77 ± 0.05	0.79 ± 0.05	0.76 ± 0.05	0.77 ± 0.05	0.77 ± 0.06	0.77 ± 0.07	-
aging_complex	SVCLinear	Nenhum	-	-	-	-	-	-	0.78 ± 0.05
aging_both	RandomForest	ReliefF	0.80 ± 0.06	0.81 ± 0.05	0.80 ± 0.06	0.80 ± 0.03	0.79 ± 0.06	0.80 ± 0.07	-
aging_both	RandomForest	Gini	0.79 ± 0.06	0.79 ± 0.05	0.78 ± 0.05	0.82 ± 0.07	0.80 ± 0.04	0.81 ± 0.05	-
aging_both	RandomForest	Nenhum	-	-	-	-	-	-	0.82 ± 0.04
aging_both	CART	ReliefF	0.74 ± 0.06	0.75 ± 0.07	0.73 ± 0.06	0.76 ± 0.05	0.78 ± 0.06	0.78 ± 0.06	-
aging_both	CART	Gini	0.76 ± 0.07	0.74 ± 0.07	0.75 ± 0.04	0.76 ± 0.08	0.75 ± 0.06	0.77 ± 0.05	-
aging_both	CART	Nenhum	-	-	-	-	-	-	0.77 ± 0.05
aging_both	XGBoost	ReliefF	0.81 ± 0.05	0.81 ± 0.05	0.81 ± 0.04	0.83 ± 0.04	0.82 ± 0.04	0.82 ± 0.04	-
aging_both	XGBoost	Gini	0.83 ± 0.05	0.81 ± 0.05	0.80 ± 0.04	0.81 ± 0.04	0.82 ± 0.04	0.82 ± 0.04	-
aging_both	XGBoost	Nenhum	-	-	-	-	-	-	0.81 ± 0.04
aging_both	SVCRadial	ReliefF	0.82 ± 0.03	0.83 ± 0.04	0.82 ± 0.04	0.82 ± 0.04	0.83 ± 0.04	0.82 ± 0.04	-
aging_both	SVCRadial	Gini	0.81 ± 0.03	0.81 ± 0.04	0.82 ± 0.04	0.83 ± 0.04	0.84 ± 0.04	0.83 ± 0.04	-
aging_both	SVCRadial	Nenhum	-	-	-	-	-	-	0.82 ± 0.04
aging_both	SVCLinear	ReliefF	0.81 ± 0.03	0.80 ± 0.03	0.80 ± 0.05	0.80 ± 0.06	0.79 ± 0.04	0.79 ± 0.04	-
aging_both	SVCLinear	Gini	0.80 ± 0.04	0.80 ± 0.04	0.80 ± 0.05	0.77 ± 0.05	0.77 ± 0.04	0.80 ± 0.04	-
aging_both	SVCLinear	Nenhum	-	-	-	-	-	-	0.81 ± 0.05

as amostras da classe negativa aleatoriamente entre genes que não sabemos se estão ou não ligados ao envelhecimento pode ter feito com que tivéssemos genes que são de fato ligados ao envelhecimento na classe negativa.

Ademais, ainda não está completa a catalogação de todas as funções biológicas de todos os genes humanos. De fato, o banco de dados do GO está em constantes mudanças. Quanto às interações entre proteínas, a base de dados que usamos não contém apenas interações experimentalmente verificadas em laboratórios mas também interações previstas. Todos esses pontos discutidos se traduzem em ruído em nossos *datasets*, o que pode degradar a performance dos classificadores treinados.

Por fim, uma análise mais detalhada da relação dos atributos de alta relevância - segundo os algoritmos utilizados - com o mecanismo de envelhecimento exige um conhecimento mais aprofundado na área. Isso nos leva a crer que este projeto poderia ser muito mais frutuoso se realizado com a colaboração de um biológo.

Capítulo 4

CONCLUSÃO

4.1 Contribuições

As principais contribuições deste trabalho foram (i) analisar e comparar o emprego de diferentes algoritmos de aprendizado de máquina na tarefa de identificar genes ligados ao envelhecimento, distinguindo-os dos não ligados, (ii) comparar o poder preditivo individual e combinado de diferentes tipos de atributos nessa tarefa de classificação, mais especificamente de atributos que carregam informações sobre interações entre proteínas e caracteristicas topológicas do interactoma e atributos relacionados aos papéis biológicos de genes individuais, (iii) analisar a relevância dos atributos utilizados em nossa tarefa de classificação.

