

Universidade de São Paulo

Faculdade de Saúde Pública

**Aplicação de algoritmos de *Machine Learning* na avaliação do  
consumo alimentar: resultados da linha de base do Estudo  
Longitudinal de Saúde do Adulto (ELSA-Brasil)**

Vanderlei Carneiro da Silva

Tese apresentada ao Programa de  
Epidemiologia da Faculdade de Saúde  
Pública da Universidade de São Paulo para  
obtenção do título de Doutor em Ciências.

Linha de Pesquisa: Epidemiologia de  
doenças e agravos à saúde (LP1).

Orientador: Profa. Dra. Isabela J M Benseñor.

São Paulo

2021

**Aplicação de algoritmos de *Machine Learning* na avaliação do  
consumo alimentar: resultados da linha de base do Estudo  
Longitudinal de Saúde do Adulto (ELSA-Brasil)**

Vanderlei Carneiro da Silva

Tese apresentada ao Programa de  
Epidemiologia da Faculdade de Saúde  
Pública da Universidade de São Paulo para  
obtenção do título de Doutor em Ciências.

Linha de Pesquisa: Epidemiologia de  
doenças e agravos à saúde (LP1).

Orientador: Profa. Dra. Isabela J M Benseñor.

Versão Original

São Paulo

2021

É expressamente proibida a comercialização deste documento tanto na sua forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título e ano da tese.



## Dedicatória

À minha mãe, irmã, aos meus  
sobrinhos e amigos pelo incentivo.

## **Agradecimentos**

Aos professores da Faculdade de Saúde Pública e de outras unidades da Universidade de São Paulo que tanto me ensinaram ao longo desses últimos anos.

Às professoras Regina Fisberg e Dirce Marchioni que desde os meus primeiros meses nesta instituição compartilharam suas experiências e me apresentaram o universo complexo e fascinante da avaliação do consumo alimentar.

À professora Lígia M. Cavalheiro por sua amizade e disponibilidade em permitir que eu pudesse construir um diálogo prático entre tecnologia e nutrição. O amadurecimento deste trabalho, também, resulta da experiência na disciplina “Inquéritos Alimentares”.

Aos amigos que tive a oportunidade de conhecer e conviver. A troca de aprendizados, culturas e inquietações foram enriquecedores. Carinhosamente agradeço a Carol, Batanero, Chiara, Fran e Tânia.

Ao Lukas Marinho pelo apoio e incentivo constante. Companheiro de todas as horas.

Às professoras Ana Paula Sayuri, Daniela Canella e ao professor Itamar Santos pela colaboração em suas participações na banca como suplentes.

À professora Tatiana Toporcov por seus ensinamentos, colaboração em diferentes momentos e por me fazer despertar o interesse pela epidemiologia.

Ao professor Antônio Vidal por sua participação na banca e pelo aprendizado adquirido em sua disciplina de “Análise Preditiva de Dados”, origem do projeto e que me permitiu pensar, construir, experimentar, buscar e desenvolver este trabalho.

À professora Bartira Gorgulho que, mesmo antes da sua atual condição como docente, se prontificou em colaborar com as minhas análises. Me fez acreditar que juntos somos mais fortes, podemos ir mais longe e a essência do ensino também envolve generosidade.

À professora e amiga Julicristie Machado – Juli – por me fazer acreditar, anos atrás, que seria possível. Muito obrigado.

Ao professor Paulo Lotufo por me receber no ELSA-Brasil, esse extraordinário estudo, e permitir que eu desfrutasse do valor da ciência e da análise de dados.

Finalmente, à minha orientadora Isabela Benseñor por acreditar em mim, neste trabalho, por compartilhar o seu tempo, experiência e conhecimentos. Muito obrigado por sua paciência e por me ajudar a crescer.

*O aprendizado deve ser livre, assim como os  
nossos corações e à ciência.*

*Que os sonhos guiem e nos permitam  
compartilhar. Que as pessoas sejam ponte,  
sejam parceiras.*

*Que o ensino seja livre, sempre, a quem deseja,  
a quem busca. Ensinar exige ternura e respeito.*



## RESUMO

Silva, V.C. da. **Aplicação de algoritmos de *Machine Learning* na avaliação do consumo alimentar: resultados da linha de base do Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil)** [Tese de doutorado]. Programa de Pós-Graduação em Epidemiologia, Faculdade de Saúde Pública, Universidade de São Paulo; 2021.

**Introdução:** A avaliação do consumo alimentar permite gerar conhecimento sobre a alimentação de indivíduos e populações, além de identificar os determinantes e tendências no consumo. Com ela é possível planejar ações, orientar serviços e implementar políticas públicas de saúde adequadas as necessidades da população. Com o apoio da tecnologia é possível automatizar algumas etapas do processo de análise de dados, com redução do tempo e recursos necessários, especialmente em grandes grupos. Entretanto, em países como o Brasil, ainda são escassas as aplicações de algoritmos de *machine learning* na avaliação da dieta. **Objetivo:** Aplicar algoritmos de *machine learning* na avaliação do consumo alimentar de servidores públicos em um grande estudo brasileiro. **Métodos:** Este estudo analisou transversalmente os dados da linha de base do Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil). A partir destes dados, para explorar e classificar padrões alimentares, foi utilizado o algoritmo de cluster – *K-Means*. Na sequência, quatro algoritmos preditivos – *Support Vector Machines (SVM)*, *Decision Trees (DT)*, *Naïve Bayes (NB)*, *K-Nearest Neighbours (Knn)* – foram aplicados incluindo variáveis demográficas, socioeconômicas e clínicas para prever padrões alimentares. Adicionalmente, Sistemas de Recomendações foram construídos com algoritmos de Filtragem Colaborativa Baseada em Usuário e Itens (UBCF / IBCF) para o aconselhamento personalizado de dieta. As análises foram realizadas com a utilização do ambiente R. **Resultados:** Dois padrões alimentares foram derivados na amostra. O primeiro padrão, rotulado como “Padrão Ocidental”, no qual os participantes apresentaram ingestões médias superiores para cereais refinados, feijões, carnes vermelhas e processadas, leite e produtos lácteos com alto teor de gorduras e bebidas adoçadas, quando comparados aqueles incluídos no outro padrão. O segundo padrão, rotulado como “Padrão Prudente”, os participantes apresentaram consumo superior de frutas, vegetais, cereais integrais, aves, peixes, leite e produtos lácteos com redução de gorduras. Para a construção dos Sistemas de Recomendações foi fixado o limite de cinco itens, por participante, para evitar

recomendações extensas e inespecíficas sobre a dieta (precisão entre 90% [IBCF] e 91% [UBCF]). **Conclusão:** Através da aplicação de algoritmos de *machine learning* foi possível realizar a análise de dados sobre o consumo, prever padrões e personalizar recomendações sobre a dieta. Com o apoio das técnicas utilizadas, é possível subsidiar profissionais na gestão e no planejamento de ações de educação alimentar e nutricional personalizadas.

**Descritores:** análise de dados, *clustering*, sistema de recomendação, dieta, epidemiologia nutricional.

## ABSTRACT

Silva, V.C. da. **Application of Machine Learning algorithms in the assessment of food consumption: baseline results from the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil)** [Thesis]. Programa de Pós-Graduação em Epidemiologia, Faculdade de Saúde Pública, Universidade de São Paulo; 2021.

**Introduction:** The evaluation of food consumption allows generating knowledge about the diet of individuals and populations, in addition to identifying the determinants and trends in consumption. With it is possible to plan actions, guide services and implement public health policies appropriate to the needs of the population. With the support of technology, it is possible to automate some stages of the data analysis process, reducing the time and resources needed, especially in large groups. However, in countries like Brazil, the applications of machine learning algorithms in diet assessment are still scarce. **Objective:** Apply machine learning algorithms in the evaluation of food consumption by public servants in a large Brazilian study. **Methods:** This study cross-sectionally analyzed the baseline data from the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil). From these data, to explore and classify dietary patterns, the cluster algorithm K-Means was used. Next, four predictive algorithms - Support Vector Machines (SVM), Decision Trees (DT), Naïve Bayes (NB), K-Nearest Neighbors (Knn) - were applied including demographic, socioeconomic and clinical variables to predict dietary patterns. Additionally, Recommendation Systems were built with User- and Items-Based Collaborative Filtering algorithms (UBCF / IBCF) for personalized diet advice. The analyzes were performed using the environment R. **Results:** Two dietary patterns were derived in the sample. The first pattern, labeled as “Western Pattern”, in which the participants had higher average intakes for refined cereals, beans, red and processed meats, milk and dairy products with a high fat content and sweetened drinks, when compared to those included in the other pattern. The second pattern, labeled “Prudent Pattern”, participants showed a higher consumption of fruits, vegetables, whole grains, poultry, fish, milk and dairy products with reduced fats. For the construction of the Recommender Systems, a limit of five items was set, per participant, to avoid extensive and unspecific recommendations on the diet (accuracy between 90% [IBCF] and 91% [UBCF]). **Conclusion:** Through the application of machine learning algorithms, it was possible to perform data analysis on consumption, predict patterns and personalize diet recommendations. With the support

of the techniques used, it is possible to subsidize professionals in the management and planning of personalized food and nutrition education actions.

**Keywords:** data analysis, clustering, recommendation system, diet, nutritional epidemiology.

## Sumário

1. INTRODUÇÃO .....	20
1.1. Justificativa.....	21
2. OBJETIVOS.....	23
2.1. Geral: .....	23
2.2. Específicos:.....	23
3. REFERENCIAL TEÓRICO.....	24
2.1. Fontes de Dados sobre Consumo Alimentar.....	24
2.2. Análise de Dados sobre a Dieta .....	28
2.3. Tecnologias Aplicadas à Avaliação do Consumo Alimentar.....	30
2.4. Algoritmos para Análise de Dados baseados em <i>Machine Learning</i> .....	33
2.5. Sistemas de Recomendação.....	36
2.6. Agenda de Alimentação e Nutrição no Brasil no contexto da inovação .....	38
4. METODOLOGIA .....	42
4.1. Desenho do estudo .....	42
4.2. Amostra .....	42
4.3. Coleta de dados .....	45
4.4. Preditores demográficos, socioeconômicos e clínicos .....	46
4.5. Avaliação da dieta .....	47
4.6. Análise estatística .....	48
4.7. Descrição dos modelos de análise e dos algoritmos utilizados.....	48

4.8. Treinamento e validação dos modelos.....	57
4.9. Limitações do estudo .....	58
4.10. Pontos fortes do estudo .....	59
4.11. Aspectos éticos .....	60
4.12. Conflito de interesses.....	60
4.13. Fontes de financiamento .....	61
5. RESULTADOS.....	62
5.1. Padrões Alimentares – Artigo 1 .....	63
5.2. Sistema de Recomendações – Artigo 2 .....	96
6. CONCLUSÕES E CONSIDERAÇÕES FINAIS.....	124
7. REFERÊNCIAS BIBLIOGRÁFICAS.....	126
ANEXO 1 – Questionário de Frequência Alimentar.....	142
ANEXO 2 – Carta CONEP.....	163
ANEXO 3 – Aprovação CEP HU .....	166
ANEXO 4 – Aprovação CEP FSP.....	169
COMPROVANTE DE SUBMISSÃO – ARTIGO 1.....	173
COMPROVANTE DE SUBMISSÃO – ARTIGO 2.....	174
CURRÍCULO LATTES DO ALUNO .....	175
CURRÍCULO LATTES DA ORIENTADORA.....	176

## LISTA DE FIGURAS

Figura 1. Modelo conceitual dos determinantes do consumo alimentar. ....	25
Figura 2. Tipos de aprendizagem. ....	34
Figura 3. Aplicação de motores de recomendação. ....	37
Figura 4. Amostra do estudo. ....	44

## LISTA DE QUADROS

Quadro 1. Descrição dos algoritmos utilizados na etapa de aprendizagem supervisionada. ....	56
---	----



## LISTA DE ABREVIATURAS E SIGLAS

- CC – Circunferência da Cintura
- CEP – Comitê de Ética em Pesquisa
- CQ – Circunferência do Quadril
- DM – Diabetes Melito
- DT – Decision Trees
- ELSA-Brasil – Estudo Longitudinal de Saúde do Adulto
- HAS – Hipertensão Arterial Sistêmica
- Kcal – Kilo calorias
- Knn – K-Nearest Neighbours
- ML – Machine Learning
- MS – Ministério da Saúde
- NB – Naïve Bayes
- OMS – Organização Mundial de Saúde
- PNAN – Política Nacional de Alimentação e Nutrição
- QFA – Questionário de Frequência Alimentar
- R24h – Recordatório de 24 horas
- RA – Registro Alimentar
- RCQ – Relação Cintura-Quadril
- RS – Recommender System
- SISVAN – Sistema de Vigilância Alimentar e Nutricional
- SVM – Support Vector Machines
- TCLE – Termo de Consentimento Livre e Esclarecido
- TIC – Tecnologia da Informação e Comunicação

VIGITEL – Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico

## APRESENTAÇÃO

Esta tese foi elaborada como requisito para obtenção do título de Doutor em Ciências no Programa de Pós-Graduação em Epidemiologia (PPG-EPI) da Faculdade de Saúde Pública da Universidade de São Paulo. Ela está em conformidade com as diretrizes estabelecidas pelo programa e o formato escolhido para a sua apresentação, dentre os reconhecidos pelo PPG-EPI, é o de coletânea de artigos. A tese encontra-se organizada com os seguintes capítulos: 1. Introdução. 2. Objetivos geral e específicos. 3. Referencial teórico, onde são desenvolvidos alguns temas para o embasamento do trabalho, incluindo avaliação do consumo alimentar, as inovações tecnológicas, *machine learning*, sistemas de recomendações e a agenda de alimentação e nutrição no Brasil no contexto da inovação. 4. Materiais e métodos. 5. Resultados – apresentação dos artigos produzidos e atualmente submetidos em revistas arbitradas. 6. Conclusões e considerações finais.

## 1. INTRODUÇÃO

Com o desafio de enfrentar os desfechos associados à carência e os excessos na ingestão de alimentos, a avaliação do consumo alimentar é fundamental, tanto na prática clínica como em estudos epidemiológicos (HÉBERT et al., 2014; SATIJA et al., 2015). A partir dela é possível identificar as características da dieta e os determinantes do consumo, bem como estudar a relação entre alimentação e saúde. O desenvolvimento de políticas públicas de alimentação e nutrição e a avaliação de estratégias de intervenção, em muitos casos, são orientados pelos trabalhos publicados nesta área (MOZAFFARIAN; ROSENBERG; UAUY, 2018).

Nos últimos anos, aplicações analíticas com Inteligência Artificial e *Machine Learning* (ML) têm revolucionado as relações sociais e culturais globalmente (SPICER; SANBORN, 2019). A evolução de ferramentas analíticas tem proporcionado o desenvolvimento de modelos sensíveis e precisos em diversas áreas, incluído a nutrição (ZHANG et al., 2015; ZHU et al., 2011). Com isso, a avaliação do consumo alimentar encontrou novos parceiros; os algoritmos. Especialistas em saúde, alimentação e ciência da computação têm trabalhado na aplicação de técnicas de análise de dados baseadas em algoritmos que são capazes de modelar e de aprender sobre padrões complexos (PANARETOS et al., 2018). Aplicações digitais com sistemas inteligentes já podem reconhecer e classificar alimentos em imagens a partir de dispositivos móveis, bem como realizar a análise de grandes bancos de dados de inquéritos populacionais sobre padrões de dieta (ILLNER et al., 2012; MCCULLOUGH, 2018).

O panorama atual reforça a epidemiologia como uma ciência fundamental para a descrição de eventos de interesse à saúde e a identificação dos seus determinantes em populações, incluindo os hábitos alimentares (GLYMOUR; BIBBINS-DOMINGO,

2019; SATIJA; HU, 2014). Gestores e profissionais de saúde podem ser apoiados por sistemas informatizados com modelos preditivos para triagem de grupos prioritários para intervenções em saúde e prevenção de doenças crônicas. Publicações nos Estados Unidos (NAVARRO SILVERA et al., 2014), Reino Unido (GIABBANELLI; ADAMS, 2016), Japão (SHIOKAWA et al., 2016) e Grécia (KASTORINI et al., 2013), já apresentam alguns dos avanços na avaliação do consumo alimentar com o apoio da tecnologia.

No Brasil, servidores públicos em instituições de ensino e pesquisa em três regiões do país são acompanhados no Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil) (SCHMIDT et al., 2015). Este estudo prospectivo e multicêntrico foi proposto, inicialmente, com o objetivo de estudar a incidência e os fatores associados com o desenvolvimento de diabetes e doenças cardiovasculares na população brasileira. Na linha de base do estudo foram realizados exames e a aplicação de questionários padronizados (BENSENOR et al., 2013). O banco de dados do ELSA-Brasil permite análises com informações sobre o consumo alimentar, aspectos sociais, saúde e a construção de modelos preditivos preenchendo uma importante lacuna no Brasil sobre aplicação de algoritmos na avaliação da dieta.

### **1.1. Justificativa**

A avaliação do consumo alimentar é uma atividade fundamental no cuidado em saúde, através dela é possível conhecer a alimentação de indivíduos e grupos. Neste aspecto, a análise de regressão, frequentemente utilizada em estudos epidemiológicos, permite identificar fatores associados a um dado evento de

interesse. No entanto, é limitada ao lidar com relações não lineares observadas em inquéritos alimentares (HEARTY; GIBNEY, 2008a; PANARETOS et al., 2017).

Algoritmos baseados em *Machine learning* ainda são pouco utilizados na epidemiologia nutricional, no entanto oferecem a oportunidade de modelar problemas complexos com centenas de variáveis preditoras. Embora resgatem muitos dos conceitos aplicados na matemática e estatística, apresentam bom desempenho mesmo quando há interação entre variáveis ou quando identificados ruídos nos dados (DUGAN et al., 2015; PANARETOS et al., 2018).

A hipótese desta pesquisa é que o treinamento de algoritmos baseados em *Machine Learning* e a construção de modelos preditivos, permitirá classificar os indivíduos e grupos em padrões de dieta com base em dados indiretos sobre o seu consumo alimentar. Com os resultados desta pesquisa também se pretende subsidiar gestores e profissionais na implementação de ações personalizadas. A gestão dos cuidados em saúde e nutrição adequada às necessidades dos indivíduos, pode auxiliar na implementação de mudanças gradativas na alimentação, mais efetivas, de modo que novas práticas sejam incorporadas com a promoção da saúde e a prevenção do risco de doenças crônicas.

## 2. OBJETIVOS

### 2.1. Geral:

- ✓ Aplicar algoritmos de análise de dados baseados em *machine learning* para avaliação do consumo alimentar de servidores públicos no Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil).

### 2.2. Específicos:

- ✓ Predizer padrões alimentares a partir de dados demográficos, socioeconômicos e clínicos.
- ✓ Propor sistema de recomendações para personalizar o aconselhamento alimentar.
- ✓ Comparar o desempenho dos algoritmos aplicados utilizando diferentes métricas de avaliação.

### 3. REFENCIAL TEÓRICO

#### 2.1. Fontes de Dados sobre Consumo Alimentar

A dieta é um fator de exposição complexo para mensuração e avaliação em estudos epidemiológicos, porém de grande importância dada a sua relação com a saúde. As dietas inadequadas, com predominância de produtos ultra-processados, alta densidade energética e pobre em alimentos *in natura*, exercem impactos na qualidade de vida das pessoas e demandam ações dos sistemas de saúde pública globalmente, em razão do aumento das prevalências das doenças crônicas não transmissíveis e complicações decorrentes do excesso de peso (MOZAFFARIAN; ROSENBERG; UAUY, 2018). Desta forma, a avaliação do consumo alimentar permite conhecer as condições de alimentação e nutrição de indivíduos e grupos e, se realizadas continuamente, monitorar tendências na população (SATIJA et al., 2015).

Os resultados dos inquéritos de saúde, que são conduzidos em diferentes grupos e escalas, têm orientado a implementação de ações e políticas públicas para a proteção da saúde, redução do risco e prevenção de doenças como obesidade, hipertensão, câncer, diabetes e doenças cardiovasculares (ANDRADE et al., 2013). Esse conjunto de doenças crônicas não transmissíveis são relacionadas à uma ampla rede de determinantes sociais e biológicos, incluindo as práticas alimentares que os indivíduos adotam ao longo da vida (BOEING, 2013).

A dieta é influenciada por muitos fatores, cuja análise requer métodos sensíveis para mensurar a sua diversidade de possibilidades. O estudo da alimentação contribui para a compreensão de comportamentos, estilos de vida e riscos à saúde entre os estratos da população (PRENTICE; HUANG, 2018). A Figura 1 apresenta os principais fatores relacionados com a alimentação.





Figura 1. Modelo conceitual dos determinantes do consumo alimentar.<sup>1</sup>

<sup>1</sup> Fonte: Story M, Kaphingst KM, Robinson-O'Brien R, Glanz K. Creating Healthy Food and Eating Environments: Policy and Environmental Approaches. *Annu Rev Public Health*. 2008;29(1):253-272. doi:10.1146/annurev.publhealth.29.020907.090926

Além dos aspectos citados, considerando que alterações fisiológicas com impactos na saúde não são identificadas pela ingestão observada em um único dia, especialmente pela variabilidade dos nutrientes nos alimentos e da própria alimentação intra-indivíduo, o objetivo de grande parte dos inquéritos alimentares é permitir avaliação da dieta habitual (FREEDMAN et al., 2004). Há intenso debate sobre o desafio de obter relato sobre o consumo com o mais elevado nível de precisão possível. Dentre os métodos desenvolvidos, os mais utilizados são o Recordatório de 24 horas (R24h), Registro Alimentar (RA)/Diário Alimentar (DA) e o Questionário de Frequência Alimentar (QFA) (SHIM; OH; KIM, 2014). Contudo, independentemente do método utilizado para a coleta dos dados, é reconhecido que todos apresentam limitações e o foco está na aproximação da alimentação habitual (DODD et al., 2006).

Por um outro lado, a combinação de diferentes métodos de coleta e o uso de avaliações repetidas são estratégias frequentemente adotadas para superar essas limitações (VERLY-JR et al., 2012). Em relação a estimativa da realidade que se busca, o acompanhamento do indivíduo pelo entrevistador diretamente no momento que a alimentação ocorre, permitiria reduzir imprecisões no relato sobre os itens e quantidades consumidas, no entanto é inviável por custos elevados e interferiria na alimentação do indivíduo. Assim, métodos estatísticos têm sido desenvolvidos para reduzir a variabilidade nos dados e estimar a ingestão habitual (GUENTHER; KOTT; CARRIQUIRY, 1997; HARTTIG et al., 2011; TOOZE et al., 2006).

Apesar disso, recordar informações sobre a alimentação envolve a habilidade de o entrevistado detalhar o local de preparo, onde os alimentos foram consumidos, horários, composição das refeições, itens e quantidade de alimentos ingeridos (PEREIRA; SICHIERI, 2007). Essa habilidade costuma variar segundo faixa etária, sendo menos precisa nos extremos de idade. As crianças por desconhecimento sobre

composição e preparo das refeições, além de o vocabulário ser incompleto, adolescentes por menor disposição e os idosos por prejuízos cognitivos e memória (FISBERG et al., 2005).

