Classificando perfis de longevidade de bases de dados longitudinais usando Floresta Aleatória

G. A. Riqueti[1], C. E. Ribeiro[2], L. E. Zárate[1]

[1] Pontifícia Universidade Católica de Minas Gerais, Brazil
giovannariqueti@gmail.com,zarate@pucminas.br
[2] University of Kent, United Kingdom
cer28@kent.ac.uk

Abstract. Estudos populacionais sobre envelhecimento humano frequentemente geram bases de dados longitudinais de alta dimensionalidade. O processo de descoberta de conhecimento precisa ser adaptado para lidar com as características especiais dessas bases de dados, para se beneficiar do seu aspecto temporal. Neste trabalho, apresentamos os resultados de um processo de descoberta de conhecimento em bases de dados aplicado nos dados do *English Longitudinal Study of Ageing* (ELSA), um proeminente estudo britânico que acompanha milhares de indivíduos por um longo período de tempo, coletando informações de diferentes dimensões, como saúde, socioeconômica, e bem-estar. O objetivo do nosso estudo é classificar os participantes do estudo ELSA, de acordo com o perfil apresentado por eles, como longevos, que são indivíduos com idade acima de 82,9 anos, ou não-longevos. Para isso, foi utilizada uma abordagem de agrupamento semi-supervisionado para encontrar grupos de representantes dos perfis, e usamos esses grupos como base de dados para a execução de um algoritmo de aprendizado supervisionado. O modelo de classificação teve bons resultados, e interpretando este modelo foi constatado que aspectos de diferentes dimensões influenciam na diferenciação entre os perfis.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Data Mining; H.2.m [**Miscellaneous**]:

Keywords: data mining, knowledge discovery, random forests, supervised machine learning

1. INTRODUÇÃO

O envelhecimento humano é um tema complexo, com vários fatores genéticos e ambientais desempenhando um papel no envelhecimento biológico e nas mudanças que ocorrem na vida das pessoas, à medida que elas envelhecem. Recentemente, a demanda por conhecimento acerca do envelhecimento tem aumentado, devido ao aumento da população idosa no mundo. Com uma maior proporção de idosos na sociedade, aumenta o interesse na construção de políticas públicas, descoberta de hábitos saudáveis, e em programas sociais para aumentar o bem-estar dessa parcela da população [Malloy-Diniz et al. 2013]. Uma das iniciativas para se descobrir conhecimento acerca do envelhecimento humano são estudos populacionais longitudinais. Estudos longitudinais acompanham um conjunto fixo de pessoas que compartilham determinada característica, como idade e localização, ao longo de vários anos. São coletados valores de uma série de informações relacionadas a um domínio, repetidamente, em períodos fixos de tempo denominados ondas. As bases de dados geradas por estes estudos são bases de dados longitudinais, nas quais os atributos possuem um índice de tempo adicional, referente à onda na qual a coleta foi feita.

O objetivo deste trabalho é relatar a aplicação de um processo de descoberta de conhecimento à base de dados de um estudo populacional, o *English Longitudinal Study of Ageing* (ELSA), com o uso de técnicas de aprendizado de máquina, para descrever os perfis dos indivíduos longevos e

Copyright©2018 Permission to copy without fee all or part of the material printed in KDMiLe is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

não-longevos. Depois do pré-processamento dos dados oriundos do estudo para gerar uma base de dados composta dos exemplos de participantes longevos e não-longevos do ELSA, foram aplicados algoritmos de agrupamento para separar os melhores representantes destes perfis, e gerar uma base de dados para treinar e testar um modelo de Floresta Aleatória. Os resultados de classificação obtidos foram satisfatórios e, foram interpretados por meio de regras de associação extraídas da construção da Floresta Aleatória para descrever os perfis de indivíduos longevos e não-longevos.