De maneira geral, todos os algoritmos utilizados apresentaram bom desempenho na classificação de genes ligados ao envelhecimento, obtendo acurácias bastante superiores a 50%. Logo, isso sugere que os algoritmos foram de fato capazes de encontrar padrões que são capazes de revelar a ligação de cada gene com o processo de envelhecimento. Em particular, XGBoost e SVC com kernel radial apresentaram os melhores desempenhos, enquanto CART obteve o pior desempenho.

Quanto ao poder preditivo dos atributos, notou-se que para todos os algoritmos de aprendizado, exceto XGBoost, infelizmente a diferença de desempenho dos classificadores para diferentes *datasets* não foi estatísticamente significante. No caso de XGBoost, tanto os dois tipos de atributos em conjunto quanto os termos GO individualmente se mostraram com poder preditivo significantemente maior do que apenas para atributos relacionados a interações entre proteínas. Isso é o contrário do que era esperado, que a combinação dos dois tipos de atributos melhorasse o poder preditivo dos classificadores treinados.

Chegou-se a conclusão de que o número de interações de uma proteína e seu grau de centralidade, mais especificamente o *coreness*, no interactoma humano são relevantes para o problema de classificação de genes, o que concorda com resultados de estudos similares. Ademais, atributos de interações binárias entre proteínas se mostraram pouco relevantes, especialmente se colocados em conjunto com atributos de papéis biológicos.

Há muitas possibilidades de trabalhos que poderiam ser desenvolvidos em continuação deste. Poderia-se estudar maneiras de gerar atributos que carreguem outras informações sobre cada gene para utilizar na tarefa de classificação de sua classificação, como por exemplo informa-

ções contidas na hierarquia dos termos GO, visto que aqui apenas consideramos relações "é um", ou então informações de como genes regulam a expressão de outros genes. Seria interessante também realizar a mesma análise feita neste projeto, no entanto usando algoritmos de aprendizado com hiperparâmetros ajustados ao invés de usar os parâmetros padrão, como foi nosso caso. Outra continuação possível seria uma análise dos resultados aqui obtidos por um biólogo.

4.2 Considerações sobre o Curso de Graduação

O campo do aprendizado de máquina tem como ideia central extrair conhecimento de dados de maneira automática. Pessoalmente, considero isso fascinante. Mas não sou só eu: tem-se visto um grande interesse nessa área, tanto por parte de pessoas que querem adquirir conhecimentos quanto de empresas buscando profissionais capazes de empregar ferramentas do campo em problemas reais. Minha escolha em particular por esse tema se deu não apenas por meu interesse, mas também para suprir a falta desse tipo de conteúdo no curso de Engenharia de Computação.

Minha primeira crítica ao curso é que a carga horária é muito alta, o que acaba por deixar alunos sobrecarregados e impedir muitos alunos de desenvolverem atividades extracurriculares, que têm se tornado cada vez mais importantes no mercado de trabalho. A sobrecarga também prejudica o desempenho geral dos estudantes. Segundo, o oferecimento de disciplinas do campo de aprendizado de máquina é extremamente limitado. O curso seria enriquecido por uma oferta de mais cursos desse tipo.

Críticas a parte, acredito que o curso foi excelente, preparando-me adequadamente para minha vida profissional. O corpo docente é composto por muitos pesquisadores e professores muito bem qualificados, que muitas vezes durante minha vida acadêmica observei que também são empenhados e dispostos a auxiliar alunos quando necessário. Algo que acredito ser um diferencial no curso é que obtive uma base sólida de física, matemática e estatística.

O curso, por ser ofertado pela USP, também herda algumas de suas melhores vantagens: um amplo oferecimento de oportunidades de intercâmbio, iniciação científica com pesquisadores renomados e participação em grupos extracurriculares interessantes.

Com relação à realização deste trabalho, o curso me ofereceu a maior parte das ferramentas necessárias para seu desenvolvimento, em particular por meio das disciplinas do ciclo básico de programação, como Introdução a Ciência da Computação e Algoritmos e Estruturas de Dados.

REFERÊNCIAS

ALBERT, R.; ALBERT, I.; NAKARADO, G. Structural vulnerability of the north american power grid. **Physical review E**, APS, v. 69, n. 2, p. 025103, 2004. Citado na página 27.

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of modern physics**, APS, v. 74, n. 1, p. 47, 2002. Citado na página 27.