Descrever o que se come também é influenciado pela forma como cada pessoa enxerga a sua alimentação, em alguns casos, reflete a expectativa da pessoa e se distancia daquela real. Distorções entre o relato e as práticas adotadas nem sempre são intencionais, elas podem resultar da imagem que o indivíduo constrói sobre a sua alimentação (GARCIA, 2004). Algumas pessoas reagem conforme são abordadas, tendo o entrevistador um papel importante. Assim, no momento da entrevista é desaconselhado o direcionamento, manifestação ou qualquer tipo de juízo por parte do entrevistador que implique em omissões ou alterações no relato. A imparcialidade nos questionamentos deve garantir que o entrevistado não se sinta pressionado em responder o que acredita ser o correto, bem como relatar segundo as expectativas do pesquisador (BERTIN et al., 2006).

Além dos aspectos já citados, é importante destacar que a estimativa do consumo depende não apenas da memória, a qualidade da medida também é influenciada pela compreensão do método e dos objetivos da pesquisa (SHIM; OH; KIM, 2014). Compreender as perguntas durante o processo de coleta dos dados é um outro aspecto fundamental, grupos com menor nível de escolaridade podem apresentar maiores dificuldades de participar de inquéritos sem a presença de um entrevistador (COSTA et al., 2006). Além disso, questões ambíguas, linguagem excessivamente técnica e o número de perguntas podem reduzir a taxa de participação e prejudicar a precisão das informações (CHIARA et al., 2007; PEREIRA; SICHIERI, 2007).

O tempo necessário para concluir a pesquisa está relacionado com a disponibilidade de participar do inquérito (LIBERATO; BRESSAN; HILLS, 2009). A busca para alcançar a melhor e mais completa informação pode levar a métodos complexos, com extenso número de perguntas e elevado grau de detalhamento (CHIARA et al., 2007). Nestes casos, o volume de dados coletados pode não ser válido. O esforço para descrever as características da dieta tende a diminuir com listas extensas ou em razão de muitos dias de registros. Além disso, um grande volume de dados exige recursos e logística para coleta, tabulação, armazenamento, visualização e análise (GIABBANELLI; ADAMS, 2016).

Além disso, o padrão de dieta socialmente reconhecido como ideal, caracterizado por alimentos classificados como “bons” e “ruins” ou “permitidos” e “proibidos” podem pressionar o participante do inquérito a superestimar o consumo de alimentos como frutas, legumes e verduras e subestimar alimentos com alto grau de processamento e elevado conteúdo de açúcares, gorduras e sal (GARCIA, 1997; SCAGLIUSI; LANCHI, 2003).

## **2.2. Análise de Dados sobre a Dieta**

Feitas essas considerações, depois de coletados os dados, a fase de análise não é uma tarefa menos desafiadora. A dieta é influenciada por muitos fatores com atuação em rede, destacam-se os aspectos demográficos, sociais, econômicos, clínicos, percepção de saúde, diferenças de gênero, cultura, religião, disponibilidade e oferta de alimentos, aspectos ambientais e regulatórios, sistemas alimentares, entre outros (SILVA et al., 2010). Para avaliar a dieta, independentemente do parâmetro

utilizado, a amplitude de análise não se esgota com a abordagem escolhida (CANESQUI, 2009).

Os métodos de avaliação do consumo alimentar avançaram nos últimos anos com o desenvolvimento de ferramentas híbridas, que se debruçam sobre aspectos quantitativos e qualitativos. O principal avanço está associado ao reconhecimento de que a avaliação de nutrientes e alimentos isoladamente são abordagens limitadas (HU, 2002; PANARETOS et al., 2017). Portanto, faz-se necessário considerar diversos componentes da dieta dos indivíduos, resumir e comparar com o conhecimento científico disponível, bem como diretrizes dietéticas (BOWMAN et al., 1998; KREBS-SMITH et al., 2018).

O campo de estudo da alimentação e nutrição combina conhecimento de muitas áreas: ciências da saúde, bioquímica, dietética, epidemiologia, economia, sociologia, antropologia, filosofia e política (DE VASCONCELOS, 2010; VASCONCELOS, 2007). A quantidade de métodos resultantes permite abordar a alimentação e responder diferentes perguntas com o enriquecimento do debate, sem que se pretenda impor nível de importância entre eles. No entanto, os estudos têm se debruçado, principalmente, sobre os métodos estatísticos para derivar padrões alimentares a partir de informações diretas sobre o consumo (DEVLIN et al., 2012; THORPE et al., 2016). Aspectos demográficos, socioeconômicos e condição de saúde são informações relacionadas com a dieta e poderiam auxiliar gestores e profissionais de saúde quando dados sobre o consumo não estão disponíveis (GIABBANELLI; ADAMS, 2016; PANARETOS et al., 2018; PENCINA et al., 2009).

### 2.3. Tecnologias Aplicadas à Avaliação do Consumo Alimentar

A Tecnologia da Informação e Comunicação (TIC) tem contribuído com o desenvolvimento e gestão de várias áreas, fornecendo ferramentas, *softwares*, sistemas, aplicativos, *hardware*, processadores, entre outros (FORNAZIN; JOIA, 2015). Uma de suas aplicações em nutrição envolve o processamento de informações necessárias para o diagnóstico, acompanhamento e planejamento de ações (CARVALHO et al., 2013). Desta forma, a coleta, tabulação, armazenamento, análise e apresentação de dados relacionados a alimentação humana ganharam reforço com a informática, tanto em nível individual como coletivo (CURIONI; BRITO; BOCCOLINI, 2013).

Há no mercado muitos *softwares* que permitem avaliar o consumo de alimentos e preparações. Além disso, o auxílio de porções de referência e medidas caseiras, como usualmente são relatadas pelos pacientes, otimizam o processo de análise da dieta, bem como os cálculos de ingestão de nutrientes com base em tabelas de composição de alimentos (CUPPARI; ANÇÃO, 2007). A complexidade dessa atividade se amplia com a oferta de uma diversidade alimentos industrializados com redução de calorias, gorduras, açúcares, glúten, lactose e acréscimo de substâncias para o seu enriquecimento. Uma outra vantagem é a possibilidade de trabalhar com preparações com diferentes opções de receitas. A relevância no uso desses *softwares* não é apenas cálculo mais ágil de todo esse volume de dados, mas também redução de erros pelo entrevistador e entre os entrevistadores com a padronização da coleta. Por outro lado, é necessário que os aplicativos disponham de bancos de dados válidos e atualizados (CRISPIM; SAMOFAL; FERREIRA, 2019).

Versões mais recentes de *softwares* oferecem módulos para antropometria, registro de exames e testes laboratoriais, condutas, evolução e rotinas para

exportação de bancos de dados para análise em outros programas estatísticos (GURINOVIC et al., 2010, 2018; KRAVCHYCHYN et al., 2018). Imagens fotográficas de alimentos com tamanhos diversos auxiliam a estimativa da porção ingerida durante a coleta de dados (BARUFALDI et al., 2016).

Um outro tipo de aplicação são as plataformas interativas para usuários de dispositivos móveis, com lembretes e sugestões para alimentação mais saudável, em alguns casos associada à prática de atividade física (KRAVCHYCHYN et al., 2018). Considerando o uso atual de *smartphones* e a cultura de registrar todos os momentos, incluindo as refeições, aplicativos em dispositivos móveis podem coletar um grande volume de dados sobre o que é consumido, independente da memória do entrevistado, realizar segmentação, predição sobre itens consumidos, estimar quantidade e o valor nutricional da refeição (ZHANG et al., 2015; ZHU et al., 2011).

Com o avanço da tecnologia houve maior facilidade de conexão entre as pessoas e dispositivos eletrônicos, além de maior agilidade no processamento e disponibilidade de grandes volumes de dados para análise (SIDEY-GIBBONS; SIDEY-GIBBONS, 2019). O progresso tem refletido no número de publicações sobre tecnologia (WONG et al., 2018), saúde (PASHAZADEH; NAVIMIPOUR, 2018) e dieta (WEBER; ACHANANUPARP, 2016). Neste sentido, o desenvolvimento de técnicas na área de Inteligência Artificial tem impulsionado pesquisas para construção de modelos relacionando informações sobre alimentação e o desenvolvimento de doenças crônicas ou ainda pela utilização de algoritmos para automatizar o processo de avaliação do consumo alimentar (KASTORINI et al., 2013; NAVARRO SILVERA et al., 2014; RICHTER; KHOSHGOFTAAR, 2018).

Para a condução de inquéritos alimentares são necessários recursos, logística, tempo e profissionais, especialmente quando envolvem muitas pessoas e medidas repetidas da dieta. O uso da linguagem de programação de computadores tem contribuído, sobretudo na expansão da capacidade de manipulação e análise de grandes bases (HAND; MANNILA; SMITH, 2001). No entanto, ferramentas baseadas em tecnologia, comparadas aos métodos tradicionais, não eliminam o viés que é inerente ao relato sobre o consumo de alimentos.

Apesar do crescente interesse, ainda são poucos os estudos de comparação e avaliação crítica da validade das medidas de dieta coletadas com o apoio de inovações tecnológicas (ILLNER et al., 2012; THOMPSON et al., 2015). Illner et al. em revisão da literatura descreveram que os principais grupos de inovações tecnológicas utilizados em estudos epidemiológicos compreendem à assistentes pessoais digitais (PDA), telefone móvel, plataformas baseadas em computador e na *web*, gravadores de voz e dispositivos com câmera embutida, sensores e *scanner* (ILLNER et al., 2012). De acordo com os autores, as vantagens dessas inovações são a redução do tempo e trabalho para coleta de dados, especialmente em grandes grupos e populações dispersas geograficamente, além de melhorar a qualidade do relato por depender menos da memória do entrevistado. Timon et al. acrescentam que a tecnologia oferece soluções para muitas das limitações das metodologias tradicionais na avaliação da dieta; menor custo, maior taxa de participação e acurácia na coleta de dados (TIMON et al., 2016). Thompson et al. encontraram que os participantes preferiram métodos autoaplicáveis que os administrados por um entrevistador. O sucesso dessas abordagens tem sido atribuída a fatores como possibilidade de a entrevista ocorrer de forma remota em um ambiente neutro e uma



sequência padronizada de perguntas, bem como auxílio na avaliação de porções de alimentos em formato digital (ILLNER et al., 2012).

Tecnologias baseadas em Inteligência Artificial e *Machine Learning* também têm apresentado resultados promissores na epidemiologia nutricional, especialmente pelo poder preditivo e capacidade de subsidiar a tomada de decisão na assistência e orientação de pacientes (AGAPITO et al., 2018; AWAYSHEH et al., 2019; NOROUZI et al., 2018). Há algumas iniciativas para classificação de indivíduos de acordo com a sua ingestão (GIABBANELLI; ADAMS, 2016; HEARTY; GIBNEY, 2008a), além do uso de ferramentas tecnológicas para coleta de dados com maior padronização e qualidade das informações (BOUSHEY et al., 2017; CRISPIM; SAMOFAL; FERREIRA, 2019).

#### **2.4. Algoritmos para Análise de Dados baseados em *Machine Learning***

A capacidade da inteligência humana para a resolução de problemas despertou, há muito tempo, o desejo de reproduzir nas máquinas esse comportamento (CONTRERAS; VEHI, 2018; HEARTY; GIBNEY, 2008a). Inteligência Artificial é a ciência que tem por objetivo simular a inteligência humana em sistemas artificiais, para que possam executar tarefas em eficiência equivalente ou superior ao homem (CONTRERAS; VEHI, 2018). *Machine Learning* ou aprendizagem de máquina, é um subcampo da Inteligência Artificial que combina conhecimentos da matemática, estatística, ciência da computação, linguagem de programação de computadores, educação e neurociência. Ela estuda a construção de modelos por meio do treinamento de algoritmos a partir de instâncias de dados que representam um problema real (WITTEN; FRANK; HALL, 2011). Em outras palavras, trata-se da análise de dados que permita encontrar funções matemáticas que expliquem a relação

entre variáveis e deem, ao computador, habilidade de aprender e fazer previsões (HAND; MANNILA; SMITH, 2001).

O aprendizado realizado com as máquinas busca simular a forma como os seres humanos aprendem. Um conjunto de atributos é fornecido para o algoritmo para que ele aprenda sobre determinado problema a ser resolvido e, uma vez treinado e construído o modelo preditivo, espera-se generalizar o aprendizado em dados desconhecidos (CONTRERAS; VEHI, 2018).

Aprendizado envolve a capacidade de transformação, adaptação e evolução. *Machine Learning* desenvolve modelos por meio de processos iterativos que encontram padrões e, em alguns casos, relacionamentos ocultos nos dados, auxiliando o processo de tomada de decisão. Ademais, é possível ajustar as análises customizando parâmetros, atribuindo pesos para os acertos e penalizações aos erros, de modo que a taxa de aprendizado aumente (WITTEN; FRANK; HALL, 2011). Dentre as classificações existentes, a aprendizagem de máquina é dividida basicamente em três tipos, conforme apresentado na Figura 2.



Figura 2. Tipos de aprendizagem.

O aprendizado supervisionado, o primeiro tipo, é muito utilizado e caracteriza-se pelo treinamento do algoritmo com o fornecimento das variáveis de saída (*target*) e dos preditores. No segundo, denominado de não supervisionado, apenas os atributos são entregues. Neste caso, o próprio algoritmo se encarrega de encontrar funções e padrões nos dados, o que permite a modelagem de problemas cuja variável

de saída não é conhecida previamente pelo pesquisador (ALANAZI; ABDULLAH; QURESHI, 2017). Já o terceiro tipo, aprendizado por reforço, envolve a tomada de decisão com base na interação da máquina com um agente externo, fornecendo *feedback* como premiações e penalizações. Nesse último caso, o sistema é guiado e navega pelo espaço de possibilidades para resolução do problema, aumentando a sua aprendizagem, tal como ocorre quando computadores aprendem a jogar interagindo com um jogador (DUNJKO; BRIEGEL, 2018).

A avaliação do modelo e do processo de aprendizagem ocorre com a divisão dos dados em subconjuntos de treino e teste. Apesar de não existirem valores fixos, usualmente 70% dos dados são utilizados para treinamento do algoritmo e 30% para teste. Nos dados de teste as previsões são realizadas e comparadas com os valores reais ou previamente disponíveis. Independentemente de a previsão ocorrer com variáveis contínuas ou em tarefas de classificação, com base nos erros e acertos do modelo, acurácia e outras métricas de avaliação são calculadas (RICHTER; KHOSHGOFTAAR, 2018).

Assume-se que após a construção do modelo, a máquina pode realizar inferência em dados desconhecidos generalizando o aprendizado obtido, assim como seres humanos recorrem ao seu *background* quando expostos a situações que exigem tomada de decisões (DUNJKO; BRIEGEL, 2018).

O poder preditivo e o potencial de uso dessa tecnologia aumentaram muito nos últimos anos. A maioria das técnicas já existem há várias décadas, no entanto com a expansão do volume de dados disponíveis para análise e a evolução da informática com estruturas mais eficientes para armazenamento e processamento ágil dos dados, atualmente os algoritmos alcançam alto desempenho na modelagem de problemas

complexos, com baixo custo e em diversas áreas (PASHAZADEH; NAVIMIPOUR, 2018; WONG et al., 2018).

Dentre as aplicações, citam-se a análise e reconhecimento de imagens, análise de sentimento baseada em texto, reconhecimento de voz, treinamento de carros autônomos, predição de fraudes em transações financeiras, desenvolvimento de sistemas de recomendação para compras, campanhas de marketing, detecção de falhas em equipamentos (HAN; KAMBER; PEI, 2011; WITTEN; FRANK; HALL, 2011). Em saúde, diagnóstico de doenças, análise de dados de consumo alimentar e predição de qualidade da dieta (HEARTY; GIBNEY, 2008a; MEHTA; PANDIT, 2018; RAJKOMAR; DEAN; KOHANE, 2019; WONG et al., 2018).

## **2.5. Sistemas de Recomendação**

Sistemas de Recomendação (SR) são técnicas para análise de dados que fornecem aos usuários recomendações sobre produtos ou serviços. O campo de SR tem suas origens na década de 1990 (ADOMAVICIUS; TUZHILIN, 2005). As recomendações podem ser baseadas em vários parâmetros, tais como, itens populares, características dos usuários como localização geográfica ou outra informação demográfica (SASE et al., 2015). Recentemente, estes sistemas têm utilizado algoritmos de *machine learning* obtendo resultados muito eficientes (PORTUGAL; ALENCAR; COWAN, 2018).

Os sistemas de recomendação têm recebido crescente atenção por suas aplicações relevantes e personalizadas. Para citar alguns exemplos, há sistemas de recomendações para livros, música, redes sociais e transportes (CHEUNG et al., 2019). Entretanto, apesar do seu uso extensivo em domínios de *e-commerce* e lazer,

sua aplicação na saúde ainda encontra-se no início (HORS-FRAILE et al., 2018a). A Figura 3 ilustra a arquitetura e aplicação de um sistema de recomendações para promoção de serviços/produtos.



Figura 3. Aplicação de motores de recomendação.

O método utilizado na maioria dos sistemas de recomendação mostra a informação mais relevante por usuário (WIESNER; PFEIFER, 2014). Por exemplo, as pessoas podem estar dispostas a experimentar e incluir novos alimentos na dieta se receberem recomendações personalizadas. Neste sentido, um estudo prévio tem mostrado que 95% dos usuários pesquisados gostariam obter recomendações de alimentos mais saudáveis e melhorar a qualidade da dieta (GE; RICCI; MASSIMO, 2015).

Não obstante, embora a utilização de sistemas de saúde possa poupar tempo e reduzir a quantidade de recursos necessários em casos de tratamentos, a automatização da avaliação da dieta e o aconselhamento alimentar com o apoio tecnológico não devem substituir o contato e a interação do paciente com um profissional habilitado (KHAN; HOFFMANN, 2003; MUSTAQEEM; ANWAR; MAJID, 2020). Por um outro lado, sistemas de recomendação talvez possam ser utilizados sob a supervisão de um especialista para criar intervenções individualizadas,

reduzindo custos e promovendo um estilo de vida mais saudável na população (HORS-FRAILE et al., 2018a).

O sistema pode ser altamente valioso para profissionais que precisam de uma abordagem clínica baseada em dados (GAYVERT; MADHUKAR; ELEMENTO, 2016; TSOUKALAS; ALBERTSON; TAGKOPOULOS, 2015). Na verdade, os sistemas de recomendação podem ser usados para auxiliar na gestão do cuidado em saúde (CHEUNG et al., 2019). O especialista humano é capaz de identificar com mais precisão os componentes da dieta do paciente sujeitos ao manejo, sem realizar orientação extensa e inespecífica (KELDERS et al., 2015).

Considerando a orientação da sociedade para tecnologias modernas, como atendimento personalizado e inteligente, recomendações para o consumo de alimentos saudáveis são úteis para orientar consumidores e profissionais (NEALE; TAPSELL, 2019). O aconselhamento alimentar apoiado por sistemas de recomendação pode ser mais eficiente (NOROUZI et al., 2018). Entretanto, este conhecimento sozinho tem efeitos relativamente limitados sobre o comportamento e os sistemas podem não ser suficientes para mudar a dieta usual e estratégias integradas pelo conhecimento do especialista com o apoio da tecnologia são necessárias para ações mais adequadas (MOZAFFARIAN; ROSENBERG; UAUY, 2018).

## **2.6. Agenda de Alimentação e Nutrição no Brasil no contexto da inovação**

A promoção da alimentação adequada e saudável é reconhecida como fator importante para redução do risco de doenças crônicas, como diabetes e doenças cardiovasculares e de morbidade e mortalidade associadas (WHO, 2016). Nesse

sentido, a Política Nacional de Alimentação e Nutrição (PNAN) já em sua primeira edição, em 1999, clamava por estratégias que assegurassem o direito humano a alimentação adequada pela população brasileira (MS, 1999). Mais recentemente, em 2011, em sua segunda edição, a política atual as demandas em saúde e nutrição, reforçou a “Promoção da Alimentação Adequada e Saudável” como a sua segunda diretriz (MS, 2011).

Deste modo, a utilização de algoritmos baseados em *machine learning* pode auxiliar a análise dados de inquéritos alimentares, extrair padrões e identificar os determinantes (barreiras) do consumo (PANARETOS et al., 2018). As vantagens desse método sobre as técnicas de análise já utilizadas consistem em extrair padrões não óbvios, otimizar modelos, recursos tecnológicos e humanos necessários para avaliação de grandes bases de dados, especialmente inquéritos populacionais, além de automatizar algumas tarefas e auxiliar profissionais de saúde e tomadores de decisão (BOLAND; BRONLUND, 2019).

No contexto da PNAN, o fomento a estratégias envolvendo Tecnologias da Informação e Comunicação e a avaliação do consumo, tem como propósito a melhoria das condições de alimentação e nutrição da população, mediante a promoção de práticas alimentares adequadas e saudáveis e se constitui como uma oportunidade para qualificar e aperfeiçoar a “Vigilância Alimentar e Nutricional” (VAN).

A “Vigilância Alimentar e Nutricional” e a “Pesquisa, Inovação e Conhecimento em Alimentação e Nutrição” são outras duas diretrizes da PNAN que justificam a inserção de tecnologias mais avançadas em sistemas de informação em saúde. Essas tecnologias podem ser utilizadas não só com a finalidade acadêmica, mas de aperfeiçoamento em plataformas como o SISVAN web (NASCIMENTO; SILVA;

JAIME, 2019). Esse sistema, embora com longa trajetória, enfrenta desafios no sentido de ampliar a sua utilização como guia para organização da “Atenção Nutricional” na atenção básica, uma outra diretriz da PNAN. A reorientação do cuidado em alimentação e nutrição pode ser aprimorada por modelos preditivos incorporados em plataformas de apoio às equipes de saúde da família para identificação de indivíduos, grupos e territórios em situação de vulnerabilidade e insegurança alimentar.

Neste sentido, é necessário o fortalecimento das ações de alimentação e nutrição no Sistema Único de Saúde, especialmente no tocante a promoção de práticas que considerem não apenas o componente biológico e nutricional. Há relevância para o acesso a alimentação como um direito que deve ser garantido sem prejuízo de outros direitos sociais previstos na Constituição Federal (BRASIL, 1988). Este reconhecimento está entre as bases diretivas da Lei 11.346/2006 que cria o Sistema de Segurança Alimentar e Nutricional (BRASIL, 2006) e Decreto 7.272/2010 que regulamenta a Lei e institui a Política Nacional de Segurança Alimentar e Nutricional (BRASIL, 2010).