2. REFERENCIAL TEÓRICO

2.1 Florestas Aleatórias

Florestas de decisão são conjuntos de Árvores de Decisão (ADs) criadas a partir de uma base de dados. O principal desafio ao gerar uma floresta de decisão é como obter uma boa variabilidade nas árvores que a compõem, levando a um maior poder de generalização (classificar instâncias desconhecidas) para o modelo. Um dos métodos de se gerar uma floresta de decisão é o algoritmo de Florestas Aleatórias (FA), introduzido por [Breiman 2001]. Em FAs, a variabilidade é obtida de duas maneiras: a) cada AD na floresta é treinada com um subconjunto das instâncias da base de dados, amostrado aleatoriamente com repetição; e b) em cada nó interno das ADs, um subconjunto dos atributos da base de dados é amostrado, para que a função de divisão avalie apenas aqueles atributos. Um algoritmo de FA possui dois parâmetros principais: *ntree*, que corresponde ao número de ADs que compõem a FA, e *mtry*, que corresponde à quantidade de atributos amostrados em cada nó interno das árvores.

Por acrescentar essa variabilidade controlada em suas ADs, as FAs são capazes de atingir boa generalização sem a necessidade do uso de métodos de poda (reduções nas ADs feitas para introduzir um poder de generalização artificialmente, geralmente com alto custo computacional), o que torna o algoritmo mais eficiente. A classificação de uma nova instância por uma FA se dá por um sistema de votação envolvendo cada AD que compõe a floresta, onde a classe escolhida por mais árvores é assinalada à instância.

Em geral, as FAs atingem boa acurácia preditiva quando comparadas a outros métodos de aprendizado de máquina supervisionados. [Fernández-Delgado et al. 2014] realizou experimentos comparando 17 famílias de classificadores (179 classificadores no total) em 121 bases de dados, e concluiu que a família das FAs obteve os melhores resultados de predição.

2.2 Interpretabilidade na Floresta Aleatória

A FA possui resultados satisfatórios quanto à acurácia de predição, no entanto, perde em interpretabilidade. Uma boa interpretação do modelo ajuda no entendimento de seu aprendizado, em uma melhor exploração dos dados e na compreensão e apropriação dos resultados [Louppe 2014]. Uma das opções de interpretação da FA é por meio da lista de variáveis e sua importância, essa importância é calculada de acordo com o ganho de informação obtido pela variável em cada nodo da AD que a utiliza. No entanto, essa alternativa ainda não oferece clareza quanto ao funcionamento da FA, e uma alternativa para se extrair informações sobre o processo do modelo e melhorar a compreensão sobre os resultados seria a geração de um novo classificador por meio da extração das regras de associação do modelo. As regras de associação buscam encontrar eventos que influenciam na ocorrência de outros eventos e representar essa dependência por meio de padrões descritivos. A extração dessas regras foi feita de acordo com as etapas descritas no artigo [Deng 2014].

Primeiramente, todas as árvores construídas para a execução da FA foram agrupadas em uma lista de árvores. Dessa lista de árvores, todas as regras de associação que representam cada caminho entre o nó raiz ao nó folha da árvore foram agrupadas. Todas essas regras foram analisadas de acordo com a sua frequência, tamanho e erro. A frequência representa a proporção das instâncias de dados em que a regra aparece. O tamanho da regra é o número de condições dessa. O erro é medido pelo número de

classificações incorretas feitas pela regra dividido pela quantidade de instâncias satisfeitas pela mesma. Tendo duas regras com mesma frequência e erro, a de menor tamanho é escolhida devido à sua melhor interpretabilidade. Em seguida, é necessário podar essas regras, o que acontece usando o critério de melhoria da métrica E que, no caso, quanto menor, melhor. Alguns dos modos de calcular o valor E são o erro de aplicar a regra ao conjunto de treinamento ou teste, ou calcular o erro pessimista. O E_i representa a qualidade da regra sem essa fazer parte do conjunto de regras e o E_0 a qualidade da regra com essa fazendo parte do conjunto. Os dois valores são usados na equação do *decay* definida na Equação 1 obtida do artigo [Deng 2014], sendo $s = 10^{-6}$.