ASHBURNER, M.; BALL, C.; BLAKE, J.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J.; DAVIS, A.; DOLINSKI, K.; DWIGHT, S.; EPPIG, J. *et al.* Gene ontology: tool for the unification of biology. **Nature genetics**, Nature Publishing Group, v. 25, n. 1, p. 25–29, 2000. Citado na página 31.

BATAGELJ, V.; ZAVERSNIK, M. An o (m) algorithm for cores decomposition of networks. **arXiv preprint cs/0310049**, 2003. Citado na página 27.

BRANDES, U. On variants of shortest-path betweenness centrality and their generic computation. **Social Networks**, Elsevier, v. 30, n. 2, p. 136–145, 2008. Citado na página 28.

BROWN, K.; JURISICA, I. Online predicted human interaction database. **Bioinformatics**, Oxford University Press, v. 21, n. 9, p. 2076–2082, 2005. Citado na página 32.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. **CoRR**, abs/1603.02754, 2016. Disponível em: <http://arxiv.org/abs/1603.02754>. Citado 2 vezes nas páginas 23 e 34.

CONSORTIUM, G. O. *et al.* Expansion of the gene ontology knowledgebase and resources. **Nucleic acids research**, Oxford Univ Press, v. 45, n. D1, p. D331–D338, 2017. Citado na página 31.

FABRIS, F.; MAGALHÃES, J. D.; FREITAS, A. A review of supervised machine learning applied to ageing research. **Biogerontology**, Springer, p. 1–18, 2017. Citado 3 vezes nas páginas 16, 28 e 29.

FACELI, K.; LORENA, A.; GAMA, J.; CARVALHO, A. Inteligência artificial: Uma abordagem de aprendizado de máquina. **Rio de Janeiro: LTC**, v. 2, p. 192, 2011. Citado 3 vezes nas páginas 21, 22 e 23.

FREITAS, A.; VASIEVA, O.; MAGALHÃES, J. D. A data mining approach for classifying dna repair genes into ageing-related or non-ageing-related. **BMC genomics**, BioMed Central, v. 12, n. 1, p. 27, 2011. Citado 7 vezes nas páginas 16, 28, 29, 33, 34, 40 e 41.

GOLDMAN, D.; CUTLER, D.; ROWE, J.; MICHAUD, P.; SULLIVAN, J.; PENEVA, D.; OLSHANSKY, S. Substantial health and economic returns from delayed aging may warrant a new focus for medical research. **Health affairs**, Health Affairs, v. 32, n. 10, p. 1698–1705, 2013. Citado na página 15.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **J. Mach. Learn. Res.**, JMLR.org, v. 3, p. 1157–1182, mar. 2003. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944919.944968>. Citado 2 vezes nas páginas 23 e 24.

HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J. Exploring network structure, dynamics, and function using NetworkX. In: **Proceedings of the 7th Python in Science Conference (SciPy2008)**. Pasadena, CA USA: [s.n.], 2008. p. 11–15. Citado na página 33.

HANNUM, G.; GUINNEY, J.; ZHAO, L.; ZHANG, L.; HUGHES, G.; SADDA, S.; KLOTZLE, B.; BIBIKOVA, M.; FAN, J.-B.; GAO, Y. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. **Molecular cell**, Elsevier, v. 49, n. 2, p. 359–367, 2013. Citado na página 28.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer New York, 2013. (Springer Series in Statistics). ISBN 9780387216065. Disponível em: <https://books.google.com.br/books?id=yPfZBwAAQBAJ>. Citado 3 vezes nas páginas 22, 24 e 35.

HORVATH, S. Dna methylation age of human tissues and cell types. **Genome biology**, BioMed Central, v. 14, n. 10, p. 3156, 2013. Citado na página 28.

JEONG, H.; MASON, S.; BARABÁSI, A.-L.; OLTVAI, Z. Lethality and centrality in protein networks. **Nature**, Nature Publishing Group, v. 411, n. 6833, p. 41–42, 2001. Citado na página 27.

KENYON, C. The genetics of ageing. **Nature**, Nature Research, v. 464, n. 7288, p. 504–512, 2010. Citado 2 vezes nas páginas 19 e 20.

LI, J.; CHENG, K.; WANG, S.; MORSTATTER, F.; ROBERT, T.; TANG, J.; LIU, H. Feature selection: A data perspective. **arXiv:1601.07996**, 2016. Citado na página 34.