Sistemas de Recomendação permitem analisar dados individuais e apresentar sugestões (HAHSLER, 2017). Não se trata de recomendar alimentos e preparações populares, mas de avaliar as características do indivíduo, considerar práticas coletivas e culturais, alinhando ao disposto em documentos como o Guia Alimentar para a População Brasileira (MS, 2014) e incentivar escolhas que priorizem alimentos frescos, minimamente processados e que sejam sustentáveis econômica e ambientalmente. No caso em tela, algoritmos de recomendação também podem ser implementados em sistemas como o SISVAN. Dado um conjunto de marcadores do consumo e de possíveis indicações, eles podem destacar recomendações adaptadas



às diferentes realidades dos indivíduos e integradas às políticas vigentes. O sistema deve ser implementado e orientado para subsidiar um aconselhamento amplo, gradual e que preserve a autonomia dos indivíduos para fazer escolhas alimentares adequadas e saudáveis, conforme previsto no Marco de Referência de Educação Alimentar e Nutricional para as Políticas Públicas (MDS, 2012).

Finalmente, cumpre destacar que essa área de estudo, embora com grande potencial de uso para aperfeiçoamento do cuidado em alimentação e nutrição, não se configura (e nem se pretende) suficiente, tampouco titular da prescrição. Porém, complementar ao conhecimento e técnicas existentes para apoiar profissionais de saúde. Sendo assim, deve ser reconhecida, experimentada e validada pelo especialista humano, sob pena de resultar em predições contraproducentes para a saúde e alimentação.

## **4. METODOLOGIA**

### **4.1. Desenho do estudo**

O estudo possui desenho transversal com os dados da linha de base do Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil). Foram analisadas as informações coletadas na primeira fase de avaliações do ELSA-Brasil, realizada entre agosto de 2008 e dezembro de 2010. Na segunda coleta de dados, realizada após quatro anos de acompanhamento, questões sobre a dieta não foram incluídas no questionário. Portanto, não foi possível estabelecer um desenho prospectivo e os dados analisados são provenientes da linha de base do estudo. Na terceira coleta de dados, nove anos após a linha de base, informações sobre dieta foram coletadas, mas ainda não estavam disponíveis para análise por ocasião do desenvolvimento desta tese.

### **4.2. Amostra**

O ELSA-Brasil (SCHMIDT et al., 2015) é um estudo longitudinal, multicêntrico, recrutou 15.105 servidores públicos de instituições de ensino e pesquisa em seis cidades (Belo Horizonte, Porto Alegre, Rio de Janeiro, Salvador, São Paulo e Vitória) em três diferentes regiões no Brasil. O estudo foi criado a partir de uma chamada pública lançada, em 2005, pelo Ministério da Saúde e pelo Ministério da Ciência e Tecnologia. Além disso, é a maior pesquisa multicêntrica de coorte realizada por um país fora do eixo dos países desenvolvidos. Participam do ELSA-Brasil: a Universidade Federal da Bahia (UFBA), a Universidade Federal de Minas Gerais (UFMG), a Universidade Federal do Espírito Santo (UFES), a Fundação Oswaldo Cruz (Fiocruz, Rio de Janeiro), a Universidade de São Paulo (USP) e a Universidade Federal do Rio Grande do Sul (UFRGS).

Os funcionários e docentes responderam questionários sobre suas condições gerais de saúde, história familiar de doenças, uso de medicamentos, saúde mental, entre outros tópicos. Além disso, foram submetidos a uma série de exames laboratoriais e físicos. Originalmente a coorte foi desenvolvida com o objetivo de estudar a incidência de diabetes e doenças cardiovasculares na população brasileira e os seus determinantes biológicos e sociais (DUNCAN et al., 2012; SCHMIDT et al., 2015).

Na linha de base, trabalhadores ativos e aposentados, homens e mulheres, de qualquer das seis instituições foram elegíveis. Foram recrutados para participar do estudo servidores públicos, entre 35 e 74 anos, sendo excluídos: aqueles com intenção de deixar a instituição, gestantes ou com gestação prévia inferior a quatro meses, com dificuldades de comunicação ou prejuízo cognitivo grave e, se aposentado, viver fora da região metropolitana correspondente.

Para a realização deste estudo foram incluídos todos os participantes da primeira onda, de ambos os sexos, que concluíram o Questionário de Frequência Alimentar (n=15.105, 100%). Como critério de seleção, a partir do cálculo da dieta, os participantes com ingestão energética diária implausível, ou seja, inferior a 500 ou superior a 4.000 Kcal/dia foram excluídos (n=2.438, 16%) (WILLET, 1998a). Na Figura 4, a composição da amostra final do estudo é apresentada, n=12.667 participantes, 5.217 (41%) homens e 7.450 (59%) mulheres.

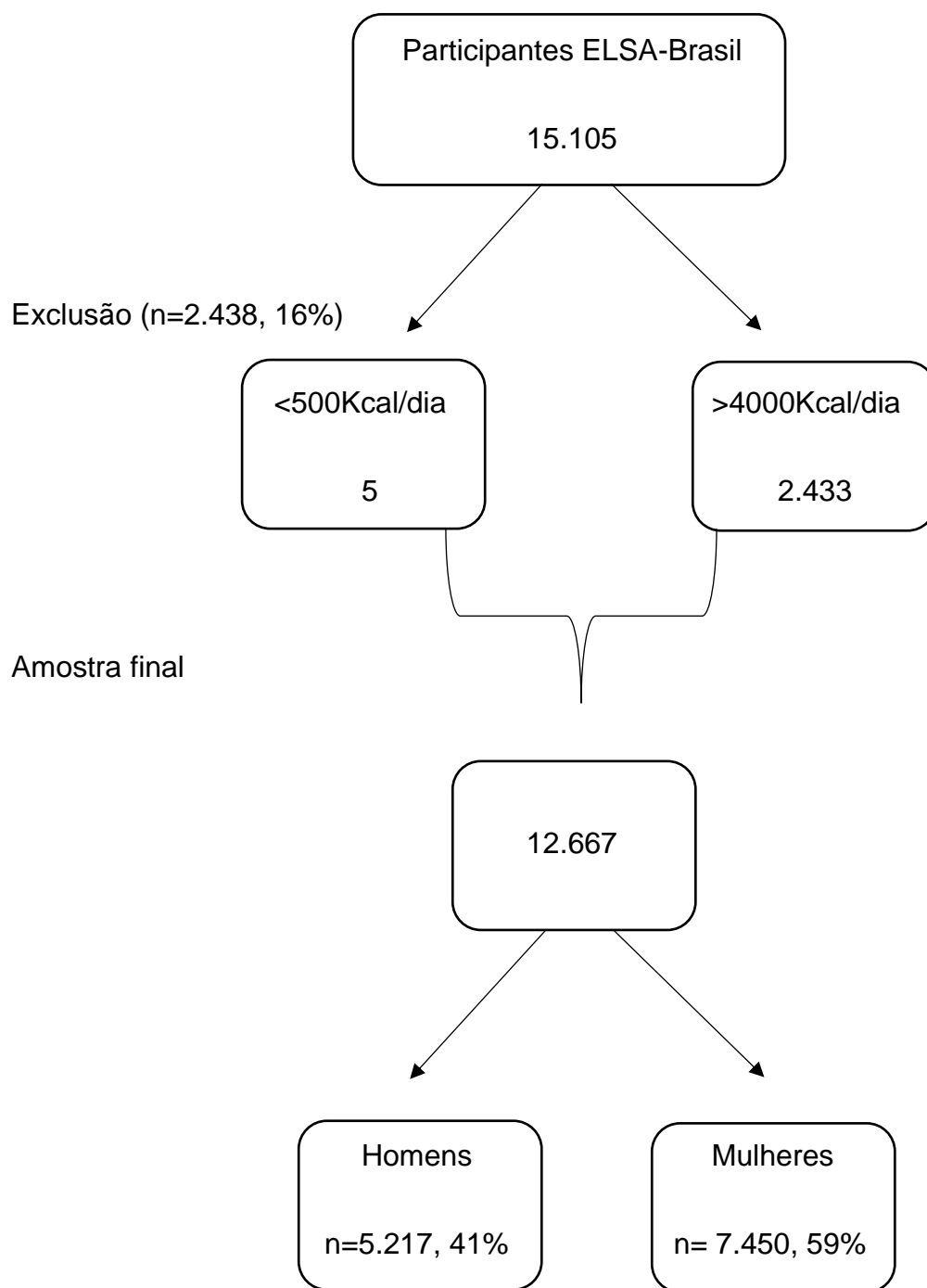


Figura 4. Amostra do estudo.

### 4.3. Coleta de dados

O ELSA-Brasil incluiu em seu protocolo de entrevistas a aplicação de questionários para coleta de informações demográficas, sociais e de saúde, além de os participantes realizarem exames de sangue, urina e avaliação antropométrica (CHOR et al., 2013; SCHMIDT et al., 2015). Todos os participantes foram convidados a comparecer em um centro de pesquisa clínica vinculado ao estudo para as entrevistas e testes clínicos. Esse procedimento assegurou um alto padrão de qualidade nos dados coletados (BENSENOR et al., 2013).

O peso corporal e a altura foram coletados com o participante vestindo roupas leves e sem sapatos. O peso foi mensurado com a utilização de uma balança calibrada (Toledo 2096PP) e precisão de 0.1Kg, enquanto a altura mensurada com um estadiômetro (Seca-SE-216) e precisão de 0.1cm. O Índice de Massa Corporal (IMC) foi calculado pela divisão do peso do participante em quilogramas pela altura em metros ao quadrado ( $\text{Kg/m}^2$ ). A circunferência da cintura (CC) foi coletada no ponto médio entre a crista ilíaca e a linha inferior da última costela, com uma fita métrica inelástica e precisão de 0.1cm. A circunferência do quadril (CQ) foi mensurada no ponto de maior circunferência na região glútea. O cálculo da relação cintura-quadril (RCQ) foi realizado pela divisão da cintura pelo quadril em centímetros.

A pressão arterial foi mensurada com um oscilômetro validado (Omron HEM 705CPINT). Três medidas foram coletadas no intervalo de um minuto e a média dos dois últimos valores foi considerada como definição para o valor de hipertensão.

#### 4.4. Preditores demográficos, socioeconômicos e clínicos

A análise resultou em predições a partir destas variáveis:

- ✓ Sexo (masculino vs. feminino);
- ✓ Idade (em anos);
- ✓ Escolaridade (fundamental [ou menos], médio, superior);
- ✓ Aposentadoria (não vs. sim);
- ✓ Raça/cor (branco, pardo, preto, outros [asiático, indígena]);
- ✓ Renda per capita (categorizada em tercís [utilizando US\$ 1.00 = R\$ 2.00 valor aproximado para conversão à época]);
- ✓ Se vive sozinho (não vs. sim);
- ✓ Estado civil (solteiro vs. casado);
- ✓ Tabagismo (nunca fumou, ex-fumante, fumante atual);
- ✓ Nível de atividade física (sedentário, insuficientemente ativo, ativo [utilizando tempo de lazer na versão longa do Questionário Internacional de Atividade Física – IPAQ]);
- ✓ Avaliação de saúde autorreferida (bom, regular, ruim);
- ✓ Local do centro de pesquisa clínica;
- ✓ IMC (em Kg/m<sup>2</sup>);
- ✓ RCQ (em cm);
- ✓ Dislipidemia (não vs. sim [LDL ≥130mg/dL ou uso de hipolipemiantes]);
- ✓ Hipertensão arterial (não vs. sim [pressão sistólica ≥140mm Hg, pressão diastólica ≥90mm Hg ou tratamento com anti-hipertensivo nas duas últimas semanas]);

- ✓ Diabetes (não vs. sim [diagnóstico prévio de diabetes, uso de medicação para tratamento de diabetes, glicose sanguínea  $\geq 126$  mg/dL, glicose 2h  $\geq 200$  mg/dL ou HbA<sub>1c</sub>  $\geq 6.5\%$ ]);
- ✓ Doença cardiovascular (não vs. sim [definida como infarto prévio autorreferido, acidente vascular encefálico ou revascularização]);

#### **4.5. Avaliação da dieta**

Para coleta de dados sobre a dieta foi utilizado um Questionário de Frequência Alimentar (QFA) semi-quantitativo (MOLINA et al., 2013b). O QFA utilizado no ELSA-Brasil possui uma lista que foi previamente estabelecida com 114 alimentos/preparações. O participante foi questionado pelo entrevistador sobre frequência do consumo da porção padrão de cada alimento nos últimos 12 meses. O período utilizado como parâmetro teve o objetivo foi estimar o hábito alimentar dos participantes considerando variações no consumo entre as estações do ano.

A reprodutibilidade e a validade relativa do QFA foram avaliadas em um estudo com 281 participantes pelo coeficiente de correlação intraclassa (ICC). Os resultados mostraram confiabilidade satisfatória para todos os nutrientes e validade relativa aceitável para energia, macronutrientes, cálcio, potássio e vitaminas E e C. Os detalhes sobre a elaboração e a validação do questionário podem ser obtidos em publicações anteriores (MOLINA et al., 2013a, 2013b).

Após coleta dos dados, o software *Nutrition Data System for Research* da *University of Minnesota*, Minneapolis, USA, foi utilizado para calcular a composição nutricional dos alimentos/preparações (NCC, 2010). A ingestão

diária de alimentos e bebidas foi determinada pelo número de porções consumidas por dia X peso (porção padrão em gramas) X frequência de consumo X composição nutricional do alimento. Os coeficientes utilizados foram: 3 para mais que 3 vezes/dia, 2.5 para 2 a 3 vezes/dia, 1 para 1 vez/dia, 0.8 para 5 a 6 vezes/semana, 0.4 para 2 a 4 vezes/semana, 0.1 para 1 vez/semana, 0.07 para 1 a 3 vezes/mês e 0 para nunca/quase nunca.

#### **4.6. Análise estatística**

O ambiente R, versão 4.0.2, foi utilizado para a execução da análise exploratória, pré-processamento dos dados e construção dos modelos de análise. As variáveis contínuas foram apresentadas com média e desvio padrão e as variáveis categóricas com frequência e proporção. Para comparação de médias o *Student's t-test* foi utilizado. No caso das variáveis categóricas, comparação entre os grupos foi realizada com o *chi-square test*. O valor de significância considerado foi  $p < 0.05$ . As análises são apresentadas na seção Resultados dos artigos.

#### **4.7. Descrição dos modelos de análise e dos algoritmos utilizados**

Considerando os objetivos do estudo, o modelo de análise foi desenvolvido com a utilização de diferentes técnicas e treinamento de algoritmos baseados em *Machine Learning*, sendo subdividido em modelos:

- ✓ Modelo 1: *k-Means* (classificação) + *Support Vector Machines*, *Decision Trees*, *K-Nearest Neighbours* e *Naive Bayes* (predição).
- ✓ Modelo 2: Sistema de Recomendações (personalização).



Para classificação dos participantes do estudo por padrão de dieta, foi utilizado um algoritmo de aprendizagem não supervisionada (*K-means*). A partir dos dados de consumo coletados com o Questionário de Frequência Alimentar, o algoritmo destacou os padrões existentes na amostra e identificou cada pessoa como pertencente a apenas um agrupamento/cluster. Na sequência, quatro outros algoritmos de aprendizagem supervisionada (*Support Vector Machines*, *Decision Trees*, *K-Nearest Neighbours* e *Naive Bayes*) foram treinados para prever os padrões utilizando parâmetros indiretos de dieta, ou seja, dados demográficos, sociais, econômicos e clínicos. Finalmente, a construção de sistemas de recomendação para personalização do aconselhamento da dieta ocorreu com algoritmos de Filtragem Colaborativa Baseada em Itens e Usuário (IBFC / UBCF). O modelo de análise foi desenhado considerando os objetivos do estudo, características dos dados disponíveis para análise e funcionalidades dos algoritmos conforme descrito em detalhes na sequência.

#### 4.7.1. *K-Means*

Algoritmos de agrupamento ou clusterização fazem parte do conjunto de técnicas de aprendizagem não supervisionada. Permitem dividir a amostra original em grupos de modo que os indivíduos dentro de um mesmo agrupamento apresentam maior homogeneidade entre si e heterogeneidade em relação àqueles inseridos em outro grupo (JUNG; KANG; HEO, 2014). A vantagem deste tipo de análise está na possibilidade de extrair informações, mais especificamente, os padrões ou grupos de indivíduos existentes no conjunto de dados original.

O principal objetivo dos algoritmos de cluster é encontrar padrões e criar um modelo que agrupe os dados automaticamente com base em suas similaridades. As classes ou os subgrupos identificados permitem abordar as características e os preditores a eles associados (XU; WUNSCH, 2010). *K-Means* é um dos algoritmos mais populares, utilizado para descoberta de conhecimento. Esse algoritmo de cluster em particular divide os dados por partição, cada instância é alocada em um agrupamento e faz parte apenas daquele grupo, sem qualquer relação hierárquica entre os *K-clusters* (DEMIDENKO, 2018; JUNG; KANG; HEO, 2014).

Os algoritmos particionais, caso do *K-Means*, usualmente são mais rápidos e menos subjetivos do que os algoritmos aglomerativos, pois os clusters são gerados e avaliados a partir de critérios específicos. O processo consiste em identificar as regras internas, geralmente baseadas em distância, que permitam separar os dados em subgrupos com itens mais homogêneos. Para essa atividade é imprescindível que o pesquisador conheça bem os dados, o processo de modelagem e as especificações técnicas do algoritmo. Há dois grandes desafios em análises com algoritmos de aprendizagem não supervisionada (JUNG; KANG; HEO, 2014). O primeiro, é lidar com impacto do número de cluster *versus* a variabilidade dos dados na amostra. O segundo, é atribuir significado aos agrupamentos gerados pelo algoritmo, pois ele identifica os clusters sem saber previamente como seriam as classificações. Mesmo com acesso as regras que alocam os dados em um grupo ou em outro, como os padrões são detectados automaticamente, não é uma tarefa fácil descrever os grupos e representá-los significativamente, especialmente em bases com muitos atributos, exigindo conhecimento sobre o tema de análise por parte do

pesquisador ou apoio de especialistas. Este algoritmo é guiado pelo princípio de similaridade, assim os elementos são relacionados e inseridos em um mesmo cluster. A definição de similaridade pode variar entre aplicações, entretanto o conceito pode ser representado pela medida de distância. Quanto menor a distância entre um par de elementos, maior é sua similaridade (JUNG; KANG; HEO, 2014). Por meio de cálculos matemáticos e considerando as características dos atributos, o algoritmo encontra várias formas de medir a similaridade e calcular a distância: Euclidiana, Manhattan, Minkovisk e Cosseno para atributos numéricos e Hamming para não numéricos. A qualidade de um processo de agrupamento depende do algoritmo escolhido, da função de distância e da aplicação. Quando a distância intra-cluster é minimizada e inter-cluster maximizada, obtém-se um modelo de alta qualidade (XU; WUNSCH, 2010).

#### 4.7.2. *Support Vector Machines (SVM)*

Algoritmos SVM fazem parte do grupo de algoritmos de aprendizado supervisionado. Dado um conjunto de exemplos de treinamento, é realizada a identificação de cada entrada como pertencente a uma de duas ou mais classes possíveis. Máquinas de Vetores de Suporte traçam uma linha de separação entre os dados que é chamada de hiperplano. O modelo de SVM é uma representação de pontos mapeados de modo que os exemplos de cada categoria são divididos por um espaço claro e bem definido (UPPU; KRISHNA; GOPALAN, 2017).

Os novos pontos são inseridos no mesmo espaço e preditos como pertencentes a uma categoria com base em qual lado do espaço eles sejam

colocados (UPPU; KRISHNA; GOPALAN, 2017). Além da linha de separação entre as classes, construída com regressão, outras duas linhas resultam de pontos de dados identificados na fronteira de cada classe. Essas linhas são chamadas de margens e os pontos de dados, vetores de suporte (ABE, 2010).

O algoritmo SVM permite a construção de excelentes classificadores com dados complexos e com muitas dimensões. São utilizados pelo seu poder preditivo, em alguns casos superiores a outros algoritmos bastante populares, como as Redes Neurais Artificiais, embora com limitada interpretabilidade da saída obtida. A classificação é realizada mesmo quando os dados apresentam valores discrepantes, se necessário o algoritmo filtra os dados atípicos e desconsidera os *outliers* (RICHTER; KHOSHGOFTAAR, 2018).

Baseado na teoria do aprendizado estatístico e em métodos de Kernel, trata-se de um classificador linear, binário, não probabilístico (EHRENTAUT et al., 2018). Apesar disso, também é utilizado em problemas de regressão e com dados não linearmente separáveis. Sua utilidade tem sido explorada em reconhecimento de *spam*, imagens, classificação de texto, caracteres escritos manualmente, textura e expressão de genes. É capaz de extrair os padrões existentes, mesmo quando não há conhecimento prévio do pesquisador ao lidar com dados muito complexos e não óbvios (ABE, 2010; RICHTER; KHOSHGOFTAAR, 2018).

#### 4.7.3. *Decision Trees (DT)*

Algoritmos de árvore de decisão fazem parte do grupo de algoritmos de aprendizado supervisionado. A representação gerada por esse tipo de algoritmo

é de muito fácil visualização e entendimento. Como o próprio nome sugere, a sua ilustração remete a uma árvore. Atributos que identificam a probabilidade de ocorrência de uma classe, dentre as classes possíveis, auxiliam na subdivisão dos dados por meio de nós de decisão. Cada nó pode ter dois ou mais ramos e este vai conduzir a uma folha terminal (HEARTY; GIBNEY, 2008a; KASTORINI et al., 2013).

Árvores de decisão são utilizadas como ferramenta de apoio a decisão que naturalmente podem induzir regras e para cada regra uma decisão precisa ser tomada. A estrutura remete a uma árvore de cabeça para baixo, com o topo representando a raiz e as folhas encontram-se na parte de baixo. Para que uma decisão ocorra, o fluxo começa na raiz, que é o ponto de partida, e os atributos que dividem os dados, representados pelos nós da árvore, determinam o próximo passo do fluxo. Os atributos, também conhecidos como condições de checagem, partem com os ramos da árvore identificando os valores que cada nó pode assumir até chegar a uma decisão, no caso, uma folha terminal (HAND; MANNILA; SMITH, 2001).

Diferentemente de outras técnicas preditivas, os resultados permitem a identificação de um conjunto de regras bem definidas. Assim como as SVM, esse algoritmo é útil tanto em problemas de regressão como de classificação, além de fazer previsões (NAVARRO SILVERA et al., 2014). A poda (*pruning*) é um procedimento utilizado para reestruturar a árvore e evitar o super ajuste (*overfitting*), além de diminuir o seu tamanho sem que haja comprometimento da acurácia. Alguns exemplos de algoritmos de árvores de decisão no ambiente R são o ID3, C5.0, CART e Random Forest (GIABBANELLI; ADAMS, 2016; RICHTER; KHOSHGOFTAAR, 2018).

#### 4.7.4. *K-Nearest Neighbours (K-NN)*

Algoritmos K-NN fazem parte do grupo de algoritmos de aprendizado supervisionado. O algoritmo K-NN foi proposto em 1966, inicialmente para resolver o problema do caixeiro viajante ao traçar a melhor rota entre diversas cidades, pois é capaz de gerar rapidamente o caminho mais curto entre dois pontos. É um algoritmo de fácil compreensão e implementação, sendo a base para métodos mais avançados e pode ser utilizado em etapas iniciais de modelagem de problemas complexos (PARR et al., 2008).