$$decay = (E_i - E_0)/max(E_0.s) \tag{1}$$

Se esse valor obtido pela Equação 1 for menor que 0.05, então a regra é retirada de seu conjunto. O último processo para se realizar a extração das regras de maneira eficiente, é a seleção de regras. Essa seleção ocorre por meio da retirada das regras redundantes e não relevantes. Em seguida, as selecionadas são usadas para se construir um novo classificador baseado na FA. Primeiramente, todas as regras são agrupadas em uma lista chamada S, e todos os dados usados para o treinamento são agrupados em D. As regras de S são testadas com a base de dados de D. O critério de seleção é definido com base no erro mínimo da regra, e depois por sua frequência, caso ocorra empate, pelo seu tamanho. A regra escolhida como melhor passa a fazer parte da lista R e as instâncias que a satisfazem são removidas de D. O processo continua até que D esteja vazio.

3. BASE DE DADOS ELSA E PREPARAÇÃO DO CONJUNTO DE DADOS

Com o objetivo de gerar um modelo de predição que classifique as instâncias do ELSA em longevos ou não-longevos, preparamos uma base de dados para treinamento e teste com os melhores representantes dos perfis de cada classe. O processo de geração da base de dados é descrito nesta Seção.

3.1 Descrição da base ELSA

O ELSA é atualmente um dos estudos populacionais de envelhecimento mais proeminentes do mundo [Marmot et al. 2015]. O estudo conta com milhares de respondentes (todos habitantes do Reino Unido) de 50 anos ou mais, visitados de dois em dois anos (duração de uma onda do estudo) por profissionais para a coleta de dados. O ELSA teve início em 2002, e sua base de dados principal compreende variáveis demográficas, econômicas, sociais, de saúde físicas, saúde mental e psicológica, e função cognitiva [Banks et al. 2016].

Neste trabalho, foram consideradas as 6 primeiras ondas do ELSA para a geração da base de dados. Foi gerada uma base de dados com as instâncias referentes aos indivíduos que participaram do estudo durante essas ondas e, ou ultrapassaram a expectativa de vida do Reino Unido (82,9 anos[1]), ou faleceram antes de atingir essa idade. Essas instâncias foram classificadas como indivíduos longevos e não-longevos, respectivamente.

Note que apenas uma instância foi mantida para cada indivíduo nesta base de dados, para evitar redundância (em uma base de dados longitudinal, há repetidas instâncias referentes a cada indivíduo). Foram mantidos o último registro disponível, para indivíduos da classe não-longevos, e o primeiro registro disponível, para indivíduos da classe longevos. Isso foi feito para aproximar as médias de idade dos indivíduos das duas classes, com o intuito de reduzir as diferenças entre representantes de classes distintas que são reflexos da diferença de idade. Um exemplo, seria comparar a saúde de idosos de 90 anos com a saúde de idosos de 70 anos, essa análise poderia resultar que pessoas não-longevas possuem uma melhor saúde que longevas, o que é uma conclusão equivocada.

[1]Fonte: World Bank website, 2014: http://data.worldbank.org/indicator/SP.DYN.LE00.FE.IN?locations=GB

3.2 Pré-processamento da base de dados

Uma preparação adequada da base de dados reduz a distorção dos dados, auxilia no desempenho dos algoritmos de mineração de dados, e colabora para resultados mais valiosos e confiáveis dos processos de KDD [Pyle 1999]. É recomendado que as metodologias tradicionais de descoberta de conhecimento em bases de dados sejam adaptadas às características especiais dos dados longitudinais [Last et al. 2001]. Portanto, foi realizado um processo de preparação de dados na base de dados do ELSA considerando a informação temporal dos dados, para garantir que o conhecimento representado na base de dados fosse correto e suficiente, evitando dados ausente.