LI, Y.-H.; DONG, M.-Q.; GUO, Z. Systematic analysis and prediction of longevity genes in caenorhabditis elegans. **Mechanisms of ageing and development**, Elsevier, v. 131, n. 11, p. 700–709, 2010. Citado na página 28.

LI, Y.-H.; ZHANG, G.-G.; GUO, Z. Computational prediction of aging genes in human. In: IEEE. **Biomedical Engineering and Computer Science (ICBECS), 2010 International Conference on**. [S.l.], 2010. p. 1–4. Citado 6 vezes nas páginas 16, 28, 29, 33, 34 e 41.

LÓPEZ-BIGAS, N.; OUZOUNIS, C. Genome-wide identification of genes likely to be involved in human genetic disease. **Nucleic acids research**, Oxford University Press, v. 32, n. 10, p. 3108–3114, 2004. Citado na página 33.

LÜ, L.; ZHOU, T.; ZHANG, Q.-M.; STANLEY, H. The h-index of a network node and its relation to degree and coreness. **Nature communications**, Nature Publishing Group, v. 7, p. 10168, 2016. Citado na página 27.

MAGALHÃES, J. D. The biology of ageing: A primer. p. 24–47, 01 2011. Citado 4 vezes nas páginas 7, 15, 19 e 20.

MATLIN, A.; CLARK, F.; SMITH, C. Understanding alternative splicing: towards a cellular code. **Nature reviews Molecular cell biology**, Nature Publishing Group, v. 6, n. 5, p. 386–398, 2005. Citado na página 32.

MITCHELL, T. **Machine Learning**. [S.l.]: McGraw-Hill Science/Engineering/Math, 1997. Citado na página 20.

NAKAMURA, E.; MIYAO, K. A method for identifying biomarkers of aging and constructing an index of biological age in humans. **The Journals of Gerontology Series A: Biological Sciences and Medical Sciences**, Oxford University Press, v. 62, n. 10, p. 1096–1105, 2007. Citado na página 28.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado 2 vezes nas páginas 22 e 34.

PIRAVEENAN, M.; PROKOPENKO, M.; ZOMAYA, A. Local assortativeness in scale-free networks. **EPL (Europhysics Letters)**, IOP Publishing, v. 84, n. 2, p. 28002, 2008. Citado na página 28.

PRESS, C. **The Top Ten Algorithms in Data Mining**. CRC Press, 2009. (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series). ISBN 9781420089653. Disponível em: <https://books.google.com.br/books?id=_kcEn-c9kYAC>. Citado na página 22.

ROBNIK-ŠIKONJA, M.; KONONENKO, I. Theoretical and empirical analysis of relieff and rrelieff. **Machine learning**, Kluwer Academic Publishers, v. 53, n. 1-2, p. 23–69, 2003. Citado 2 vezes nas páginas 24 e 25.

SILVA, T.; ZHAO, L. **Machine Learning in Complex Networks**. Springer International Publishing, 2016. ISBN 9783319172897. Disponível em: <https://books.google.com.br/books?id=WdDurQEACAAJ>. Citado na página 27.

SOLÉ, R.; VALVERDE, S. Information theory of complex networks: on evolution and architectural constraints. In: **Complex networks**. [S.l.]: Springer, 2004. p. 189–207. Citado na página 28.

TACUTU, R.; CRAIG, T.; BUDOVSKY, A.; WUTTKE, D.; LEHMANN, G.; TARANUKHA, D.; COSTA, J.; FRAIFELD, V.; MAGALHÃES, J. D. Human ageing genomic resources: integrated databases and tools for the biology and genetics of ageing. **Nucleic acids research**, Oxford University Press, v. 41, n. D1, p. D1027–D1033, 2012. Citado na página 32.

VAPNIK, V. **The Nature of Statistical Learning Theory**. Springer, 1995. ISBN 9780387945590. Disponível em: <https://books.google.com.br/books?id=r_ayQgAACAAJ>. Citado na página 23.

WAN, C.; FREITAS, A.; MAGALHÃES, J. D. Predicting the pro-longevity or anti-longevity effect of model organism genes with new hierarchical feature selection methods. **IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)**, IEEE Computer Society Press, v. 12, n. 2, p. 262–275, 2015. Citado na página 28.

WUCHTY, S.; ALMAAS, E. Peeling the yeast protein network. **Proteomics**, Wiley Online Library, v. 5, n. 2, p. 444–449, 2005. Citado na página 27.