A classificação de uma entrada desconhecida é realizada com base no rótulo dos exemplos de treinamento dos vizinhos mais próximos no espaço de características. A determinação do rótulo do exemplo desconhecido pode ser feita com base no voto majoritário dos  $k$  vizinhos mais próximos. Para sua utilização é necessário um conjunto de dados para treinamento, definir a métrica do cálculo de distância entre os dados e o número de  $K$  vizinhos mais próximos que serão utilizados pelo algoritmo no momento da classificação (HAND; MANNILA; SMITH, 2001; SABERIOON et al., 2018).

A distância euclidiana é a mais utilizada, no entanto a distância Manhattan, Minkowsky e Hamming também se encontram implementadas nos algoritmos. As medidas de distância são muito sensíveis a escala observada nos dados, portanto a padronização dos dados deve ser realizada antes de qualquer treinamento do algoritmo (HAND; MANNILA; SMITH, 2001).

#### 4.7.5. *Naive Bayes (NB)*

Algoritmos Naive Bayes fazem parte do grupo de algoritmos de aprendizado supervisionado. Baseado no teorema de probabilidade de Bayes, o algoritmo tem o objetivo de calcular a probabilidade de uma amostra desconhecida pertencer a uma classe possível, ou seja, prever a classe mais provável. Esta previsão é conhecida como classificação estatística em razão de ser totalmente baseada em probabilidade. Além disso, a classificação também é chamada de ingênua, pois considera que o efeito de um atributo na ocorrência de uma classe é independente da presença ou ausência de qualquer outro atributo (NEAPOLITAN, 2004).

O algoritmo encontra um modelo ou função que descreve diferentes classes de dados, com a finalidade de rotular automaticamente novas instâncias apresentadas com base no treinamento realizado. É um classificador probabilístico simples, útil em grandes bases de dados e pode apresentar resultados superiores a outras técnicas mais sofisticadas. Utilizado para previsões multi-classes, classificação de textos, organização automática de *e-mails*, filtragem de *spam*, identificação de páginas com conteúdo adulto, análise de sentimento, detecção de expressões multi-palavras, previsões em tempo real, sistema de recomendação (HAND; MANNILA; SMITH, 2001).

No Quadro 1 são apresentadas as principais características dos algoritmos de aprendizagem supervisionada.

Quadro 1. Descrição dos algoritmos utilizados na etapa de aprendizagem supervisionada.

<b>Algoritmo</b>	<b>Vantagens</b>	<b>Desvantagens</b>
SVM	Lida com dados não linearmente separáveis, desconsidera outliers	Abstrato, pouco poder de interpretação
Decision Trees	Regras claras, boa interpretabilidade	Overfitting
Knn	Simples, acurado	Sensível a escala nos dados, pode ser lento em grandes conjuntos
Naïve Bayes	Simples, acurado	Assume independência entre atributos

SVM: Support Vector Machines; DT: Decision Trees; Knn: K-Nearest Neighbours; NB: Naïve Bayes.

#### 4.7.6. Collaborative Filtering algorithms

A filtragem colaborativa é uma das técnicas mais utilizadas entre os sistemas de recomendação (SRs), é baseada no conceito de que os indivíduos que têm interesse em um determinado item podem também ter interesse semelhante em outros itens (FU et al., 2019). Em sistemas de recomendação colaborativos, as sugestões são baseadas em classificações dadas por grupos de pessoas. Ele localiza usuários com um histórico de classificação semelhante ao do usuário atual e gera recomendações para este usuário (SASE et al., 2015).

Esses sistemas podem ser baseados em memória, comparando os usuários uns contra os outros, diretamente usando correlação ou outras medidas, ou baseados em modelo, em que um modelo é derivado dos dados de classificação históricos e usado para fazer previsões (THOMAS; SUJATHA, 2016). Por exemplo, ao acessar os perfis dos participantes do estudo, o SR tem



acesso a todas as suas características. Assim, o sistema pode identificar pessoas que compartilham a mesma preferência alimentar e, então, sugerir itens alimentares por semelhança (MUSTAQEEM; ANWAR; MAJID, 2020).

O algoritmo recommender lab representa um modelo de recomendação aprendido para um determinado conjunto de dados (uma matriz de classificação). Os objetos são criados pela função *recommender* (HAHSLER, 2017). Valores *missing* no banco de dados podem ser utilizados para recomendação. Portanto, se uma pessoa possui valores *missing, ratings* (escala de preferência/gostos) são preditos e àqueles com maior valor são recomendados. O sistema prediz qual seria o nível de preferência de uma pessoa por um item, que pode ser uma marca, serviço ou um alimento, com base no seu perfil e histórico e compara com outros indivíduos.

#### **4.8. Treinamento e validação dos modelos**

Para todos os algoritmos citados há algumas etapas que são comuns, como a fatorização de variáveis categóricas e a padronização de escala nas variáveis contínuas. Além disso, como o objetivo dos modelos é a generalização do aprendizado, o banco de dados original (n=12.667) foi dividido em dois subconjuntos com a função “createDataPartition”, assegurando-se aleatoriedade dos indivíduos e proporcionalidade entre ambos na variável dieta [padrão alimentar], objeto de predição nos modelos. O primeiro subconjunto foi utilizado para treino e o segundo para teste, sendo 70% e 30%, respectivamente.

A avaliação dos modelos, incluindo o cálculo da acurácia e de outras métricas de erro, ocorreram no subconjunto de teste. Os valores reais –

correspondente a classificação do padrão de dieta realizada com o *K-Means* e dados sobre o consumo alimentar – foram cruzados com as predições realizadas pelos demais algoritmos a partir dos dados demográficos, socioeconômicos e clínicos. Desta forma, comparação do padrão alimentar extraído com o algoritmo de cluster vs. padrão predito pelo algoritmo supervisionado.

#### **4.9. Limitações do estudo**

Os dados utilizados para avaliação do consumo foram provenientes da coleta realizada com a aplicação do Questionário de Frequência Alimentar. Assim como ocorre com outros métodos de coleta de dados, há possibilidade de o relato ser sub ou superestimado. O viés de memória está presente quando aplicados instrumentos que requerem que o entrevistado faça a descrição da alimentação. Além disso, a generalização dos resultados deste estudo deve ser cuidadosa. Embora a amostra seja constituída por um número grande de participantes e que eles venham de seis localidades diferentes, ela é composta apenas por servidores públicos, que possuem características que não representam o restante da população, especialmente no que se refere a renda e escolaridade mais elevadas.

Embora os algoritmos baseados em *machine learning* implementem técnicas com capacidade para a predição de eventos, devem ser utilizadas como ferramenta complementar aos métodos de análise já estabelecidos. Mesmo com a contribuição da tecnologia para o desenvolvimento de metodologias analíticas alternativas, o elemento humano deve permanecer como aquele que detém

responsabilidade para validar os resultados obtidos com a aplicação de algoritmos baseados em *machine learning* (MEHTA; PANDIT, 2018).

Um outro ponto de reflexão, cujo objeto de regulação da Lei Geral de Proteção de Dados (LGPD) recentemente passou a vigorar no Brasil, recai sobre o conhecimento obtido com dados privados e as intenções de uso. Algoritmos baseados em *machine learning* são capazes de produzir modelos para prever a aceitação de produtos e serviços, informação valiosa a muitas empresas. O problema não está, necessariamente, no poder preditivo do modelo, mas nas circunstâncias que envolvem a sua exploração (PASHAZADEH; NAVIMIPOUR, 2018). Conforme descrito em item específico (item 4.11. Aspectos Éticos), este estudo foi submetido à apreciação pelo Comitê de Ética em Pesquisa (CEP) da Faculdade de Saúde Pública da Universidade de São Paulo. Os riscos que foram identificados e caracterizados como de “nível baixo”, estão relacionados principalmente com vazamento de informações. Entretanto, os dados utilizados na análise e divulgação dos seus resultados não permitem qualquer identificação das pessoas participantes do estudo. Além disso, a predição do padrão alimentar não foi utilizada com finalidade para estigmatizar os participantes em grupos com comportamentos “bons’ ou “ruins” e o ambiente de discussão é estritamente acadêmico.

#### **4.10. Pontos fortes do estudo**

A alimentação é um importante fator de exposição para a saúde e amplamente estudado em estudos epidemiológicos. Para a sua avaliação, demanda o desenvolvimento de técnicas sensíveis, de fácil interpretação, reprodução e baixo custo. Considerado o investimento necessário para

avaliação da alimentação em grandes populações como estudos de coorte e inquéritos populacionais, o desenvolvimento de modelos preditivos para a dieta pode representar avanço nessa área, especialmente por utilizar dados indiretos e de mais fácil mensuração quando comparado ao relato sobre o consumo. Para o nosso conhecimento, este se constitui como um dos primeiros estudos que tem aplicado algoritmos na predição de padrões alimentares, bem como pela construção de um sistema de recomendações alimentares em um país de baixa e média renda como o Brasil.

#### **4.11. Aspectos éticos**

O estudo ELSA-Brasil foi submetido ao Comitê de Ética em Pesquisa (CEP) de cada instituição envolvida, tanto na primeira (Anexo 3) como na segunda e terceira onda. Todos os participantes foram informados sobre os objetivos da pesquisa e ao concordar em participar no estudo assinaram o Termo de Consentimento Livre e Esclarecido (TCLE). Além disso, este trabalho foi submetido ao Comitê de Ética em Pesquisa da Faculdade de Saúde Pública da Universidade de São Paulo, conforme parecer emitido sob o número 2.566.286 (Anexo 4).

#### **4.12. Conflito de interesses**

O autor declara não haver conflito de interesses.

#### **4.13. Fontes de financiamento**

O ELSA-Brasil recebeu financiamento do Ministério da Saúde do Brasil (Departamento de Ciência e Tecnologia) e Ministério de Ciência, Tecnologia e Inovação (FINEP, Financiadora de Estudos e Projetos), processos nº 01 06 0010.00 RS, 01 06 0212.00 BA, 01 06 0300.00 ES, 01 06 0278.00 MG, 01 06 0115.00 SP e 01 06 0071.00 RJ e CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) (SCHMIDT et al., 2015). No entanto, este trabalho não recebeu apoio financeiro. Os órgãos de financiamento do estudo ELSA-Brasil não tiveram qualquer influência sobre a concepção, desenho, análise, interpretação ou redação deste estudo.

## 5. RESULTADOS

Foram incluídos, na sequência, 2 (dois) artigos relativos ao tema da tese e de autoria principal que foram submetidos em revista arbitrada. Ambos os trabalhos foram desenvolvidos e orientados pelos objetivos apresentados inicialmente e se debruçaram sobre a aplicação de algoritmos para apoiar algumas etapas da avaliação do consumo alimentar, além de apresentar comparações entre os algoritmos e técnicas utilizados.

Artigo 1 “Clustering Analysis and Machine Learning algorithms in prediction of dietary patterns: baseline results of the ELSA-Brasil study” tem como objetivo prever padrões alimentares a partir de um conjunto de variáveis demográficas, socioeconômicas e clínicas.

Artigo 2 “Recommender System Based on Collaborative Filtering for Personalized Dietary Advice: A Cross-sectional Analysis of the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil)” tem como objetivo propor recomendações dietéticas personalizadas com base em algoritmos de recomendação.

Os trabalhos foram desenvolvidos em parceria entre pesquisadores vinculados aos departamentos de Epidemiologia e de Nutrição da Faculdade de Saúde Pública da Universidade de São Paulo; do Centro de Pesquisa Clínica e Epidemiológica do Hospital Universitário da USP; da Universidade Federal do Mato Grosso; e da Universidade Federal de Uberlândia. Todos os pesquisadores contribuíram com os desenhos dos estudos, análise dos dados, interpretação dos resultados, redação e revisão da versão final dos manuscritos.

## 5.1. Padrões Alimentares – Artigo 1

### **Clustering Analysis and Machine Learning algorithms in prediction of dietary patterns: baseline results of the ELSA-Brasil study**

Short title: Dietary Patterns and Machine Learning

Vanderlei Carneiro da Silva<sup>1,2</sup>, Bartira Gorgulho<sup>3</sup>, Dirce Maria Marchioni<sup>4</sup>, Tânia Aparecida de Araujo<sup>5</sup>, Itamar de Souza Santos<sup>2</sup>, Paulo Andrade Lotufo<sup>2</sup>, Isabela Martins Benseñor<sup>2</sup>

1. Postgraduate student, Department of Epidemiology, School of Public Health, University of São Paulo, Brazil.
2. Center of Clinical and Epidemiological Research, University Hospital, University of São Paulo, Brazil.
3. Department of Food and Nutrition, School of Nutrition, Federal University of Mato Grosso, Brazil.
4. Department of Nutrition, School of Public Health, University of São Paulo. Brazil.
5. School of Medicine, Federal University of Uberlandia, Brazil.

**Abstract:**

**Introduction:** Machine learning (ML) is particularly understudied in the dietary pattern literature. This study aimed to predict dietary patterns using sociodemographic and clinical data from participants in a large Brazilian cohort study and to compare algorithm performance.

**Methods:** We analyzed data from public employees (n= 12,667) participating in the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil) from 2008-2010. The sample consisted of 5,217 men (41%) and 7,450 women (59%) aged 35-74 years old, who completed a 114-item semiquantitative food frequency questionnaire (FFQ) at baseline. The dietary patterns were derived by the K-means clustering algorithm, and four other classifiers (support vector machine [SVM], naïve Bayes [NB], K-nearest neighbors [KNN] and decision tree [DT] classifiers) were used to predict the patterns of each participant.

**Results:** K-means clustering identified two dietary patterns. Cluster 1, labeled the Western pattern, was characterized by higher intakes of refined cereals, beans, processed and red meats, high-fat milk and dairy products, and sugary beverages; Cluster 2, labeled the Prudent pattern, was characterized by higher intakes of fruit, vegetables, whole cereals, poultry, fish, and milk and reduced-fat milk derivatives. The most important predictors were age, sex, per capita income, education level, retirement status, and physical activity. The accuracy of the models varied from moderate to good: 69% (KNN), 70% (NB), 70% (DT), and 71% (SVM).

**Conclusion:** The algorithms' performance for dietary pattern prediction was similar, and the models presented may provide support in screener tasks using a few



sociodemographic and clinical variables and guide health care professionals in dietary interventions.

**Keywords:** dietary patterns; machine learning; classification algorithms; clustering analysis.

## **Introduction**

Diet has been shown to play a fundamental role in the prevention of chronic diseases, and a healthy lifestyle can help to reduce the risk of diseases (WHO, 2019). Many dietary assessment methods attempt to estimate total food and nutrient intake, but this detailed estimation may lead to the collection of much data, which can be complex and difficult to summarize into dietary patterns (GIABBANELLI; ADAMS, 2016). In addition, dietary intake data are not always available, and the assessment methods require structure and preparation from the interviewers, especially in large samples (NEUHOUSER et al., 2001; SHIM; OH; KIM, 2014).

Machine learning (ML) is a fast-growing discipline that investigates how computers can automatically learn to recognize complex patterns and make intelligent data-based decisions (HAN; KAMBER; PEI, 2011). This area comprises numerous different types of algorithms that can process large amounts of data, such as nutrition information, and ultimately transform data into knowledge (PANARETOS et al., 2018). The approach of dietary patterns allows us to evaluate diets as a whole, instead of food or nutrients alone (MUGA et al., 2016). Cluster analysis is an applied method for characterizing dietary patterns (DEVLIN et al., 2012; NEWBY; TUCKER, 2004). It aggregates individuals with similar dietary patterns into mutually exclusive categories according to the mean of the food intake variables, such as the frequency of food consumed, the percentage of energy contributed by each food, or the average grams of food intake (THORPE et al., 2016).

Sociodemographic and clinical data are basic information always collected in health care and epidemiological surveys. It is widely understood that sociodemographic and health characteristics are associated with dietary patterns and can influence food

choices (BARALDI et al., 2018; KRIEGER et al., 2019). These data could be used as a complementary method for screening and guiding health care professionals when dietary intake records are not available. Few studies have examined the predictive accuracy of alternative ML methodologies to predict dietary patterns based on self-reported dietary intake or even considering sociodemographic and health data (HEARTY; GIBNEY, 2008a; HOFFMANN et al., 2004; PANARETOS et al., 2018; PENCINA et al., 2009).

Dietary assessment can be used to guide public health policies and population interventions (SUBAR et al., 2015). Moreover, when resources and time are limited for data collection and a less detailed assessment meets the research objectives, faster and more objective tools can be useful (PÉREZ RODRIGO et al., 2015). The aim of this study was to predict dietary patterns and to compare the performance of various ML algorithms. We will test the hypothesis that patterns derived by clustering analysis can be predicted from sociodemographic and clinical datasets from public employees participating in the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil).

## **Materials and Methods**

The ELSA-Brasil study was designed to investigate the incidence of cardiovascular diseases, diabetes, and their biological and social determinants. Initially, 15,105 participants aged 35–74 years old from six cities – Belo Horizonte, Porto Alegre, Rio de Janeiro, Salvador, São Paulo, and Vitória – were included. This analysis included cross-sectional data from the baseline examination, which was carried out between August 2008 and December 2010 (AQUINO et al., 2012; BENSENOR et al., 2013). The study was conducted according to the guidelines in the Declaration of Helsinki and approved by the Ethics Committee under number 2.566.286. The participants were

informed about the research objectives, and written consent was obtained from all the participants.

### Dietary Assessment

A semiquantitative food frequency questionnaire (FFQ) developed for the ELSA-Brasil study was used. Subjects were asked to estimate how often, on average, they consumed a standardized portion of a given food item, in the 12-month period preceding the interview. The reproducibility and relative validity of the FFQ was evaluated in a study with 281 participants by the intraclass correlation coefficient (ICC). The FFQ showed satisfactory reliability for all nutrients and reasonable validity, especially for energy, macronutrients, calcium, potassium, and vitamins E and C. Details about the elaboration (MOLINA et al., 2013c) and validation of the questionnaire (MOLINA et al., 2013a) can be obtained in previous publications.

The daily intakes of foods and drinks were quantified by the number of servings consumed per day x the weight (standard portion in grams) x the frequency of consumption x the nutritional composition of the food serving. The daily equivalent coefficients used were 3 for more than 3 times/day, 2.5 for 2 to 3 times/day, 1 for once/day, 0.8 for 5 to 6 times/week, 0.4 for 2 to 4 times/week, 0.1 for once/week, 0.07 for 1 to 3 times/month, and 0 for never/almost never (MOLINA et al., 2013c).

The Nutrition Data System for Research (NDSR) software (University of Minnesota, Minneapolis, USA, 2010) was used to determine the nutritional composition of foods and preparations and daily energy intake in kilocalories (NCC, 2010). Of the total sample, we excluded n=2,438 (16%) participants with implausible energy intake of less than 500 or greater than 4,000 Kcal/day (WILLET, 1998b). The final sample was

composed of 12,667 public employees, of which 5,217 (41%) were men and 7,450 (59%) were women.

### Sociodemographic and Clinical Predictors

The ELSA-Brasil study included clinical tests and interviews in its protocol, which required volunteers to visit a clinical research center. This protocol ensured a high standard of quality in data collection. Each participant was previously assigned to a specific sequence of clinical tests and interviews (BENSENOR et al., 2013).

The following sociodemographic and clinical data were collected and included in the analysis: sex (male vs. female), age (years), education level (elementary [or less], high school, college), retirement status (no vs. yes), self-reported race/ethnicity (white, brown, black, other [Asian, indigenous]), per capita income in US\$ (categorized in terciles [using US\$ 1.00 = R\$ 2.00 as the approximate baseline examination exchange rate]), living alone (no vs. yes), marital status (not married vs. married), smoking habit (never, ex-smoker, current smoker), physical activity (sedentary, insufficiently active, active [using the leisure time section of the long version of the International Physical Activity Questionnaire]), health self-assessment (good, regular, bad), and local clinical research center.

Weight and height measurements were obtained with the participant wearing light clothes and no shoes. We measured body weight to the nearest 0.1 kg with a calibrated scale (Toledo 2096PP) and height with a vertical stadiometer (Seca-SE-216) to the nearest 0.1 cm. Body mass index (BMI) was calculated by dividing weight in kilograms by height in meters squared ( $\text{kg}/\text{m}^2$ ). Waist circumference was measured with a tape measure to the nearest 0.1 cm around the midpoint between the inferior costal border and the iliac crest,

while hip circumference was measured at the point of greatest circumference in the gluteal region. The waist-to-hip ratio (WHR) was calculated by dividing the waist circumference by the hip circumference in centimeters.

Blood pressure (BP) was measured using a validated Omron HEM 705CPINT oscillometer device. Three measurements were taken at one-minute intervals, and the mean of the two latter BP measurements was considered the value for defining hypertension, defined as systolic BP  $\geq 140$  mm Hg, diastolic BP  $\geq 90$  mm Hg or verified treatment with antihypertensive drugs during the previous two weeks. Dyslipidemia was defined as low-density lipoprotein (LDL) cholesterol  $\geq 130$  mg/dL or the use of medication to treat dyslipidemia. Diabetes was defined as a previous diagnosis of diabetes, use of medication to treat diabetes, fasting plasma glucose  $\geq 126$  mg/dL, 2-h plasma glucose  $\geq 200$  mg/dL, or HbA<sub>1C</sub>  $\geq 6.5\%$ . Cardiovascular disease was defined as self-reported prior myocardial infarction, stroke or revascularization.

### Statistical Analysis

The continuous variables are presented as means and standard deviations (SDs), and the categorical variables are presented as frequencies. The continuous variables were standardized, and adjustments were made to keep the data in the same format (scale function). The categorical variables were treated as factors. All analyses were performed in R, version 4.0.2. The data preprocessing steps were performed with the support of the base functions and the dplyr package. Associations between categorical variables were tested through chi-square tests. Comparisons of the mean values of continuous variables by dietary pattern (i.e., Western or Prudent) were performed using Student's t-tests.

## Machine Learning Algorithms

To implement the models, the open-source library of the R environment for ML (caret, version 4.0.2) was employed. The K-means clustering algorithm was used to identify dietary patterns. Next, four different algorithms (support vector machine [SVM], naïve Bayes [NB], K-nearest neighbors [KNN], and decision tree [DT] classifiers) were used to predict the dietary patterns of each participant. We calculated the agreement between the patterns predicted by the supervised techniques using sociodemographic data and the classification scheme provided by the unsupervised algorithm (COHEN, 1960; LANDIS; GG, 1977). In general, values equal to 0.5 correspond to the performance of a random classifier, values less than 0.6 (and greater than 0.5) indicate to moderate predictive performance, whereas values greater than 0.7 indicate good predictive performance (HANLEY; MCNEIL, 1982; KASTORINI et al., 2013; LANDIS; GG, 1977; PENCINA et al., 2009). Furthermore, other evaluation metrics, such as the sensitivity, specificity, positive and negative predictive value, were also calculated. In the following, the functionality of each ML algorithm is briefly described.