A base de dados passou então pelas seguintes etapas de pré-processamento (descritas em detalhes em [Ribeiro and Zárate 2017]):

—Filtragem de instâncias e atributos inconsistentes, ou considerados não confiáveis. O atributo idade é desconsiderado e retirado da base de dados.

—Aplicação da técnica de dados ausentes *Last-Observation-Carried-Forward* [Minhas et al. 2015]. Atributos com dados ausentes foram substituídos por valores do mesmo atributo, para a mesma instância, em uma onda anterior do estudo, quando disponível.

—Seleção conceitual de atributos, guiada por um estudo prévio para identificar os aspectos ambientais relacionados ao envelhecimento humano [Ribeiro et al. 2017]. Todos os atributos mantidos na base de dados descrevem aspectos utilizados em outros estudos do envelhecimento humano.

—Foi realizada uma fusão de atributos correlatos, para reduzir a dimensionalidade da base de dados. Questões dos questionários do ELSA que eram diretamente dependentes umas das outras foram fundidas em um único atributo que representa toda a informação obtida no conjunto de questões dependentes.

—Todos os atributos tiveram seus valores transformados para um valor numérico entre 0 e 1. Esta recodificação foi feita para que a base de dados pudesse ser usada como entrada nos algoritmos de agrupamento, e obedece a uma lógica interna nos atributos, onde todos os valores menores são menos desejáveis para uma vida longeva do que os maiores. Ou seja, todos os atributos têm o valor 0 para a opção de "pior caso", e o valor 1 para a opção de “melhor caso”.

—Grupos de atributos da base de dados foram então reduzidos em atributos especiais, denominados Blocos, seguindo uma metodologia usada no estudo ELSA. Cada Bloco da base de dados final representa um aspecto relacionado ao envelhecimento humano.

Ao final da etapa de preparação, foi gerada uma base de dados com 1333 participantes do ELSA, e 128 atributos representados por 28 Blocos. Das instâncias selecionadas, 242 (18%) representam a classe de não-longevos, e 1091 (82%) representam a classe de longevos. Na Tabela 1, é possível observar o significado de cada Bloco.

3.3 Seleção de Registros

O objetivo do trabalho é encontrar o perfil de indivíduos longevos e não-longevos, ou seja, representantes da classe que seguem um determinado padrão. Porém, o estudo do envelhecimento humano é muito complexo e envolve diversos fatores, como social, econômico, saúde física e mental, e genéticos. O aspecto genético não é levado em consideração na base de dados ELSA. A sua ausência juntamente com a presença de dados de diversos campos de estudos contribuem para que as características de indivíduos longevos e não-longevos sejam muito semelhantes, sendo necessário uma busca dos melhores perfis que descrevem cada classe, a fim de diferenciá-las adequadamente. Uma vez que os grupos não são facilmente discriminados, utilizamos a clusterização por densidade, o que levou a grupos com mais representantes de apenas uma única classe. Esse procedimento introduziu um carácter semi-supervisionado [Grira et al. 2004] ao processo de descrição dos perfis.

Table I. Significado dos Blocos

Blocos	Descrição	Blocos	Descrição	Blocos	Descrição	Blocos	Descrição
G	Gênero	A7	Eficiência sensorial	B5	Ansiedade e estresse	C3	Relação com crianças
A1	Frequência de atividade física	A8	Consumo de álcool e tabaco	B6	Boa autoavaliação	C4	Relação com família
A2	Problemas sérios de saúde	A9	Resultados de teste de memória	B7	Sentimentos negativos	C5	Relação com amigos
A3	Limitações físicas	B1	Sintomas de depressão	B8	Sentimentos positivos	C6	Relação com o parceiro(a)
A4	Histórico médico	B2	Questões cognitivas	B9	Satisfação com a vida e perspectiva	C7	Casa e automóvel
A5	Uso de medicamento	B3	Limitações físicas	C1	Educação formal recente	C8	Carreira
A6	Dor e sua intensidade	B4	Resultados de testes de mobilidade	C2	Estrutura social geral	C9	Status econômico atual