### Clustering Analysis (K-Means)

The K-means clustering algorithm was used to divide the participants into groups based on their dietary intake data. K-means clustering is one of the most popular algorithms (HEARTY; GIBNEY, 2008b; VILLEGAS et al., 2004; WINKVIST et al., 2009) and is used for knowledge discovery. This method of clustering partitions the data; that is, each instance is placed in a cluster and is only part of that group, and there is no hierarchical relationship between the K clusters (NEWBY; TUCKER, 2004; REDDY; VINZAMURI, 2014). The frequencies of food intake were converted to z-scores and

input into the algorithm. We retained two dietary patterns considering homogeneity in the derived groups, the balance between classes and the range of clusters found previously in the literature (DEVLIN et al., 2012; NEWBY; TUCKER, 2004). The factor interpretability was examined to confirm the final number of clusters and if a group was sufficiently large for adequate statistical power, that is, at least 10% of the total sample (THORPE et al., 2016). The class imbalance problem is closely related to cost-sensitive learning (HAN; KAMBER; PEI, 2011).

Next, the original database (n=12,667) was randomly divided into two subsets. The first subset was used for training, and the second subset was used for testing, with sample sizes of 8,866 (70%) and 3,801 (30%), respectively. The training stage of the classifier algorithms was based on sociodemographic and clinical data: sex, age, education level, retirement status, race/ethnicity, marital status, per capita income, living alone, local clinical research center, smoking habit, physical activity, health self-assessment, BMI, WHR, dyslipidemia, hypertension, diabetes and cardiovascular disease.

#### Support Vector Machine (SVM)

SVM classifiers operate by separating classes using linear decision boundaries called hyperplanes (SIDEY-GIBBONS; SIDEY-GIBBONS, 2019). The model is a representation of mapped points so that the examples of each category are divided by a clear and well-defined space. Given a set of training examples, the algorithm identifies each input as belonging to one of two possible classes (ABE, 2010). The SVM classifier is based on the theory of statistical learning and kernel methods and is a linear, binary, no



probabilistic classifier. Despite this characteristic, it is also used in regression problems and in nonlinearly separable data (OLSON; DELEN, 2008).

#### Naïve Bayes (NB)

The NB algorithm is extremely simple but powerful. NB is based on Bayes' theorem and aims to calculate the probability that an unknown sample belongs to a certain class; that is, it predicts the most likely class (HAND; MANNILA; SMITH, 2001). In addition, the classification is also called naïve, as it considers that the effect of an attribute on the occurrence of a class is independent of the presence or absence of any other attribute. This algorithm is useful in large databases and can yield results superior to other more sophisticated techniques (HAN; KAMBER; PEI, 2011).

#### K-Nearest Neighbors (KNN)

The KNN algorithm predicts an unknown entry based on the label of the training examples of the closest neighbors in the characteristics space. The determination of the label of an unknown example is usually based on the majority vote of the K closest neighbors. For its use, a training dataset is necessary to define the distance calculation metric between the data and the number of K closest neighbors that will be used by the algorithm at the time of classification (HAND; MANNILA; SMITH, 2001).

#### Decision Tree (DT)

The DT algorithm has a flowchart-like tree structure, where each internal node (nonleaf node) denotes a test on an attribute, each branch represents an outcome of the

test, and each leaf node (or terminal node) holds a class label (HAN; KAMBER; PEI, 2011). The structure refers to an upside-down tree, with the roots at the top and the leaves at the bottom. The DT algorithm is used as a decision support tool that can naturally induce rules, and for each rule, a decision needs to be made. Unlike other predictive algorithms, the results allow the identification of a set of well-defined rules. The construction of a DT is appropriate for exploratory knowledge discovery, and the algorithm can handle multidimensional data (WITTEN; FRANK; HALL, 2011). In general, DT classifiers achieve good accuracy.

## Results

From the original dataset ( $n= 12.667$ ), two major dietary patterns were derived: one pattern with 7,157 (57%) participants and the other pattern with 5,510 participants (43%). Subdivisions with three or more clusters did not substantially improve intracluster homogeneity. Figure 1 shows the retained dietary patterns. The smallest clustering pattern, in black (the “Prudent” diet), includes fewer processed foods, while the largest clustering pattern, in gray (the “Western” diet) is characterized by higher intakes of high energy density foods and fats.

Table 1 shows the characteristics of the study sample by dietary pattern. The Prudent dietary pattern group had a higher proportion of women, older people ( $54 \pm 9.2$  years old vs.  $51 \pm 8.7$  years old), people with a high level of education, white people, single people, people living alone, people with higher income, retired people, nonsmokers, and physically active people than the Western dietary pattern group ( $p < 0.001$ ). Among the health issues, the group following the Prudent dietary pattern

contained a higher proportion of people with chronic diseases (dyslipidemia, diabetes, cardiovascular) disease and health rated as good.

Table 2 shows the dietary intakes by pattern. The participants in cluster 1 (n= 5,510, 43% of the total sample) presented a higher consumption of fruit, vegetables, whole cereals, poultry, fish, lower-fat dairy products and milk (Prudent pattern). The subjects that were in cluster 2 (n= 7,157, comprising 57% of the population) presented a higher mean consumption of refined cereals, beans, processed and red meats, eggs, high-fat dairy products and milk, salted snacks, and sugary beverages (Western pattern).

Table 3 shows the performance of the models. Using only sociodemographic and clinical data to predict the dietary pattern in the test set, the accuracy results were 0.71, 0.70, 0.69 and 0.70 for the SVM, NB, KNN and DT classifiers, respectively. Furthermore, the sensitivity, specificity, positive and negative predictive value are presented; each metric measures the classification ability related to one of the two clusters.

Table 4 shows the predictors ranked for each algorithm. All the variables were used in the initial training stage; however, in the table, the variables are presented in descending order by level of importance. The most common features selected are sex, age, education level, per capita income, and physical activity. On the other hand, BMI, diabetes, hypertension, cardiovascular disease, and clinical research center were the least important.

Supplementary Table 1 shows the confusion matrix used to verify the hits and errors of the model. This method is the most common statistical approach in previous studies. The hits are on the main diagonal, that is, when the pattern predicted by the algorithm corresponded with the true pattern (derived from the diet).

## Discussion

Two major dietary patterns were identified in our sample, and ML algorithms were trained to predict them using the only sociodemographic and clinical characteristics of the sample data. The Prudent pattern was characterized by higher intakes of whole cereals, fruit, vegetables, poultry, fish, milk and reduced-fat milk derivatives. The Western pattern was characterized by higher intakes of refined cereals, beans, processed and red meats, eggs, high-fat dairy products and milk, salted snacks, and sugary beverages. Confirming our hypothesis, after a training stage, all the algorithms were able to classify individuals into a dietary pattern based on sociodemographic and clinical characteristics with moderate-to-good accuracy (69% to 71%).

Comparisons of our findings with those of previous studies are limited by the absence of studies that have predicted dietary patterns using only sociodemographic and clinical data. Although using a different method, Pencina et al. evaluated the performance of Fisher's discriminant functions for identifying the dietary patterns of women (n= 1,828) and men (n= 1,666), aged 18-76 years old, who were participants of the Framingham Nutrition Studies (PENCINA et al., 2009). The model classified approximately 80% of participants correctly. The precision was obtained in new individuals whose dietary pattern label was not known by the algorithm, only the information about food groups. Our results were less accurate, but the predictions were made without any information about the diet, bringing new possibilities for the prediction of diet using sociodemographic and clinical health data.

Among the potential predictors, sex, age, education level, per capita income, and physical activity were the most important features in the final models. Associations between dietary patterns and socioeconomic and demographic characteristics have been

observed earlier in low- and middle-income countries (ARRUDA et al., 2014; MAYÉN et al., 2014, 2016) and in high-income countries. Thorpe et al. aimed to compare dietary patterns derived by principal component analysis (PCA) and cluster analysis (CA) and to examine their associations with sociodemographic and health behaviors. They showed that PCA and CA identified comparable dietary patterns for older Australians (n=1,888 men and n=2,071 women). Furthermore, those with dietary patterns characterized by a higher consumption of red and processed meat and refined grains were more likely to be Australian-born, have a lower level of education, have a higher BMI, smoke and not meet physical activity recommendations. Sánchez-Villegas showed that more physically active subjects were less likely to follow a “Western” dietary pattern and more likely to follow a “Spanish Mediterranean” dietary pattern (SÁNCHEZ-VILLEGAS et al., 2003). Guyot et al., in a sample of 5,194 women and men aged 45–60 years old living in France, found that a prudent diet was directly associated with a high level of education and being older than 55 years old and negatively correlated with current smoking status (KESSE-GUYOT et al., 2009).

Our results presented in Table 1 confirm that there are differences between the two identified patterns. The algorithms used can classify individuals based on features present. However, although people in treatment for diet-related chronic diseases are more likely to be exposed to dietary advice and health messages, health and clinical data such as health self-assessment, BMI, dyslipidemia, hypertension, and cardiovascular disease had no significant influence on the performance of the algorithms. In this analysis, sociodemographic data were more related to the dietary pattern than the profile of comorbidities in the sample.

The dietary patterns were identified considering the balancing and homogeneity of the two major clusters (Prudent [43%] and Western [57%]). The participants were

grouped in such a way that the algorithms could capture their main differences. Cardoso et al., using PCA, also analyzed data at the baseline ELSA-Brasil study and found 4 groups: “traditional” (48%), “fruits and vegetables” (25%), “pastry shop” (24%), “diet/light” (5%) (CARDOSO et al., 2016). Our current study used a different method to identify the patterns, which explains the differences. Furthermore, other subdivisions could result in small groups, such as the one presented by Cardoso et al; the “diet/light pattern” represented only 5% of the sample. As mentioned before, class balance is an important determinant for ML algorithm performance (DEVLIN et al., 2012), and this aspect was taken into account in the present study.

Four algorithms were explored, and the results were similar. This finding demonstrates that all the classifiers used can be helpful in predicting dietary patterns. Panaretos et al. used two techniques (the KNN and random forest algorithms) to evaluate participants’ health based on the dietary information of 3,042 men and women ( $45 \pm 14$  years old) who were enrolled in the ATTICA study (PANARETOS et al., 2018). In that study, the ML techniques were superior to linear regression in the correct classification of the individuals according to the health score (accuracy of approximately 38% vs. 6%, respectively), whereas the two ML methods showed equal classification ability.

Our study fills an important gap in the literature: a need for innovative approaches, especially advanced analytical research methods that could enable researchers to predict dietary patterns based on indirect data. The algorithms used can classify individuals into subgroups by predicting the most common dietary patterns and to present features associated with specific patterns in the population. Furthermore, examining single items makes the estimation of the effect size inaccurate, mainly due to the multicollinearity effect because of the strong correlations between some foods and nutrients that are

consumed together (PANARETOS et al., 2017). In some cases, alternative approaches such as the one used in this study can be useful and can overcome these weaknesses.

Although the use of well-established instruments such as R24h and FFQs to collect dietary data can provide many details, which allows the assessment of nutrients and food items and the overall diet structure (SUBAR et al., 2015), these instruments are not always available. ML and other tools supported by technology do not replace existing dietary assessment methods but can be used as complementary approaches (BOLAND; BRONLUND, 2019; ILLNER et al., 2012). Like all other methods, they have limitations: they rely on expert technology teams to deploy the models, the presence of computational structures in health systems and the training of health care professionals on how to correctly interpret the results of the models. Therefore, the use of combined tools is recommended (SHIM; OH; KIM, 2014).

ELSA-Brasil is not a population-based study. However, there are several similarities in the prevalence of behavioral risk factors and chronic conditions evaluated by our study and those evaluated by Vigitel, a telephone-based behavioral risk factor survey that collected representative data for adults living in 27 Brazilian state capitals and the Federal District (MALTA et al., 2015; SOUZA et al., 2011). Vigitel has revealed an increase in the intake of processed foods, sugary beverages, high energy density foods, and food rich in sodium, sugar and fats, although some regions still maintain the consumption of fresh and traditional foods such as rice and beans (MS, 2020). In our study, foods that are markers of a Western diet represented 57% of the analyzed sample. In this way, some generalizability of these results to the general population living in metropolitan areas of the country is possible.

This study has some strengths. The data analyzed were from a large and multicenter sample of adult and elderly individuals. Although the sample consisted of only civil servants, it aggregated an admixed, multiethnic population recruited from six major centers and captured no isolated eating practices. The prediction of dietary patterns based on demographic, social and health factors, especially in low- and middle-income countries, such as Brazil, opens the door for interventions based not only on the components of food consumption itself but also considering the determinants in the population related to diet. The FFQ used for data collection was developed and validated in the study population. Participants were invited to attend a clinical research center for exams and clinical evaluations, which guaranteed a high standard of quality control in the predictors used in the study.

Some limitations also should be addressed. This study is a cross-sectional analysis, and although the features used can predict dietary patterns, we do not assess causality. The collected dietary data were self-reported and are subject to the interviewees' memory biases. However, the FFQ allows the collection of the participants' usual consumption regardless of intraindividual variability and the ranking of people into consumption ranges. Several subjective decisions were made in the process of analysis, such as the definition and collapsing of some food groups, the retention of the number of clusters and the labeling of the identified dietary patterns.

Our results showed that sociodemographic and clinical data can predict dietary patterns with moderate-to-good accuracy and consistency. Sociodemographic data are basic information always collected in health services and epidemiological surveys. In addition, when dietary data are not available or a complete population assessment is not possible due to lack of time or resources or logistical difficulties, the proposed algorithms



can be used as a screening approach. ML methods can be a valuable tool in the field of nutritional epidemiology.

### **Funding**

The ELSA-Brasil study received funding from the Ministry of Health of Brazil (Department of Science and Technology), Ministry of Science, Technology and Innovation (FINEP, Funding of Studies and Projects) and CNPq (National Development Council Scientific and Technological Advice) grants 01 06 0010.00 RS; 01 06 0212.00 BA; 01 06 0300.00 ES; 01 06 0278.00 MG; 01 06 0115.00 SP and 01 06 0071.00 RJ. However, this subproject did not receive funding. Funding received at the study baseline had no influence on the design, analysis, drafting, interpretation or decision of the version submitted for publication.

### **Author contributions**

V.C.S. and I.M.B. contributed to the design, acquisition, analysis, interpretation of data and the drafting and revision of the text; B.G., D.M.M. and T.A.A. contributed to the design, analysis, and revision of the text; I.S.S. and P.A.L. contributed to the acquisition, analysis, and revision of the text. All authors give final approval of the version to be published.

### **Conflict of interests**

The authors declare no conflicts of interest.

### **Data availability**

The data underlying this article cannot be shared publicly due to privacy of individuals that participated in the study and research data is confidential.

## References

ABE, S. **Support Vector Machines for pattern classification**. 1° ed. New York: York, London Springer New, 2010.

AQUINO, E. M. L. et al. Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): Objectives and Design. **American Journal of Epidemiology**, v. 175, n. 4, p. 315–324, 15 fev. 2012.

ARRUDA, S. P. M. et al. Socioeconomic and demographic factors are associated with dietary patterns in a cohort of young Brazilian adults. **BMC Public Health**, v. 14, n. 654, p. 1–13, 2014.

BARALDI, L. G. et al. Consumption of ultra-processed foods and associated sociodemographic factors in the USA between 2007 and 2012: evidence from a nationally representative cross-sectional study. **BMJ open**, v. 8, n. 3, p. e020574, 2018.

BENSENOR, I. M. et al. Routines of organization of clinical tests and interviews in the ELSA-Brasil investigation center. **Rev Saúde Pública**, v. 47, n. suppl 2, p. 37–47, jun. 2013.

BOLAND, M.; BRONLUND, J. eNutrition - The next dimension for eHealth? **Trends in Food Science & Technology**, v. 91, n. April, p. 634–639, 2019.

CARDOSO, L. DE O. et al. Eating patterns in the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): an exploratory analysis. **Cadernos de saude publica**, v. 32, n. 5, p. 1–14, jan. 2016.

COHEN, J. A Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, 2 abr. 1960.

DEVLIN, U. M. et al. The use of cluster analysis to derive dietary patterns:

methodological considerations, reproducibility, validity and the effect of energy mis-reporting. **Proc Nutr Soc.**, v. 71, n. 4, p. 599–609, 2012.

GIABBANELLI, P. J.; ADAMS, J. Identifying small groups of foods that can predict achievement of key dietary recommendations: data mining of the UK National Diet and Nutrition Survey, 2008-12. **Public health nutrition**, v. 19, n. 9, p. 1543–51, jun. 2016.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: Concepts and Techniques**. 3rd. ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

HAND, D. J.; MANNILA, H.; SMITH, P. **Principles of data mining**. 1° ed. Cambridge, Massachusetts: Press, Cambridge Mass MIT, 2001.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, v. 143, n. 1, p. 29–36, abr. 1982.

HEARTY, A. P.; GIBNEY, M. J. Analysis of meal patterns with the use of supervised data mining techniques--artificial neural networks and decision trees. **The American journal of clinical nutrition**, v. 88, n. 6, p. 1632–1642, 2008a.

HEARTY, Á. P.; GIBNEY, M. J. Comparison of cluster and principal component analysis techniques to derive dietary patterns in Irish adults. **British Journal of Nutrition**, v. 101, n. 4, p. 598–608, 25 jun. 2008b.

HOFFMANN, K. et al. Application of a New Statistical Method to Derive Dietary Patterns in Nutritional Epidemiology. **American Journal of Epidemiology**, v. 159, n. 10, p. 935–944, 15 maio 2004.

ILLNER, A.-K. et al. Review and evaluation of innovative technologies for measuring diet in nutritional epidemiology. **International Journal of Epidemiology**, v. 41, n. 4, p. 1187–1203, ago. 2012.

KASTORINI, C. M. et al. Comparative analysis of a-priori and a-posteriori dietary patterns using state-of-the-art classification algorithms: A case/case-control study. **Artificial Intelligence in Medicine**, v. 59, n. 3, p. 175–183, 2013.

KESSE-GUYOT, E. et al. Dietary patterns and their sociodemographic and behavioural correlates in French middle-aged adults from the SU.VI.MAX cohort. **European Journal of Clinical Nutrition**, v. 63, n. 4, p. 521–528, 23 abr. 2009.

KRIEGER, J. P. et al. Dietary patterns and their sociodemographic and lifestyle determinants in switzerland: Results from the national nutrition survey menuCH. **Nutrients**, v. 11, n. 1, p. 1–16, 2019.

LANDIS, J.; GG, K. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159–174, 1977.

MALTA, D. C. et al. Fatores de risco e proteção para doenças crônicas por inquérito telefônico nas capitais brasileiras, Vigitel 2014. **Revista Brasileira de Epidemiologia**, v. 18, n. suppl 2, p. 238–255, dez. 2015.

MAYÉN, A.-L. et al. Socioeconomic determinants of dietary patterns in low- and middle-income countries: a systematic review. **Am J Clin Nutr**, v. 100, p. 1520–31, 2014.

MAYÉN, A.-L. et al. Socioeconomic Differences in Dietary Patterns in an East African Country: Evidence from the Republic of Seychelles. **PLOS ONE**, v. 11, n. 5, p. e0155617, 23 maio 2016.

MOLINA, M. D. C. B. et al. Reprodutibilidade e validade relativa do Questionário de Frequência Alimentar do ELSA-Brasil. **Caderno de Saúde Pública**, v. 29, n. 2, p. 379–389, 2013a.

MOLINA, M. DEL C. B. et al. Diet assessment in the Brazilian Longitudinal Study of

Adult Health (ELSA-Brasil): Development of a food frequency questionnaire. **Rev. Nutr.**, v. 26, n. 2, p. 167–176, abr. 2013b.

**MS. Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise em Saúde e Vigilância de Doenças Não Transmissíveis. Vigitel Brasil 2019: vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico.** 1º ed. Brasília, DF: Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise em Saúde e Vigilância de Doenças Não Transmissíveis., 2020.

MUGA, M. A. et al. Association between Dietary Patterns and Cardiovascular Risk Factors among Middle-Aged and Elderly Adults in Taiwan: A Population-Based Study from 2003 to 2012. **PLOS ONE**, v. 11, n. 7, p. 1–18, 1 jul. 2016.

**NCC. National Cancer Center. Nutrition data system for research software.**Minneapolis, 2010.

NEUHOUSER, M. L. et al. A brief dietary assessment instrument for assessing target foods, nutrients and eating patterns. **Public Health Nutrition**, v. 4, n. 1, p. 73–78, 2 fev. 2001.

NEWBY, P. K.; TUCKER, K. L. Empirically Derived Eating Patterns Using Factor or Cluster Analysis: A Review. **Nutr Rev.**, v. 62, n. 5, p. 177–203, 2004.

OLSON, D. L.; DELEN, D. **Advanced Data Mining Techniques.** Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

PANARETOS, D. et al. **Repeatability of dietary patterns extracted through multivariate statistical methods: a literature review in methodological issues**International Journal of Food Sciences and Nutrition, 2017.

PANARETOS, D. et al. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002-2012): The ATTICA study. **British Journal of Nutrition**, v. 120, n. 3, p. 326–334, 2018.

PENCINA, M. J. et al. Performance of a Method for identifying the Unique Dietary Patterns of Adult Women and Men: The Framingham Nutrition Studies. **J Am Diet Assoc**, v. 108, n. 9, p. 1453–1460, 2009.

PÉREZ RODRIGO, C. et al. Screeners and brief assessment methods. **Nutricion hospitalaria**, v. 31 Suppl 3, p. 91–8, 26 fev. 2015.

REDDY, C. K.; VINZAMURI, B. A Survey of Partitional and Hierarchical Clustering Algorithms. In: AGGARWAL, C. C.; REDDY, C. K. (Eds.). . **Data Clustering. Algorithms and Applications**. 1° ed ed. [s.l.] CRC Press Taylor & Francis Group, 2014. p. 8–106.

SÁNCHEZ-VILLEGAS, A. et al. Gender, age, socio-demographic and lifestyle factors associated with major dietary patterns in the Spanish Project SUN (Seguimiento Universidad de Navarra). **European Journal of Clinical Nutrition**, v. 57, n. 2, p. 285–292, 6 fev. 2003.

SHIM, J.-S.; OH, K.; KIM, H. C. Dietary assessment methods in epidemiologic studies. **Epidemiology and health**, v. 36, p. e2014009, 2014.

SIDEY-GIBBONS, J. A. M.; SIDEY-GIBBONS, C. J. Machine learning in medicine: a practical introduction. **BMC medical research methodology**, v. 19, n. 1, p. 64, 19 mar. 2019.

SOUZA, A. D. M. et al. Avaliação dos marcadores de consumo alimentar do VIGITEL

(2007-2009). **Revista Brasileira de Epidemiologia**, v. 14, n. 1, p. 44–52, 2011.

SUBAR, A. F. et al. Addressing Current Criticism Regarding the Value of Self-Report Dietary Data. **The Journal of Nutrition**, v. 145, n. 12, p. 2639–2645, 1 dez. 2015.

THORPE, M. G. et al. A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians. **International Journal of Behavioral Nutrition and Physical Activity**, v. 13, n. 1, 2016.

VILLEGAS, R. et al. Dietary patterns in middle-aged Irish men and women defined by cluster analysis. **Public Health Nutrition**, v. 7, n. 8, p. 1017–1024, 2 dez. 2004.

WHO. **Sustainable healthy diets. Guiding Principles**. Rome: FAO and WHO, 2019.