Primeiramente foi utilizado o algoritmo DBSCAN [Ester et al. 1996]. A escolha de um algoritmo de agrupamento baseado em densidade tem dois motivos: 1) experimentos iniciais com algoritmos baseados em distância não obtiveram resultados satisfatórios, o que pode indicar uma estrutura não-convexa de agrupamento dos dados; e 2) o número de grupos é indeterminado, portanto é preciso um algoritmo capaz de determinar esse número automaticamente. Em relação aos parâmetros usados no DBSCAN, foram variados ambos o *epsilon* e o *minpts* partindo de valores próximos de zero e aumentando gradativamente alternando os dois parâmetros, o *epsilon* de 0,01 em 0,01 e o *minpts* de 10 em 10. A cada teste, a distribuição de classes dos clusters encontrados foi usada como critério, e os parâmetros foram variados até o ponto em que a distribuição não obteve avanços quanto à sua homogeneidade. Os resultados do DBSCAN mostraram tendências para grandes grupos com a classe majoritária (longevos) e alguns grupos com maior concentração da classe minoritária (não-longevos), mas não foram encontrados grupos completamente concisos, provavelmente devido à alta dimensionalidade da base de dados e o desbalanceamento de classes.

Para refinar os resultados do DBSCAN, os dois maiores agrupamentos com boa distribuição de instâncias de cada classe foram refinados em uma segunda etapa de agrupamento. Para a classe de não-longevos, foi selecionado um agrupamento de 112 instâncias com exatamente 56 de cada classe (a melhor distribuição encontrada nos resultados do DBSCAN). Para a classe longevos, foi selecionado um grupo de 776 instâncias, onde 716 (92%) pertencem à classe dos longevos.

Estes dois grupos foram utilizados, separadamente, como entrada para uma segunda etapa de agrupamento, utilizando a versão unidimensional do algoritmo *Self-Organizing Maps* (SOM) [Kohonen and Somervuo 1998]. O SOM foi escolhido por ser capaz de explorar relações lineares e não-lineares em bases de dados de alta dimensionalidade, o que o torna capaz de detectar um refinamento destes subgrupos da base de dados [Kantardzic 2011, Chapter 9]. Quanto aos parâmetros utilizados, o *neighborhood weight* foi fixado em zero, o número de cluster foi variado de 2 a 10 e o *unexplained variance* de 0,05 em 0,05. Os clusters foram avaliados a cada teste com relação à homogeneidade de classes, e o melhor resultado foi com 0,25 de *unexplained variance*, 7 clusters para a classe dos não-longevos e 8 clusters para os longevos. Após a execução do SOM com cada grupo como entrada, foram removidos os grupos com menos de 50% de instâncias da classe-alvo.

Através deste processo de refino, foram encontrados os melhores agrupamentos de instâncias, para definirmos os perfis de participantes longevos e não-longevos do ELSA. Como indicativo de qualidade desses representantes, analisamos o diâmetro do cluster, definido como a média da distância Euclidiana quadrada entre os elementos do cluster [Zaït and Messatfa 1997]. Conseguimos um diâmetro total

de 0,23 nos longevos, e de 0,2 para os representantes dos não-longevos, ambos considerados bons resultados por ser abaixo de 0.3. O ponto de referência para avaliação foi escolhido de acordo com o valor indicado para o índice *Silhouette*, que padroniza o valor 0.7 e valores acima como indicadores de um bom cluster, no entanto, a medida do diâmetro considera 0 como o melhor resultado e o valor 0.7 é invertido para 0.3. A medida do índice *Silhouette* que considera a distância inter-cluster não obteve bons resultados por causa da proximidade entre os elementos da base de dados, como já mencionado antes. Como representantes da classe de não-longevos, foi selecionado um grupo de 51 instâncias, com 71% delas pertencendo à classe de não-longevos, e como representantes da classe de longevos, foi selecionado um grupo de 723 instâncias, com 95% destas pertencendo aos longevos.