WILLET, W. Correction for the effects of measurement error. In: WILLET, W. (Ed.). . **Nutritional Epidemiology**. 2nd. ed. New York, NY, USA: Oxford University Press, 1998. p. 74–147.

WINKVIST, A. et al. More distinct food intake patterns among women than men in northern Sweden: a population-based survey. **Nutrition Journal**, v. 8, n. 1, p. 12, 19 dez. 2009.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining practical machine learning tools and techniques**. 3° ed. United States: Elsevier, 2011.



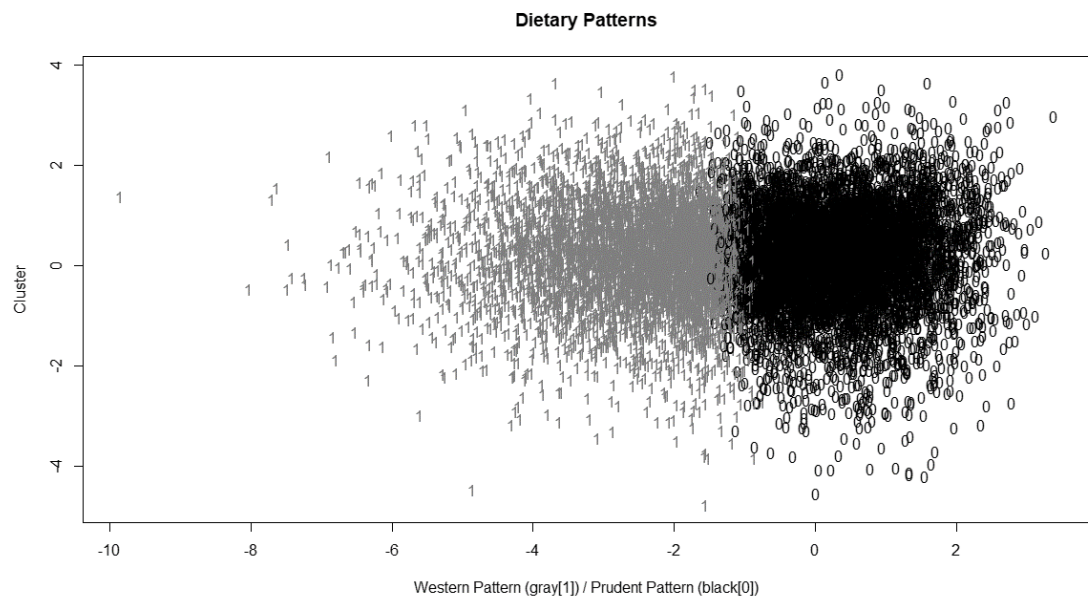


Figure 1. Groupings identified by the K-means clustering algorithm.

Table 1. Characteristics of the study population, Longitudinal Study of Adult Health (ELSA-Brasil), 2008-2010.

Variable	General		Western		Prudent	
	n	%	n	%	n	%
Study population	12,667	100.0	7,157	100.0	5,510	100.0
Data set						
Train	8,866	70.0	5,042	70.5	3,824	69.4
Test	3,801	30.0	2,115	29.5	1,686	30.6
Sex						
Male	5,217	41.2	3,543	49.5**	1,674	30.4
Female	7,450	58.8	3,614	50.5	3,836	69.6
Age (years) <sup>1</sup>	52	±9.1	51	±8.7**	54	±9.2
Education level						
Elementary (or less)	1,423	11.2	1,128	15.8**	295	5.4
High school	4,072	32.2	2,876	40.2	1,196	21.7
College	7,172	56.6	3,153	44.0	4,019	72.9
Retirement status						
No	10,046	79.3	6,064	84.7**	3,982	72.3
Yes	2,621	20.7	1,093	15.3	1,528	27.7
Race/ethnicity						
White	6,994	55.2	3,424	47.8**	3,570	64.8
Mixed	3,379	26.7	2,240	31.3	1,139	20.7
Black	1,831	14.4	1,245	17.4	586	10.6
Others <sup>2</sup>	463	3.7	248	3.5	215	3.9
Marital status						
Single	4,486	35.4	2,286	31.9**	2,200	39.9
Married	8,181	64.6	4,871	68.1	3,310	60.1
Per capita income <sup>3</sup>						
1° tercile	4,225	33.4	3,281	45.8**	944	17.1
2° tercile	4,492	35.5	2,441	34.1	2,051	37.2
3° tercile	3,950	31.1	1,435	20.1	2,515	45.7
Living alone						
No	11,043	87.2	6,471	90.4**	4,572	83.0
Yes	1,624	12.8	686	9.6	938	17.0
Smoking habit						
Never	7,306	57.7	3,946	55.1**	3,360	61.0
Ex-smoker	3,780	29.8	2,049	28.6	1,731	31.4
Current smoker	1,581	12.5	1,162	16.3	419	7.6
Physical activity <sup>4</sup>						
Sedentary	5,798	45.8	3,955	55.3**	1,843	33.5
Insufficiently active	3,354	26.5	1,776	24.8	1,578	28.6
Active	3,515	27.7	1,426	19.9	2,089	37.9

"Continue"

Table 1. Characteristics of the study population, Longitudinal Study of Adult Health (ELSA-Brasil), 2008-2010.

*“Continued”*

Variable	General		Western		Prudent	
	n	%	n	%	n	%
Health self-assessment						
Good	10,266	81.1	5,650	78.9**	4,616	83.8
Regular	2,167	17.1	1,367	19.1	800	14.5
Bad	234	1.8	140	2.0	94	1.7
BMI (kg/m <sup>2</sup> ) <sup>1</sup>	26.9	±4.7	27.0	±4.8	26.9	±4.6
Waist-to-hip ratio <sup>1</sup>	0.9	±0.1	0.9	±0.1**	0.8	±0.1
Dyslipidemia <sup>5</sup>						
No	5,237	41.3	3,206	44.8**	2,031	36.9
Yes	7,430	58.7	3,951	55.2	3,479	63.1
Hypertension <sup>6</sup>						
No	8,159	64.4	4,584	64.1	3,575	64.9
Yes	4,508	35.6	2,573	35.9	1,935	35.1
Diabetes <sup>7</sup>						
No	10,634	83.9	6,082	85.0**	4,552	82.6
Yes	2,033	16.1	1,075	15.0	958	17.4
Cardiovascular disease <sup>8</sup>						
No	12,188	96.2	6,922	96.7*	5,266	95.6
Yes	479	3.8	235	3.3	244	4.4

1. Mean, standard deviation.

2. Others = Asian + indigenous.

3. Calculation based on 2009: 1 US\$ = R\$ 2,00.

4. Sedentary: does not do physical activity; insufficiently active: &lt;150 minutes/week or exercise &lt;than 3 days a week; active: 150 minutes/week at least 3 days a week.

5. LDL ≥130 mg/dL or use of cholesterol reducers.

6. Systolic blood pressure ≥140 mmHg, diastolic blood pressure ≥90 mmHg or verified treatment with antihypertensive drugs during the previous two weeks.

7. Defined as an account of a previous diagnosis of diabetes or the use of medication for diabetes or to fulfill a diagnostic value of diabetes.

8. Defined as a report of a heart attack, stroke or revascularization.

p-values are derived from Student's t-tests or chi-square tests.

\*p &lt; 0.05, \*\*p &lt; 0.001

Table 2. Food consumption by dietary pattern, Brazilian Longitudinal Study of Adult Health (ELSA-Brasil), 2008-2010.

Food groups (g or mL/day)	Western Pattern		Prudent Pattern	
	Mean	±SD	Mean	±SD
Energy <sup>1</sup>	2,606.4	700.2**	2,503.5	678.4
Refined cereals	197.7	148.1**	81.5	96.3
Whole cereals	7.2	38.9**	28.4	63.8
Beans and other legumes	208.7	186.1**	115.6	112.9
Fruit	389.4	304.5**	616.6	391.1
Vegetables	143.4	94.4**	253.7	151.4
Potatoes and tubers	49.8	50.6*	47.4	54.7
Red meat	59.2	59.4**	45.4	48.9
Poultry	69.6	69.6**	87.9	82.9
Fish	36.3	41.9**	54.6	58.8
Eggs	12.1	14.2**	9.3	11.7
Processed meats	19.9	21.6**	17.7	20.1
Pasta	5.6	15.1	5.7	16.9
Salted snacks	10.3	14.8**	6.9	10.4
High-fat dairy products and milk	157.6	225.0**	63.3	143.7
Lower-fat dairy products and milk	76.9	169.6**	223.4	258.2
Sugary beverages	204.7	241.2**	138.3	182.8

1. Kcal/day.

p-values are derived from Student's t-tests.

\*p < 0.05, \*\*p < 0.001

Table 3. Performance measures by the ML algorithms.

	<b>SVM</b>	<b>NB</b>	<b>KNN</b>	<b>DT</b>
Accuracy	0.71	0.70	0.69	0.70
95% CI	(0.69, 0.72)	(0.69, 0.72)	(0.68, 0.71)	(0.69, 0.72)
Sensibility	0.62	0.66	0.74	0.63
Specificity	0.77	0.74	0.64	0.76
Positive predictive value	0.69	0.67	0.72	0.68
Negative predictive value	0.72	0.73	0.65	0.72

SVM: support vector machine; NB: naïve Bayes; KNN: K-nearest neighbors; DT: decision tree.

Table 4. Feature selection by the ML algorithms.

Order	Algorithm							
	SVM		DT		KNN		NB	
1	Per capita income	0.69	Sex	1.00	Per capita income	0.68	Per capita income	0.68
2	Education level	0.66	Per capita income	1.00	Education level	0.65	Education level	0.65
3	Physical activity	0.63	Physical activity	0.75	Physical activity	0.63	Physical activity	0.63
4	Age	0.61	Education level	0.72	Age	0.61	Sex	0.62
5	Sex	0.59	Age	0.62	Sex	0.60	Age	0.60
6	Race/ethnicity	0.58	Smoking habit	0.25	Race/ethnicity	0.58	Race/ethnicity	0.58
7	Waist-to-hip ratio	0.57	Clinical research center	0.23	Waist-to-hip ratio	0.57	Waist-to-hip ratio	0.57
8	Active worker	0.56	Diabetes	0.20	Active worker	0.56	Active worker	0.56
9	Smoking habit	0.54	Race/ethnicity	0.18	Smoking habit	0.55	Smoking habit	0.54
10	Marital status	0.54	Active worker	0.12	Marital status	0.54	Marital status	0.54
11	Living alone	0.54	Cardiovascular disease	0.09	Dyslipidemia	0.54	Dyslipidemia	0.54
12	Dyslipidemia	0.53	Living alone	0.09	Living alone	0.54	Live alone	0.54
13	Health self-assessment	0.52	Health self-assessment	0.07	Health self-assessment	0.53	Health self-assessment	0.52
14	Clinical research center	0.51	Hypertension	0.03	Hypertension	0.51	Diabetes	0.51
15	Diabetes	0.51	Waist-to-hip ratio	0.02	Clinical research center	0.51	Clinical research center	0.51
16	Cardiovascular disease	0.51	Marital status	0.02	Diabetes	0.51	Body mass index	0.51
17	Hypertension	0.50	Body mass index	0.02	Body mass index	0.51	Cardiovascular disease	0.51
18	Body mass index	0.50	Dyslipidemia	0.01	Cardiovascular disease	0.50	Hypertension	0.50

SVM: support vector machines; DT: decision trees; NN: K-nearest neighbors; NB: naïve Bayes.

## Supplementary Material

Supplementary Table 1. Confusion matrix.

<b>Support Vector Machine</b>			<b>Decision Tree</b>		
	Reference			Reference	
Prediction-based pattern	Western	Prudent	Prediction-based pattern	Western	Prudent
Western	<b>1,632</b>	634	Western	<b>1,604</b>	626
Prudent	483	<b>1,052</b>	Prudent	511	<b>1,060</b>

<b>K-Nearest Neighbors</b>			<b>Naïve Bayes</b>		
	Reference			Reference	
Prediction-based pattern	Western	Prudent	Prediction-based pattern	Western	Prudent
Western	<b>1,651</b>	630	Western	<b>1,564</b>	578
Prudent	496	<b>1,023</b>	Prudent	551	<b>1,108</b>

## 5.2. Sistema de Recomendações – Artigo 2

### **Recommender System Based on Collaborative Filtering for Personalized Dietary Advice: A Cross-sectional Analysis of the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil)**

Short title: Recommender System and Dietary Intake

Vanderlei Carneiro da Silva<sup>1,2</sup>, Bartira Gorgulho<sup>3</sup>, Dirce Maria Marchioni<sup>4</sup>, Tânia Aparecida de Araujo<sup>5</sup>, Itamar de Souza Santos<sup>2</sup>, Paulo Andrade Lotufo<sup>2</sup>, Isabela Martins Benseñor<sup>2</sup>

1. Postgraduate student, Department of Epidemiology, School of Public Health, University of São Paulo, Brazil.
2. Center of Clinical and Epidemiological Research, University Hospital, University of São Paulo, Brazil.
3. Department of Food and Nutrition, School of Nutrition, Federal University of Mato Grosso, Brazil.
4. Department of Nutrition, School of Public Health, University of São Paulo. Brazil.
5. School of Medicine, Federal University of Uberlandia, Brazil.



**Abstract:**

**Introduction:** Recommendation systems play an increasingly important role in modern online services. However, there are few examples of recommender systems for dietary advice. This study aimed to provide personalized dietary recommendations based on collaborative filtering for participants in a large study and to compare the recommender algorithm performance. **Methods:** This was a cross-sectional study with a sample that included 12,667 public employees from the ELSA-Brasil, aged 35-74 years, 59% women, who completed a 114-item semiquantitative Food Frequency Questionnaire (FFQ) at the study baseline. The FFQ data were collapsed into 21 food groups by food preparation or nutritional characteristics. The analysis was conducted with user-based collaborative filtering (UBCF) and item-based collaborative filtering (IBCF) algorithms. A maximum of five recommendations per participant was fixed to create a top-N recommendation list. **Results:** The eligible food groups were possible recommendations aimed to support healthier food choices. The user- and item-based approaches were similar (precision of 88-91%). However, some error metrics were lower for UBCF than for IBCF: root mean square error (RMSE: 1.49 vs. 1.67), mean square error (MSE: 2.21 vs. 2.78), and mean absolute error (MAE: 1.26 vs. 1.40). **Conclusion:** We implemented a recommender system to assess dietary intake data whose output may guide health professionals in personalized advice. Its use could impact adherence to interventions for healthier diets.

**Keywords:** recommender system; collaborative filtering; diet; dietary advice.

## **Introduction**

The adoption of a healthy lifestyle is recognized as an important component of chronic disease prevention and management (DESROCHES et al., 2013). Diseases such as diabetes, cardiovascular diseases, cancer, and obesity represent a major burden in health care systems worldwide. Among the main factors, physical exercise, reduction in alcohol consumption, smoking cessation, and healthy diet are essential to reduce the risk of these diseases. However, adherence to recommended behavior changes is generally extremely low. Nonadherence rates to the treatment of chronic diseases are estimated to be between 50 and 80% (WHO, 2003).

Over the years, health management on a personal level has evolved and been supported by technology (HORS-FRAILE et al., 2018a). Recommender systems are widely used to help users find new items or services, such as books, music, transportation or even people (PORTUGAL; ALENCAR; COWAN, 2018). Among existing recommender systems, collaborative filtering (CF) has gained significant success. Two approaches are common, although a combination of both techniques can be employed. User-based CF methods recognize people related to the analyzed individual and predict the rating to be the average ratings of similar people. In the same way, item-based CF identifies items related to the demanded product (for example, a movie) and predicts the rating to be the average of the ratings of similar products (THOMAS; SUJATHA, 2016).

In the field of health, recommender systems have been designed with a large amount of data to assist experts in making clinical decisions and treatment recommendations (MUSEN; MIDDLETON; GREENES, 2014; MUSTAQEEM; ANWAR; MAJID, 2020; WIESNER; PFEIFER, 2014). Among health recommender systems, there are applications for medication prescription (GHASEMI et al., 2019),

support for smoking cessation (HORS-FRAILE et al., 2018b) and depression and anxiety management (CHEUNG et al., 2018). Recommender systems are still not widely used in support of personalized dietary advice, which indicates a research opportunity to fill this gap (CHEN et al., 2018; NOUH et al., 2019). Effective health interventions that increase adherence to healthier dietary intake, for example, are always a great challenge (LOVEGROVE; GITAU, 2008). Moreover, most of these systems were proposed for specific health issues and lack a focus on preventing diseases.

A set of recommendations that aims to be effective in influencing behavioral changes should target segments of the population with messages personalized to their needs (TRAILL; CHAMBERS; BUTLER, 2012). Evidence has demonstrated that many people face the problem of making healthier food decisions that will impact their risk of noncommunicable diseases (GE; RICCI; MASSIMO, 2015). The use of dietary recommender systems could significantly contribute to health care and guide professionals to identify specific sociodemographic and clinical profiles as well as food items that are more likely to lead to adherence. This study aimed to provide personalized dietary recommendations based on collaborative filtering for participants in the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil) and to compare the recommender algorithm performance. We test the hypothesis that dietary advice may be assisted by the recommender system to predict a list of the best-N dietary recommendations for different groups.

## **Materials and Methods**

### **Study population**

ELSA-Brasil is a prospective cohort study designed to investigate the incidence of cardiovascular diseases and diabetes and their biological, environmental, occupational and social determinants. The participants were 15,105 public employees of teaching and research institutions, male and female, aged 35–74 years, recruited in six cities in Brazil (AQUINO et al., 2012; BENSENOR et al., 2013). In this analysis, cross-sectional data from the baseline examination (2008-2010) were included. The study was approved by the Ethics Committee under number 2.566.286 and conducted according to the guidelines of the Declaration of Helsinki. All participants were informed about the research objectives and signed a declaration of free and informed consent.

#### Dietary Assessment

The Food Frequency Questionnaire (FFQ) developed for the ELSA-Brasil study presents a list of 114 food items and is based on a previously validated questionnaire (MOLINA et al., 2013c). Study participants were asked by trained interviewers about their frequency and amount of consumption of each food over the 12-month period preceding the interview. The reproducibility and relative validity of the FFQ were evaluated in a study with 281 participants by intraclass correlation coefficient (ICC). The results demonstrated satisfactory reliability for all nutrients and acceptable relative validity for energy, macronutrients, calcium, potassium, and vitamins E and C. Details about the elaboration (MOLINA et al., 2013c) and validation of the questionnaire (MOLINA et al., 2013a) can be obtained in previous publications.

The daily intake was quantified by the number of servings consumed per day x weight (standard portion in grams) x frequency of consumption x nutritional composition of the food serving. The daily equivalent coefficients used were 3 for more than 3

times/day, 2.5 for 2 to 3 times/day, 1 for once/day, 0.8 for 5 to 6 times/week, 0.4 for 2 to 4 times/week, 0.1 for once/week, 0.07 for 1 to 3 times/month, and 0 for never/almost never. The Nutrition Data System for Research (NDSR) software (University of Minnesota, Minneapolis, USA, 2010) was used to determine the nutritional composition of foods.

We excluded 2,438 (16%) participants with an energy intake of less than 500 or greater than 4,000 Kcal/day (WILLET, 1998b). The final sample was composed of 12,667 public employees, of whom 5,217 (41%) were men and 7,450 (59%) were women.

#### Sociodemographic and clinical characteristics

The study required volunteers to visit the research center for clinical tests and interviews. This ensured a high standard of quality in data collection (BENSENOR et al., 2013). Sociodemographic and clinical data were collected and included in the analysis: sex (male vs. female), age (years), education level (elementary [or less], high school, college), retirement (no vs. yes), self-reported race/ethnicity (white, brown, black, other [Asian, Indigenous]), per capita income in US\$ categorized in terciles, using US\$1.00 = R\$2.00 as the approximate baseline examination exchange rate, living alone (no vs. yes), marital status (not married vs. married), smoking habit (never, ex-smoker, current smoker), physical activity (sedentary, insufficiently active, active [using the leisure time section of the long version of the International Physical Activity Questionnaire]), health self-assessment (good, regular, bad), and local clinical research center.

Weight and height measurements were obtained with the participant wearing light clothes and without shoes. We measured body weight to the nearest 0.1 kg with a calibrated balance (Toledo 2096PP) and height with a vertical stadiometer (Seca-SE-216)

to the nearest 0.1 cm. Body mass index (BMI) was calculated by dividing weight in kilograms by height in meters squared ( $\text{kg}/\text{m}^2$ ). Waist circumference was measured with a tape measure to the nearest 0.1 cm around the midpoint between the inferior costal border and the iliac crest, while hip circumference was measured at the point of greatest circumference in the gluteal region. The waist-to-hip ratio (WHR) was calculated by dividing waist by hip in centimeters.

Blood pressure (BP) was measured using a validated Omron HEM 705CPINT oscillometer device. Three measurements were taken at one-minute intervals, and the mean of the two latter blood pressure measurements was considered the value for defining hypertension, defined as systolic blood pressure  $\geq 140$  mm Hg, diastolic blood pressure  $\geq 90$  mm Hg or verified treatment with antihypertensive drugs during the last two weeks. Dyslipidemia was defined as LDL cholesterol  $\geq 130$  mg/dL or the use of medication to treat dyslipidemia. Diabetes was defined as a previous diagnosis of diabetes, use of medication to treat diabetes, fasting plasma glucose  $\geq 126$  mg/dL, 2-h plasma glucose  $\geq 200$  mg/dL, or  $\text{HbA}_{1\text{C}} \geq 6.5\%$ . Cardiovascular disease was defined as self-reported prior myocardial infarction, stroke or revascularization.

### Statistical Analysis

Continuous variables are presented as the mean and standard deviation (SD), and categorical variables are presented as frequencies. Associations between categorical variables were tested through the chi-square test. Comparisons of the mean values of continuous variables by data set (i.e., Train or Test) were performed using Student's t-test.

## Recommender System

All analyses were performed on the recommender lab package, R software, version 4.0.2. The dataset was mapped in the form of a rating matrix by the creator function recommender. Each row indicates a study participant, and the column indicates a food group. Participants were randomly divided into two subsets. The first was used for training (70%), and the second was used for testing (30%). The 114 foods from FFQ were collapsed into 21 groups by food preparation or nutritional characteristics: refined cereals; whole cereals; tubers and roots; breads; confectionery; beans and other legumes; oilseeds; fruits; vegetables; red meats; white meats and fish; processed meats; eggs; high-fat dairy products and milk; low-fat dairy products and milk; oils and fats; salted snacks; juices and other beverages; soft drinks; and alcoholic beverages. Dietary intake data were categorized into quintiles and were transformed into ratings, that is, a scale between 1 and 5.

We used user- and item-based collaborative filtering algorithms, and a range maximum of 5 recommendations was fixed to avoid unspecified and very extensive recommendations. The list of items eligible as recommendations included 8 of the 21 food groups used as input: whole cereals; tubers and roots; beans and other legumes; oilseeds; fruits; vegetables; white meats and fish; low-fat dairy products and milk. These food groups were chosen based on current Brazilian dietary guidelines with a focus on diets with a recognized impact on health promotion and reduction of the risk of chronic diseases, referenced by the local culture (MS, 2014). The system was designed to support personalized dietary advice and to predict a list of the best-N dietary recommendations.

Missing values (absence of consumption) represented foods that could be used for the recommendation. If a participant had a missing value in the dataset, for example,

absence of the consumption of beans, these foods could be recommended. Formally, missing values were replaced by estimated ratings. The system identified users who shared the same food preference and suggested items by similarity. For every participant, the algorithm identified the K-most similar. The Pearson correlation coefficient was used as a measure of similarity of the participants. The prediction function was used to predict ratings of unknown items. Next, to create a top-N recommendation list, the food items were ordered by predicted rating. Finally, the function `calcPredictionAccuracy` was used to calculate the accuracy of the predictions.

## Results

Table 1 shows the characteristics of the study population. There were no statistically significant differences between the two groups according to socioeconomic and health data. The most common sample consisted of women (59%), with a mean age of 52 years old ( $\pm 9$  SD), high education level, active worker, white, and not single. They reported being mostly nonsmokers and sedentary and self-reported their health as good. Among health and clinical characteristics, the body mass index mean was  $26.9 \text{ kg/m}^2$  ( $\pm 4.7$  SD) and the waist-to-hip Ratio was  $0.9$  ( $\pm 0.1$  SD). The frequency of dyslipidemia was 60%, hypertension 36%, diabetes 26% and cardiovascular disease 4%.

Figure 1 shows the recommender system architecture that was implemented. It was designed to use dietary intake data as input in the system. The algorithms based on the analysis of the participant's diet produce a list of recommendations (output) that is personalized and can guide dietary advice.

Table 2 shows the food groups and statistics as the mean and standard deviation. All groups were analyzed by recommender engines. Eligible recommendations were as



follows: whole cereals; tubers and roots; beans and other legumes; oilseeds; fruits; vegetables; white meats and fish; low-fat dairy products and milk.

Table 3 shows the error metrics by model. Root mean square error (RMSE), mean squared error (MSE), and mean absolute error (MAE) were used to compute the deviation of the prediction from the true value. Compared to item-based collaborative filtering (IBCF), user-based collaborative filtering (UBCF) had a lower error rate, RMSE: 1.49 vs. 1.67; MSE: 2.21 vs. 2.78; MAE: 1.26 vs. 1.40, respectively.

Table 4 compares the performance by model and k nearest neighbors. UBCF and IBCF showed similar performance with precision between 0.88 and 0.91 and a plateau when  $k = 10$  was used. The precision refers to the percentage of recommended food items with intake, while recall refers to the percentage of intake food items that have been recommended. Other metrics are also presented. There are no differences between the models.

The Supplementary Materials show plots for the ROC curve and precision recall. Both confirm similar performance between the two algorithms.

## **Discussion**

Our system was trained to suggest items to support healthier food choices. The recommendations were based on the participant's diet and calculations of similarity among peer users with a similar diet. The results showed that there were no differences between user- and item-based collaborative filtering approaches regarding their performance. This finding confirms our hypothesis that a recommender engine can analyze individuals' diet characteristics and provide personalized dietary recommendations.

Although previous studies have used some different methods, they have also applied recommendation systems in the field of nutrition (CHEN et al., 2018; FRANCO, 2017; NOROUZI et al., 2018). Chen et al. applied deep learning neural network models and compared different data sets from grocery products with accuracies of 72-84% (CHEN et al., 2018). Norouzi et al. analyzed Iranian women and men (n=30) and focused on the development of a food recommender system for managing diabetic patients' nutrition (NOROUZI et al., 2018). The roulette wheel algorithm was used, and a snack with a higher ranking was recommended to the patient. The results showed that the system recommended various snacks according to the season (accuracy of 100%) and personal interest (accuracy of 90%) for diabetic patients. Our results were like those of previous studies, with accuracies between 88% and 91%.

The ratings for an individual can be predicted by first finding a neighborhood of similar users and then aggregating the ratings of these users to form a prediction (DA SILVA et al., 2016; THOMAS; SUJATHA, 2016). The premise is that users who agree on the intake profile for some foods typically also agree on the rating for other items. Our recommender engine used food intake data that were transformed into ratings. The system locates peer users with a similar diet, and the foods with the highest rating predicted for an individual are recommended. Various statistical techniques, such as Euclidean distance, cosine similarity and Pearson correlation, can be used to compute the similarity between users (HAN; KAMBER; PEI, 2011). In the proposed approach, Pearson's correlation coefficient was used to find the nearest neighborhoods.

Typically, a person's dietary intake is assessed and then used as an input for decision-making to provide feedback to the person (FRANCO, 2017). This concept was implemented in the architecture of our system, as shown in Figure 1. The system was designed to support healthier and less processed food choices with a recognized impact

on risk reduction and the prevention of chronic diseases (BOEING et al., 2012; REDDY; KATAN, 2004). Although recommendation systems are expanding in many areas, they are still underutilized in diet, especially with a preventive nature. An advantage of their use is that personalized advice is more effective than general population-based recommendations for modifying health-related behavior in nutrition interventions (CELIS-MORALES et al., 2017). Furthermore, knowledge about healthy eating is not sufficient on its own to change eating behavior, but individualized feedback has been found to be associated with higher adherence to interventions to promote healthy lifestyles (HELANDER et al., 2014).

Train users were used to learn the recommender model, whereas test users were used to evaluate the recommendations. The recommender system predicts the preferences and ratings that a user would give to an item (THOMAS; SUJATHA, 2016). Some foods were withheld from the testing base before the recommendations were created. It was assumed that if a recommender algorithm performed better in predicting the withheld items, it would also perform better in finding good recommendations for unknown items (HAHSLER, 2017). The difference between the finally predicted value and the actual correct answer is defined as an error value. The results showed very small differences in favor of user-based collaborative filtering, so the two systems could be useful if applied in the context of clinical care. However, user-based CF may consider a user's social environment. This can be useful for the system to recommend healthy foods that are also part of the individual's culture, that is, foods that are present in the diet of his or her peers. This technique is based on the concept that users who have interest in a particular item may have similar interest in other items (RICCI et al., 2011).

In contrast to many common recommender systems (e.g., online shopping based on previous purchases), the food items that users like and consume the most are not

necessarily the healthiest (FRANCO, 2017). Thus, recommendations were based on the system's ability to provide suggestions for healthy foods and an emphasis on items to which the user could adhere. We limited the maximum number of recommendations, although other cut-offs could also be established. Extensive recommendations can discourage adherence, and dietary changes are more effective when they are adopted gradually (HILL, 2009). Furthermore, while habits are consolidated, a new reassessment can be conducted. Recommendations that consider changes in the patient's situation should be adapted over time.

The main contribution of our study is the presentation of a tool to promote healthier choices to which users are likely to adhere and that can be beneficial for groups of healthy people and for individuals affected by chronic diet-related diseases. In addition, health services and specialists can gain a better understanding of patient needs by obtaining more accurate models of their users (CHEUNG et al., 2019).

Further research is necessary to assess the effectiveness of existing systems for diet management. Communication technologies provide new potential and offer several advantages, such as lowering costs and improving outcomes, by reaching a larger segment of the target population (KELLY et al., 2016). On the other hand, it is a consistent finding that human support is also necessary to ensure adherence (i.e., following the intervention protocol) and to increase the effects (KELDERS et al., 2015). Therefore, technology does not replace specialists but can represent benefits in more personalized health care.

This study has some strengths. The data analyzed were from a large sample of older and middle-aged adult individuals and incorporated a catalog of typical/regional foods. Although the sample consisted only of civil servants, it aggregated an admixed multiethnic population and captured non-isolated eating practices. The study helps to fill

an important gap in the literature since there are few examples of food recommender systems that provide user contents to improve the quality of diet. The system could be used as a clinical decision support system. For data collection, an FFQ was developed and validated in the study population. Participants were invited to attend a clinical research center for exams and clinical evaluations, which guaranteed a high standard of quality control in the data used in the study.

Some limitations should also be addressed. The collected dietary data were self-reported and are subject to the interviewees' memory bias. The FFQ allows for the collection of participants' usual consumption regardless of intraindividual variability in addition to ranking people in consumption ranges. However, it has limitations like other collection methods, especially with regard to not capturing details about the diet such as tastes, preferences, negative or positive reactions to certain foods and preparations, restrictions, intolerance, allergies or even main concerns about diet and health that are relevant to dietary advice. Therefore, recommendations should be interpreted in a holistic context, especially in the strata of the population with socioeconomic restrictions.

We implemented an automated system capable of guiding health professionals in personalized dietary advice. The recommendations were established based on the analysis of participants' dietary intake data. The algorithm evaluated the set of possible recommendations and highlighted those to which participants were most likely to adhere. This work opens discussion about applications of automated intelligence systems in the field of nutrition. Future work can assess whether adherence to recommendations differs when an automated tool is used to support a human expert compared to interventions without the support of technological tools.

**Funding**

The ELSA-Brasil study received funding from the Ministry of Health of Brazil (Department of Science and Technology), Ministry of Science, Technology and Innovation (FINEP, Funding of Studies and Projects) and CNPq (National Development Council Scientific and Technological Advice) grants 01 06 0010.00 RS; 01 06 0212.00 BA; 01 06 0300.00 ES; 01 06 0278.00 MG; 01 06 0115.00 SP and 01 06 0071.00 RJ. However, this subproject did not receive funding. Funding received at the study baseline had no influence on the design, analysis, drafting, interpretation or decision of the version submitted for publication.

**Conflict of interests**

The authors declare no conflicts of interest.

**Author contributions**

V.C.S. and I.M.B. contributed to the design, acquisition, analysis, interpretation of data, drafting, and revision of the text; B.G., D.M.M. and T.A.A. contributed to the design, analysis, and revision of the text; I.S.S. and P.A.L. contributed to the acquisition, analysis, and revision of the text. All authors gave final approval of the version to be published.

## References

AQUINO, E. M. L. et al. Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): Objectives and Design. **American Journal of Epidemiology**, v. 175, n. 4, p. 315–324, 15 fev. 2012.

BENSENOR, I. M. et al. Routines of organization of clinical tests and interviews in the ELSA-Brasil investigation center. **Rev Saúde Pública**, v. 47, n. suppl 2, p. 37–47, jun. 2013.

BOEING, H. et al. Critical review: vegetables and fruit in the prevention of chronic diseases. **European Journal of Nutrition**, v. 51, n. 6, p. 637–663, 9 set. 2012.

CELIS-MORALES, C. et al. Effect of personalized nutrition on health-related behaviour change: evidence from the Food4me European randomized controlled trial. **International Journal of Epidemiology**, v. 46, n. 2, p. 578–88, 14 ago. 2017.

CHEN, C. et al. PERSON — Personalized Expert Recommendation System for Optimized Nutrition. **IEEE Trans Biomed Circuits Syst**, v. 12, n. 1, p. 151–160, 2018.

CHEUNG, K. et al. Evaluation of a recommender app for apps for the treatment of depression and anxiety: An analysis of longitudinal user engagement. **Journal of the American Medical Informatics Association**, v. 25, n. 8, p. 955–962, 2018.

CHEUNG, K. L. et al. **How recommender systems could support and enhance computer-tailored digital health programs: A scoping review** *Digital Health*, 2019.

DA SILVA, E. Q. et al. An evolutionary approach for combining results of recommender systems techniques based on collaborative filtering. **Expert Systems with Applications**, v. 53, p. 204–218, jul. 2016.

DESROCHES, S. et al. Interventions to enhance adherence to dietary advice for

preventing and managing chronic diseases in adults. **Cochrane Database of Systematic Reviews**, n. 2, 28 fev. 2013.

FRANCO, R. Z. **Online Recommender System for Personalized Nutrition Advice**. Proceedings of the Eleventh ACM Conference on Recommender Systems. **Anais...**New York, NY, USA: ACM, 27 ago. 2017

GE, M.; RICCI, F.; MASSIMO, D. Health-aware Food Recommender System. **Proceedings of the 9th ACM Conference on Recommender Systems**, n. September, p. 16–20, 2015.

GHASEMI, S. H. et al. Design and Evaluation of a Smart Medication Recommendation System for the Electronic Prescription. **Stud Health Technol Inform**, v. 260, p. 128–135, 2019.

HAHSLER, M. recommenderlab: A Framework for Developing and Testing Recommendation Algorithms. **R package version 0.2-2.**, n. <http://lyle.smu.edu/IDA/recommenderlab/>, 2017.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: Concepts and Techniques**. 3rd. ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

HELANDER, E. et al. Factors Related to Sustained Use of a Free Mobile App for Dietary Self-Monitoring With Photography and Peer Feedback: Retrospective Cohort Study. **Journal of Medical Internet Research**, v. 16, n. 4, p. 1–13, 15 abr. 2014.

HILL, J. O. Can a small-changes approach help address the obesity epidemic? A report of the Joint Task Force of the American Society for Nutrition, Institute of Food Technologists, and International Food Information Council. **The American Journal of Clinical Nutrition**, v. 89, n. 2, p. 477–484, 1 fev. 2009.



HORS-FRAILE, S. et al. Analyzing recommender systems for health promotion using a multidisciplinary taxonomy: A scoping review. **International Journal of Medical Informatics**, v. 114, n. December 2017, p. 143–155, 2018a.

HORS-FRAILE, S. et al. A recommender system to quit smoking with mobile motivational messages: study protocol for a randomized controlled trial. **Trials**, v. 19, n. 618, p. 1–12, 2018b.

KELDERS, S. M. et al. Comparing human and automated support for depression: Fractional factorial randomized controlled trial. **Behaviour Research and Therapy**, v. 72, p. 72–80, 2015.

KELLY, J. T. et al. Telehealth methods to deliver dietary interventions in adults with chronic disease: a systematic review and meta-analysis. **The American Journal of Clinical Nutrition**, v. 104, n. 6, p. 1693–1702, 1 dez. 2016.

LOVEGROVE, J. A.; GITAU, R. Personalized nutrition for the prevention of cardiovascular disease: a future perspective. **Journal of Human Nutrition and Dietetics**, v. 21, n. 4, p. 306–316, ago. 2008.

MOLINA, M. D. C. B. et al. Reprodutibilidade e validade relativa do Questionário de Frequência Alimentar do ELSA-Brasil. **Caderno de Saúde Pública**, v. 29, n. 2, p. 379–389, 2013a.

MOLINA, M. DEL C. B. et al. Diet assessment in the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): Development of a food frequency questionnaire. **Rev. Nutr.**, v. 26, n. 2, p. 167–176, abr. 2013b.

MS. **Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Atenção Básica. Guia Alimentar para a população brasileira.** Brasília, 2014.

MUSEN, M. A.; MIDDLETON, B.; GREENES, R. A. Clinical Decision-Support Systems. In: **Biomedical Informatics**. London: Springer London, 2014. p. 643–674.

MUSTAQEEM, A.; ANWAR, S. M.; MAJID, M. A modular cluster based collaborative recommender system for cardiac patients. **Artificial Intelligence in Medicine**, v. 102, n. September 2018, p. 101761, 2020.

NOROUZI, S. et al. A mobile application for managing diabetic patients' nutrition: A food recommender system. **Archives of Iranian Medicine**, v. 21, n. 10, p. 466–472, 2018.

NOUH, R. M. et al. A Smart Recommender Based on Hybrid Learning Methods for Personal Well-Being Services. **Sensors**, v. 19, n. 431, p. 1–21, 2019.

PORTUGAL, I.; ALENCAR, P.; COWAN, D. **The use of machine learning algorithms in recommender systems: A systematic review** *Expert Systems with Applications*, 2018.

REDDY, K. S.; KATAN, M. B. Diet, nutrition and the prevention of hypertension and cardiovascular diseases. **Public Health Nutrition**, v. 7, n. 1a, p. 167–186, 2 fev. 2004.

RICCI, F. et al. **Recommender Systems Handbook**. Boston, MA: Springer US, 2011.

THOMAS, A.; SUJATHA, A. . Comparative Study of Recommender systems. **2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)**, v. March, p. 1–6, 2016.

TRAILL, W. B.; CHAMBERS, S. A.; BUTLER, L. Attitudinal and demographic determinants of diet quality and implications for policy targeting. **Journal of Human Nutrition and Dietetics**, v. 25, n. 1, p. 87–94, fev. 2012.

WHO. **Adherence to long-term therapies: evidence for action**. Geneva, Switzerland:

[s.n.].

WIESNER, M.; PFEIFER, D. Health Recommender Systems: Concepts, Requirements, Technical Basics and Challenges. **International Journal of Environmental Research and Public Health**, v. 11, n. 3, p. 2580–2607, 3 mar. 2014.

WILLET, W. Correction for the effects of measurement error. In: WILLET, W. (Ed.). . **Nutritional Epidemiology**. 2nd. ed. New York, NY, USA: Oxford University Press, 1998. p. 74–147.

Table 1. Description of the study population, Longitudinal Study of Adult Health (ELSA-Brasil), 2008-10.

Variable	General		Train		Test		p-Value
	n	%	n	%	n	%	
Study population	12,667	100.0	8,866	100	3,801	100	
Sex							
Male	5,217	41.2	3,665	41.3	1,552	40.8	0.596
Female	7,450	58.8	5,201	58.7	2,249	59.2	
Age (years) <sup>1</sup>	52	±9	52	±9	52	±9	0.943
Education level							
Elementary (or less)	1,423	11.2	1,022	11.5	401	10.5	0.247
High school	4,072	32.2	2,829	31.9	1,243	32.7	
College	7,172	56.6	5,015	56.6	2,157	56.8	
Retirement							
No	10,046	79.3	7,012	79.1	3,034	79.8	0.351
Yes	2,621	20.7	1,854	20.9	767	20.2	
Race/ethnicity							
White	6,994	55.2	4,887	55.1	2,107	55.4	0.986
Mixed	3,379	26.7	2,373	26.8	1,006	26.5	
Black	1,831	14.4	1,281	14.4	550	14.5	
Others <sup>2</sup>	463	3.7	325	3.7	138	3.6	
Marital status							
Not single	8,181	64.6	5,700	64.3	2,481	65.3	0.290
Single	4,486	35.4	3,166	35.7	1,320	34.7	
Per capita income <sup>3</sup>							
1° tercile	4,225	33.4	2,994	33.8	1,231	32.4	0.103
2° tercile	4,492	35.5	3,093	34.9	1,399	36.8	
3° tercile	3,950	31.2	2,779	31.3	1,171	30.8	
Living alone							
No	11,043	87.2	1,157	13.1	467	12.3	0.239
Yes	1,624	12.8	7,709	86.9	3,334	87.7	
Smoking habit							
Never	7,306	57.7	5,137	57.9	2,169	57.0	0.440
Ex-smoker	3,780	29.8	2,643	29.8	1,137	29.9	
Current smoker	1,581	12.5	1,086	12.3	495	13.0	
Physical activity <sup>4</sup>							
Sedentary	5,798	45.8	4,022	45.4	1,776	46.7	0.354
Insufficiently active	3,354	26.5	2,371	26.7	983	25.8	
Active	3,515	27.7	2,473	27.8	1,042	27.4	

"Continue"

Table 1. Description of the study population, Longitudinal Study of Adult Health (ELSA-Brasil), 2008-10.

"Continued"

Variable	General		Train		Test		p-Value
	n	%	n	%	n	%	
Health self-assessment							
Good	10,266	81.1	7,191	81.1	3,075	80.9	
Regular	2,167	17.1	1,510	17.0	657	17.3	
Bad	234	1.8	165	1.9	69	1.8	0.930
BMI (kg/m <sup>2</sup> ) <sup>1</sup>	27.0	±4.7	26.9	±4.7	27.0	±4.8	0.679
Waist-to-hip ratio <sup>1</sup>	0.9	±0.1	0.9	±0.1	0.8	±0.1	0.254
Dyslipidemia <sup>5</sup>							
No	5,237	41.3	3,666	41.4	1,571	41.3	
Yes	7,430	58.7	5,200	58.6	2,230	58.7	0.985
Hypertension <sup>6</sup>							
No	8,159	64.4	5,708	64.4	2,451	64.5	
Yes	4,508	35.6	3,158	35.6	1,350	35.5	0.912
Diabetes <sup>7</sup>							
No	10,634	83.9	7,450	84.0	3,184	83.8	
Yes	2,033	16.1	1,416	16.0	617	16.2	0.713
Cardiovascular disease <sup>8</sup>							
No	12,188	96.2	8,529	96.2	3,659	96.2	
Yes	479	3.8	337	3.8	142	3.7	0.860

1. Mean, Standard Deviation.

2. Others = Asian + Indigenous.

3. Calculation based on 2009: 1US\$ = R\$ 2,00.

4. Sedentary: Does not perform physical activity; Insufficiently Active: &lt;150 minutes/week of exercise &lt; 3 days a week; Active: 150 minutes/week at least 3 days a week.

5. LDL ≥130 mg/dL or use of cholesterol reducers.

6. Systolic blood pressure ≥140 mmHg, diastolic blood pressure ≥90 mmHg or verified treatment with antihypertensive drugs during the last two weeks.

7. Defined as an account of a previous diagnosis of diabetes or the use of medication for diabetes or to fulfill a diagnostic value of diabetes.

8. Defined as a report of a heart attack, stroke or revascularization.

p-values derived from Student's t-test or the chi-square test.

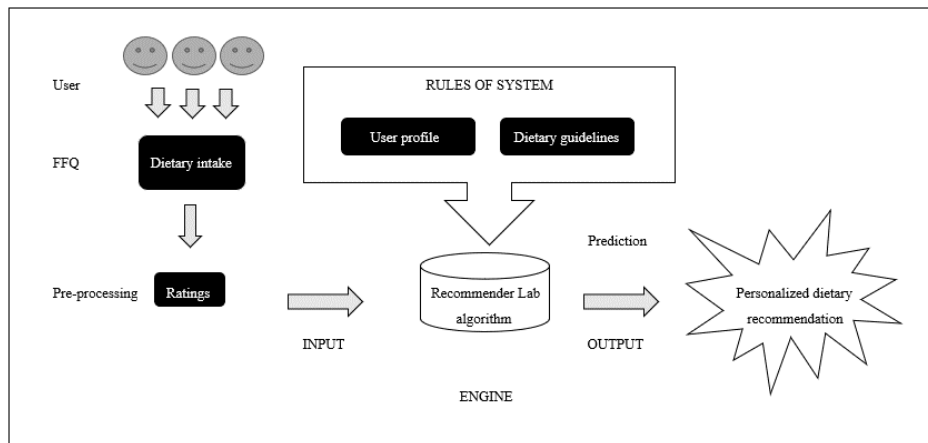


Figure 1. The architecture of the dietary recommender system for participants of the ELSA-Brasil study.

Table 2. Food groups and dietary intake.