Concluindo a etapa de agrupamento do nosso trabalho, foi realizado um teste estatístico para comprovar que os grupos de representantes dos participantes de cada classe são realmente partes de populações diferentes, confirmando a existência dos perfis destas classes. O teste realizado foi o teste T-Quadrado de Hotelling, a versão multivariada do Teste T de duas amostras, da estatística univariada, que determina se as médias populacionais de duas variáveis aleatórias são iguais [Hotelling 1992]. O teste rejeitou a hipótese nula, de que as médias populacionais são iguais, confirmando que os representantes selecionados para as classes longevo e não-longevo são padrões de população diferentes e, consequentemente, podem ser usados como base de dados para o treinamento e teste de um algoritmo de aprendizado supervisionado.

4. CLASSIFICAÇÃO FLORESTA ALEATÓRIA E INTERPRETAÇÃO DOS RESULTADOS

A fim de identificar as causas que levam uma pessoa a ser longeva ou não-longeva, o conjunto de dados obtido na Seção anterior é usado para treinar e testar o método de classificação supervisionado Floresta Aleatória, sendo 75% da base para treinamento e 25% para teste. Não foi utilizada validação cruzada uma vez que essa técnica particiona o conjunto de dados em subconjuntos, o que não é adequado para a base de dados em questão por essa possuir poucas instâncias, cerca de 100 exemplos. As FAs possuem uma boa acurácia de predição, o que contribui para que os resultados sejam mais precisos e que as causas sejam identificadas com mais exatidão. Para interpretar os resultados e melhor compreender o aprendizado da FA, foram extraídas as regras de associação geradas na construção do modelo.

Para decidir os parâmetros usados na criação das FAs, sendo os principais deles, *ntree* e *mtry*, primeiro realizou-se um experimento avaliativo com o valor de *mtry* fixo para variar a quantidade de árvores. No entanto, a quantidade de árvores não alterou significativamente a acurácia, e o valor escolhido foi de $ntree = 100$ árvores visando um baixo custo computacional [Oshiro et al. 2012]. Fixado o valor de 100 árvores, foi testado valores de *mtry* variando de 3 a 16, os resultados mostraram valores compreendidos entre 92.86% e 100%. Os valores utilizados que apresentaram melhor acurácia foram o de 14 e 16, obtendo acurácia 100%. O resultado encontrado é o esperado devido ao uso da FA, que aumenta a acurácia, e pelo fato da base de dados já ter sido pré-processada de maneira eficiente em etapas anteriores.

A FA encontrada como a melhor(parâmetros $ntree$=100 e $mtry$=14), foi testada com 10 sementes distintas escolhidas para o seu processo aleatório. Durante a montagem das ADs, de acordo com a semente que é escolhida de maneira aleatória é selecionado o conjunto de variáveis a serem analisadas para cada nó. O teste com várias sementes é necessário para determinar se os resultados obtidos pela combinação dos parâmetros apresentam um comportamento convergente. Uma confirmação de que os parâmetros são adequados é que a lista de variáveis com maior importância obtidas em cada experimento se repete, com pequenas variações, assim como a acurácia.

Após encontrar os melhores parâmetros, o modelo foi treinado e testado, obtendo os resultados mostrados na Tabela 2 referente às métricas de avaliação. As regras de associação foram extraídas de acordo com a metodologia mostrada na Seção 2.2 e podem ser visualizadas na Tabela 3. Essas regras podem ser interpretadas de acordo com a Tabela 4. Quanto à interpretação dos Blocos, o B4

Table II. Métricas de avaliação dos resultados da Floresta Aleatória

Classe	**Precisão**	**Sensibilidade**	**Medida-F**	**Porcentagem na base de dados**
Longevos	1.0	1.0	1.0	50%
Não-longevos	1.0	1.0	1.0	50%

Table III. Regras de associação extraídas da Floresta Aleatória

Tamanho	**Frequência**	**Erro**	**Condições**	**Previsão**
3	45,95%	0%	**A2 >4 & B4 >3,825 & C2 >4,995**	**Longevo**
1	27,03%	0%	**C8 <= 8,5**	**Não longevo**
2	5,41%	0%	**A1 >6,875 & A4 >4,5**	**Longevo**
1	21,62%	6,25%	**Outras condições**	**Não longevo**

Table IV. Análise das condições

A1 >6,875	O indivíduo pratica regularmente atividades físicas, combinando exercícios leves, moderados e vigorosos. Exercícios leves e rigorosos são praticados em uma frequência moderada, uma vez por semana, e exercícios moderados são praticados em uma frequência maior, mais de uma vez por semana.
A2 >4	O indivíduo não possui nenhuma doença crônica de coração e também não possui problemas psiquiátricos.
A4 >4,5	O indivíduo não recebeu tratamento de câncer nos últimos 2 anos.
B4 >3,825	O indivíduo não apresenta dificuldade para caminhar 400 metros sozinho ou possui pouca dificuldade, mas consegue caminhar sem o auxílio de algum objeto. O entrevistado não reclama de problemas quanto à sua mobilidade.
C2 >4,995	O indivíduo possui amigos e algum membro da família ainda vivo, ou possui muitos membros da família ainda vivos e não tem amigos.
C8 <= 8,5	O indivíduo não está aposentado e não possui trabalho; ou está aposentado, não podendo ter um trabalho além da aposentadoria ou realizar trabalho voluntário.

representa a dificuldade do indivíduo de caminhar 400 metros sozinho com ou sem o auxílio de um equipamento, esse Bloco pode ter grande influência quanto à longevidade devido à pouca mobilidade do entrevistado poder atrapalhar na execução de exercícios físicos, tema também abordado em A1. O C2 aponta para o problema da solidão entre os idosos e o C8 demonstra a importância do idoso se manter ativo, seja por trabalho remunerado ou voluntário. Ambos os Blocos A2 e A4 estão relacionados a doenças, sendo o A2 referente a doenças crônicas de coração ou distúrbios psiquiátricos, e sendo o A4 referente se o paciente recebeu tratamento de câncer nos últimos dois anos.

5. CONCLUSÕES

Esse trabalho propõe um processo de descoberta sobre uma base de dados longitudinal de envelhecimento humano para que possamos identificar indivíduos longevos ou não-longevos. A base de dados utilizada ELSA é formada por 6 ondas e envolve milhares de atributos e gravações. Por meio de um pré-processamento a quantidade de atributos foi diminuída e os atributos foram separados em Blocos que representam um aspecto relacionado ao envelhecimento humano. Utilizando essa base de dados já pré-processada, usamos o método de classificação supervisionada FA para identificar perfis que melhor descrevem os longevos e não-longevos. A utilização da FA apresentou resultados satisfatórios, entretanto, esse classificador é caracterizado por ser mais descritivo do que interpretável. Para uma melhor interpretabilidade, extraímos as regras de associação existentes dentro do modelo e as analisamos para que os resultados sejam entendidos de forma prática. No entanto, o classificador obtido por meio dessa extração perde um pouco da acurácia obtida pela FA.

Por meio das regras obtidas, e sua interpretação, é possível obter *insights* quanto ao que leva uma pessoa a ser longeva ou não-longeva, podendo inspirar possíveis políticas públicas e aplicações no mercado. Possíveis aplicações no mercado seriam o melhor cálculo de preços de plano de saúde e

seguro de vida. Em relação às políticas públicas, é preciso analisar cada Bloco individualmente. Sobre a temática abordada em B4 e A1, uma possível solução seria a construção em praças e diversos locais públicos de espaços propícios a exercícios físicos por meio da instalação de equipamentos de ginástica e alongamento. Quanto ao C2, uma maneira de resolver a questão seria a criação de programas sociais que incentivem o trabalho voluntário de visitas regulares a idosos. Para o problema abordado em C8, uma alternativa seria a criação de programas sociais no qual o idoso seja o voluntário. Já os Blocos A2 e A4 destacam a necessidade de investir no diagnóstico precoce e tratamento das doenças relatadas.

Agradecimentos

Os dados foram disponibilizados através do *UK Data Archive*. O ELSA foi desenvolvido por uma equipe de pesquisadores baseados no *NatCen Social Research*, no *University College London* e no *Institute for Fiscal Studies*. Os dados foram coletados pela *NatCen Social Research*. O financiamento é fornecido pelo *National Institute of Aging* nos Estados Unidos, e por um consórcio de departamentos governamentais do Reino Unido coordenados pelo Oce para Estatísticas Nacionais. Os desenvolvedores e financiadores do ELSA e do *Archive* não têm qualquer responsabilidade pelas análises ou interpretações aqui apresentadas. Os autores também agradecem o apoio da FAPEMIG no desenvolvimento deste trabalho, através da concessão de bolsa de pesquisa.

REFERENCES

Banks, J., Breeze, E., Lessof, C., and Nazroo, J. *The dynamics of ageing: Evidence from the English Longitudinal Study of Ageing 2002-15 (Wave 7)*. Institute for Fiscal Studies, 7 Ridgmount Street London WC1E 7AE, 2016.

Breiman, L. Random forests. *Machine learning* 45 (1): 5–32, 2001.

Deng, H. Interpreting tree ensembles with intrees. *arXiv preprint arXiv:1408.5456*, 2014.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*. Vol. 96. pp. 226–231, 1996.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research* 15 (1): 3133–3181, 2014.

Grira, N., Crucianu, M., and Boujemaa, N. Unsupervised and semi-supervised clustering: a brief survey. *A review of machine learning techniques for processing multimedia content, Report of the MUSCLE European Network of Excellence (FP6)*, 2004.

Hotelling, H. The generalization of student's ratio. In *Breakthroughs in Statistics*. Springer, pp. 54–65, 1992.

Kantardzic, M. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

Kohonen, T. and Somervuo, P. Self-organizing maps of symbol strings. *Neurocomputing* 21 (1): 19–30, 1998.

Last, M., Klein, Y., and Kandel, A. Knowledge discovery in time series databases. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 31 (1): 160–169, 2001.

Louppe, G. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

Malloy-Diniz, L., Fuentes, D., and Cosenza, R. *Neuropsicologia do Envelhecimento: Uma Abordagem Multidimensional*. Vol. 1, 2013.

Marmot, M., Oldfield, Z., Clemens, S., Blake, M., Phelps, A., Nazroo, J., Steptoe, A., Rogers, N., and Banks, J. English longitudinal study of ageing: Waves 0-6, 1998-2013. [data collection]. 23rd edition, 2015.

Minhas, S., Khanum, A., Riaz, F., Alvi, A., Khan, S. A., Initiative, A. D. N., et al. Early alzheimer's disease prediction in machine learning setup: Empirical analysis with missing value computation. In *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, pp. 424–432, 2015.

Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. How many trees in a random forest? In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, pp. 154–168, 2012.

Pyle, D. *Data preparation for data mining*. Vol. 1. Morgan Kaufmann, 1999.

Ribeiro, C. E., Brito, L. H. S., Nobre, C. N., Freitas, A. A., and Zárate, L. E. A revision and analysis of the comprehensiveness of the main longitudinal studies of human aging for data mining research. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7 (3), 2017.

Ribeiro, C. E. and Zárate, L. E. Data preparation for longitudinal data mining: a case study on human ageing. *Journal of Information and Data Management* 7 (2): 116, 2017.

Zaït, M. and Messatfa, H. A comparative study of clustering methods. *Future Generation Computer Systems* 13 (2-3): 149–159, 1997.