<b>Food groups</b>	<b>Foods</b>	<b>Mean</b>	<b>±SD</b>	<b>Recommender</b>
1 Refined cereals	White rice, refined grains, corn, and breakfast cereals	157.89	144.46	
2 Whole cereals	Whole-grain rice, oats, granola, whole cereals, and whole meal bread,	46.66	71.63	X
3 Tubers and roots	Potato (stewed, boiled or mashed), boiled cassava, yam, sweet potato, fried potato, cassava, polenta, sweet potato, farofa (cassava flour with complements), cornmeal cuscus; cassava flour and corn flour	48.08	52.46	X
4 Breads	Bread, rolls, light bread, pita bread, and simple cake	80.26	67.76	
5 Confectionery	Sweets, caramels, cereal bar, sweet biscuit with filling, cake with filling, sweet bread, tart, trifle, chocolate bar, bonbons, chocolates, chocolate powder, chocolate drink, cappuccino, milk pudding, milk sweet, mousse, jelly, jam, fruit sweets, honey, syrup, sweet bread, fruit ice lolly, fruit salad with complements, and ice cream	84.41	69.88	
6 Beans and other legumes	Beans (black, red, white, cowpeas, etc.), lentils, chickpeas and peas	168.20	165.03	X
7 Oilseeds	Walnuts, cashew nuts, Brazil nuts, peanuts, almonds and pistachio nuts	5.39	11.49	X
8 Fruits	Avocado, pineapple, banana, persimmon, jack, pine nut, guava; orange, tangerine, apple, pear; papaya, mango, watermelon, melon, strawberry, and fruit salad without complements	503.28	369.01	X

"Continue"

Table 2. Food groups and dietary intake.

*“Continued”*

<b>Food groups</b>	<b>Foods</b>	<b>Mean</b>	<b>±SD</b>	<b>Recommender</b>
9 Vegetables	Zucchini, chayote, eggplant, pumpkin, lettuce, garlic, beetroot, broccoli, onion, carrot, chicory, watercress, arugula, raw collards, raw spinach, escarole, braised collards, braised spinach, cauliflower, okra, tomato, runner beans, vegetable soup	191.37	134.14	X
10 Red meats	Boned beef (steak, minced meat and stewed meat), beef on the bone (rib, oxtail), pork, tripe, liver, offal, and stroganoff	76.76	63.11	
11 White meats and fish	Boiled chicken, fried chicken, chicken breast, boiled, baked or grilled fish, fried fish, shrimp, shellfish, crab, swimming crab, sardine, and tuna	125.10	97.68	X
12 Processed meats	Sausage, light cold meats (turkey sausage, turkey breast), ham, mortadella, salami, and bacon	19.93	21.43	
13 Eggs	Fried egg, omelet, scrambled egg, and boiled and poached egg	10.86	13.24	
14 High-fat dairy products and milk	Full-cream yogurt, full-cream dairy milk, yellow cheeses, and full-cream cheese spread,	132.93	202.21	
15 Lower-fat dairy products and milk	Light yogurt, semi-skimmed dairy milk, skimmed dairy milk, white cheeses, and light cheese spread	164.93	232.49	X

*“Continue”*



Table 2. Food groups and dietary intake.

*“Continued”*

<b>Food groups</b>	<b>Foods</b>	<b>Mean</b>	<b>±SD</b>	<b>Recommender</b>
16 Oils and fats	Margarine, butter, mayonnaise, and light mayonnaise	5.07	5.60	
17 Pasta	Macaroni, macaroni dishes, and instant macaroni	30.44	35.85	
18 Salted snacks	Hotdog, popcorn, fried and baked savories, pizza, and instant soup	30.86	29.25	
19 Juices and other beverages	Coffee with sugar, coffee with sweetener, coffee with no sugar, mate with sugar, mate with sweetener, sugar-free mate, soy extract, artificial juice with sugar, artificial juice with sweetener, sugar-free artificial juice, industrialized juice with sugar, industrialized juice with sweetener, sugar-free industrialized juice, natural juice with sugar, natural juice with sweetener, sugar-free natural juice, and coconut water	403.98	263.65	
20 Soft drinks	Soft drinks, and diet/light soft drinks	86.02	160.36	
21 Alcoholic beverages	Beer, wine, and distilled beverages	100.09	203.05	

Table 3 Evaluation of prediction accuracy by model.

	RMSE	MSE	MAE
User-Based CF	1.49	2.21	1.26
Item-Based CF	1.67	2.78	1.40

1. Root Mean Square Error.

2. Mean Squared Error.

3. Mean Absolute Error.

Table 4. Confusion matrix by model.

<b>User-based Collaborative Filtering (UBCF)</b>								
K	TP	FP	FN	TN	Precision	Recall	TPR	FRP
1	0.88	0.12	9.08	0.92	0.88	0.09	0.09	0.11
3	2.70	0.30	7.25	0.75	0.90	0.27	0.27	0.27
5	4.54	0.46	5.41	0.59	0.91	0.46	0.46	0.43
10	9.11	0.89	0.85	0.15	0.91	0.91	0.91	0.84

<b>Item-based Collaborative Filtering (IBCF)</b>								
K	TP	FP	FN	TN	Precision	Recall	TPR	FRP
1	0.88	0.12	9.08	0.92	0.88	0.09	0.09	0.12
3	2.66	0.34	7.30	0.70	0.89	0.27	0.27	0.33
5	4.45	0.55	5.51	0.49	0.89	0.45	0.45	0.53
10	9.03	0.97	0.93	0.07	0.90	0.91	0.91	0.93

k: The k nearest neighbors

True Positives (TP): Recommended items with intake

False Positives (FP): Recommended items that without intake

False Negatives (FN): Not recommended items with intake

True Negatives (TN): Not recommended items without intake

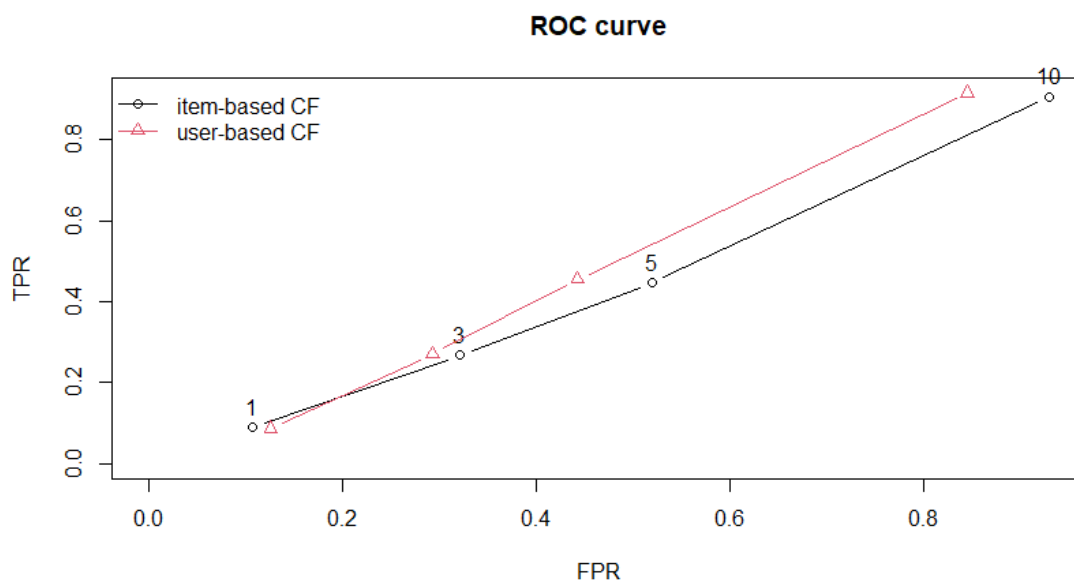
Precision: The percentage of recommended items with intake

Recall: The percentage of intake items that have been recommended

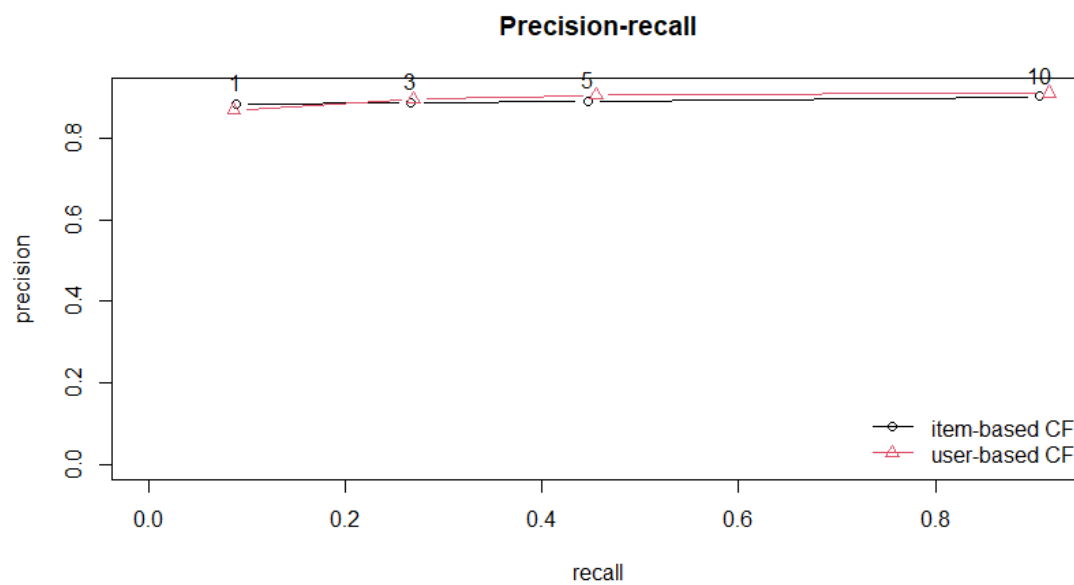
True Positive Rate (TPR): The percentage of intake items that have been recommended

False Positive Rate (FPR): The percentage of non-intake items that have been recommended

## Supplementary Data



Supplementary Figure 1. ROC curve.



Supplementary Figure 2. Precision recall.

## 6. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Este estudo apresentou resultados da aplicação de algoritmos de análise de dados baseados em *machine learning* na avaliação do consumo alimentar de servidores públicos em um grande estudo brasileiro. Foram descritas várias métricas de avaliação dos modelos o que permitiu a comparação da performance entre os algoritmos. A predição de padrões alimentares foi realizada com dados indiretos sobre a dieta, abordagem pouco explorada na literatura. Embora os modelos preditivos tenham alcançado, no geral, acurácia próximo a 70%, foi possível classificar os indivíduos a partir de um conjunto de dados demográficos, socioeconômicos e clínicos. Informações que são frequentemente coletadas em inquéritos de saúde e podem subsidiar atividades de rastreamento, especialmente quando dados sobre a dieta não estão disponíveis.

Com o modelo de análise utilizado, foi possível subdividir o processo de avaliação dos dados coletados com o Questionário de Frequência Alimentar, classificando os participantes do estudo em dois grandes *clusters* de dieta e identificar os fatores associados. Além disso, foram propostos dois sistemas de recomendação para personalizar o aconselhamento alimentar. Análises como as que foram realizadas, podem guiar o planejamento e a gestão de ações de alimentação e nutrição com recomendações personalizadas e maior aderência destas pela população por considerar o perfil do indivíduo e não apenas recomendações gerais. Contudo, o sistema foi proposto com base nos dados disponíveis e direcionado à hábitos alimentares saudáveis que podem beneficiar a amostra como um todo. Um sistema de recomendações para o tratamento de doenças e a gestão de condições clínicas não foi abordado no presente estudo, o qual demandaria um conjunto de outros parâmetros e modelos específicos, fora do escopo do estudo.

Este trabalho configura-se como um dos primeiros no Brasil na avaliação de dados de dieta apoiada por algoritmos. Estudos futuros podem comparar adesão do aconselhamento alimentar quando realizado com o apoio de sistemas personalizados e quando baseados em diretrizes gerais para a população.

Finalmente, os computadores, hoje disponíveis, podem executar cálculos em alto nível, com agilidade e precisão extraordinárias. No entanto, a avaliação do consumo alimentar é uma atividade complexa e a interpretação da relação homem-alimento não se esgota com funções matemáticas. A representação de expectativas, significados, queixas, comportamentos e sentimentos envolvidos nessa relação, demanda escuta qualificada (SCAGLIUSI, 2018). O encontro terapêutico já se reveste de cuidado e busca compreender os elementos objetivos e subjetivos da dieta, tais como: antropometria, exame físico, história de vida, condição de saúde atual e pregressa, estrutura familiar e social, acesso a serviços de saúde, aspectos econômicos, religiosos, filosóficos, crenças, entre outros. Essa atuação, indelegável à máquina, permite que a alimentação seja compreendida muito além de números e padrões. Sistemas artificiais utilizam números como linguagem de aprendizado, essa característica não inviabiliza os *insights* obtidos, entretanto deve ser compreendida como complementar (HEARTY; GIBNEY, 2008a).

## 7. REFERÊNCIAS BIBLIOGRÁFICAS

ABE, S. **Support Vector Machines for pattern classification**. 1° ed. New York: York, London Springer New, 2010.

ADOMAVICIUS, G.; TUZHILIN, A. Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. **IEEE transactions on knowledge and Data Engineering**, v. 17, n. 6, p. 734–749, 2005.

AGAPITO, G. et al. DIETOS: A dietary recommender system for chronic diseases monitoring and management. **Computer Methods and Programs in Biomedicine**, v. 153, p. 93–104, jan. 2018.

ALANAZI, H. O.; ABDULLAH, A. H.; QURESHI, K. N. A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care. **Journal of Medical Systems**, v. 41, n. 4, 2017.

ANDRADE, F. R. DE et al. Inquéritos populacionais como instrumentos de gestão e os modelos de atenção à saúde. **Revista de Saúde Pública**, v. 47, n. suppl 3, p. 154–160, dez. 2013.

AQUINO, E. M. L. et al. Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): Objectives and Design. **American Journal of Epidemiology**, v. 175, n. 4, p. 315–324, 15 fev. 2012.

ARRUDA, S. P. M. et al. Socioeconomic and demographic factors are associated with dietary patterns in a cohort of young Brazilian adults. **BMC Public Health**, v. 14, n. 654, p. 1–13, 2014.

AWAYSHEH, A. et al. Review of Medical Decision Support and Machine-Learning Methods. **Veterinary pathology**, p. 300985819829524, 13 mar. 2019.

BARALDI, L. G. et al. Consumption of ultra-processed foods and associated sociodemographic factors in the USA between 2007 and 2012: evidence from a nationally representative cross-sectional study. **BMJ open**, v. 8, n. 3, p. e020574, 2018.

BARUFALDI, L. A. et al. Programa para registro de recordatório alimentar de 24 horas: aplicação no Estudo de Riscos Cardiovasculares em Adolescentes. **REV BRAS**

**EPIDEMIOL**, v. 19, n. 2, p. 464–468, 2016.

BENSENOR, I. M. et al. Routines of organization of clinical tests and interviews in the ELSA-Brasil investigation center. **Rev Saúde Pública**, v. 47, n. suppl 2, p. 37–47, jun. 2013.

BERTIN, R. L. et al. Métodos de avaliação do consumo alimentar de gestantes: uma revisão. **Revista Brasileira de Saúde Materno Infantil**, v. 6, n. 4, p. 383–390, 2006.

BOEING, H. et al. Critical review: vegetables and fruit in the prevention of chronic diseases. **European Journal of Nutrition**, v. 51, n. 6, p. 637–663, 9 set. 2012.

BOEING, H. Nutritional epidemiology: New perspectives for understanding the diet-disease relationship? **European Journal of Clinical Nutrition**, v. 67, n. 5, p. 424–429, 2013.

BOLAND, M.; BRONLUND, J. eNutrition - The next dimension for eHealth? **Trends in Food Science & Technology**, v. 91, n. April, p. 634–639, 2019.

BOUSHEY, C. J. et al. **New mobile methods for dietary assessment: Review of image-assisted and image-based dietary assessment methods**. Proceedings of the Nutrition Society. **Anais...**2017

BOWMAN, S. A. et al. **The Healthy Eating Index 1994-96**, 1998.

BRASIL. **Constituição da República Federativa do Brasil de 1988**. Brasília, 1988.

BRASIL. **Lei 11.346, de 15 de setembro de 2006. Cria o Sistema Nacional de Segurança Alimentar e Nutricional – SISAN com vistas em assegurar o direito humano à alimentação adequada e dá outras providências**. Brasília, 2006.

BRASIL. **Decreto nº7.272, de 25 de agosto de 2010. Regulamenta a Lei no 11.346, que cria o SISAN, institui a Política Nacional de Segurança Alimentar e Nutricional - PNSAN, estabelece os parâmetros para a elaboração do PIANSAN, e dá outras providências**. Brasília, 2010.

CANESQUI, A. M. Pesquisas qualitativas em nutrição e alimentação. **Revista de Nutricao**, v. 22, n. 1, p. 125–139, 2009.

CARDOSO, L. DE O. et al. Eating patterns in the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): an exploratory analysis. **Cadernos de saude publica**, v. 32, n.

5, p. 1–14, jan. 2016.

CARVALHO, A. T. DE et al. Métodos de análise em programas de segurança alimentar e nutricional: uma experiência no Brasil. **Ciência & Saúde Coletiva**, v. 18, n. 2, p. 309–321, 2013.

CELIS-MORALES, C. et al. Effect of personalized nutrition on health-related behaviour change: evidence from the Food4me European randomized controlled trial. **International Journal of Epidemiology**, v. 46, n. 2, p. 578–88, 14 ago. 2017.

CHEN, C. et al. PERSON — Personalized Expert Recommendation System for Optimized Nutrition. **IEEE Trans Biomed Circuits Syst**, v. 12, n. 1, p. 151–160, 2018.

CHEUNG, K. et al. Evaluation of a recommender app for apps for the treatment of depression and anxiety: An analysis of longitudinal user engagement. **Journal of the American Medical Informatics Association**, v. 25, n. 8, p. 955–962, 2018.

CHEUNG, K. L. et al. **How recommender systems could support and enhance computer-tailored digital health programs: A scoping review** *Digital Health*, 2019.

CHIARA, V. L. et al. Food list reduction for a food frequency questionnaire: methodological issues. **Revista Brasileira de Epidemiologia**, v. 10, n. 3, p. 410–420, 2007.

CHOR, D. et al. Questionnaire development in ELSA-Brasil: Challenges of a multidimensional instrument. **Revista de Saude Publica**, v. 47, n. SUPPL2, p. 27–36, 2013.

COHEN, J. A Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, 2 abr. 1960.

CONTRERAS, I.; VEHI, J. **Artificial intelligence for diabetes management and decision support: Literature review** *Journal of Medical Internet Research*, 2018.

COSTA, A. G. V. et al. Questionário de frequência de consumo alimentar e recordatório de 24 horas: Aspectos metodológicos para avaliação da ingestão de lipídeos. **Revista de Nutricao**, v. 19, n. 5, p. 631–641, 2006.

CRISPIM, S. P.; SAMOFAL, P.; FERREIRA, G. R. Uso da tecnologia para a avaliação do consumo alimentar. In: MARCHIONI, D. M. L.; GORGULHO, B. M.; STELUTI, J.



(Eds.). . **Consumo Alimentar. Guia para Avaliação**. 1° ed. Barueri, SP: Manole, 2019. p. 107–122.

CUPPARI, L.; ANÇÃO, M. S. Uso de programas computadorizados na avaliação dietética. In: FISBERG, R. M. et al. (Eds.). . **Inquéritos Alimentares. Métodos e bases científicas**. 1° ed. Barueri, SP: Manole, 2007. p. 71–82.

CURIONI, C. C.; BRITO, F. DOS S. B.; BOCCOLINI, C. S. O uso de Tecnologias de Informação e Comunicação na Área da Nutrição. **J Bras Tele**, v. 2, n. December, p. 103–111, 2013.

DA SILVA, E. Q. et al. An evolutionary approach for combining results of recommender systems techniques based on collaborative filtering. **Expert Systems with Applications**, v. 53, p. 204–218, jul. 2016.

DE VASCONCELOS, F. DE A. G. A ciência da nutrição em trânsito: Da nutrição e dietética à nutrigenômica. **Revista de Nutricao**, v. 23, n. 6, p. 935–945, 2010.

DEMIDENKO, E. The next-generation K-means algorithm. **Statistical Analysis and Data Mining**, v. 11, n. 4, p. 153–166, 2018.

DESROCHES, S. et al. Interventions to enhance adherence to dietary advice for preventing and managing chronic diseases in adults. **Cochrane Database of Systematic Reviews**, n. 2, 28 fev. 2013.

DEVLIN, U. M. et al. The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy mis-reporting. **Proc Nutr Soc.**, v. 71, n. 4, p. 599–609, 2012.

DODD, K. W. et al. Statistical Methods for Estimating Usual Intake of Nutrients and Foods: A Review of the Theory. **Journal of the American Dietetic Association**, v. 106, n. 10, p. 1640–1650, out. 2006.

DUGAN, T. M. et al. Machine Learning Techniques for Prediction of Early Childhood Obesity. **Applied clinical informatics**, v. 6, n. 3, p. 506–20, 2015.

DUNCAN, B. B. et al. Doenças Crônicas Não Transmissíveis no Brasil: Prioridade para enfrentament e investigação. **Revista de Saude Publica**, v. 46, n. SUPPL.1, p. 126–134, 2012.

DUNJKO, V.; BRIEGEL, H. J. Machine learning & artificial intelligence in the quantum domain: A review of recent progress. **Reports on Progress in Physics**, v. 81, n. 7, 2018.

EHRENTAUT, C. et al. Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting. **Health Informatics Journal**, v. 24, n. 1, p. 24–42, 2018.

FISBERG, R. M. et al. **Inquéritos Alimentares. Métodos e bases científicas**. 1° ed. Barueri, SP: Manole, 2005.

FORNAZIN, M.; JOIA, A. L. Articulando perspectivas teóricas para analisar a informática em saúde no Brasil Linking theoretical perspectives to analyze health informatics in Brazil. **Saúde Soc.**, v. 24, n. 1, p. 46–60, 2015.

FRANCO, R. Z. **Online Recommender System for Personalized Nutrition Advice**. Proceedings of the Eleventh ACM Conference on Recommender Systems. **Anais...**New York, NY, USA: ACM, 27 ago. 2017

FREEDMAN, L. S. et al. Adjustments to Improve the Estimation of Usual Dietary Intake Distributions in the Population. **The Journal of Nutrition**, v. 134, n. 7, p. 1836–1843, 1 jul. 2004.

FU, M. et al. A Novel Deep Learning-Based Collaborative Filtering Model for Recommendation System. **IEEE Transactions on Cybernetics**, v. 49, n. 3, p. 1084–1096, 2019.

GARCIA, R. W. D. Representações Sociais da Alimentação e Saúde e suas Repercussões no Comportamento Alimentar. **PHYSIS: Rev. Saúde Coletiva**, v. 7, n. 2, p. 51–68, 1997.

GARCIA, R. W. D. Representações sobre consumo alimentar e suas implicações em inquéritos alimentares: estudo qualitativo em sujeitos submetidos à prescrição dietética. **Revista de Nutrição**, v. 17, n. 1, p. 15–28, mar. 2004.

GAYVERT, K. M.; MADHUKAR, N. S.; ELEMENTO, O. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. **Cell Chemical Biology**, v. 23, n. 10, p. 1294–1301, 2016.

GE, M.; RICCI, F.; MASSIMO, D. Health-aware Food Recommender System.

**Proceedings of the 9th ACM Conference on Recommender Systems**, n. September, p. 16–20, 2015.

GHASEMI, S. H. et al. Design and Evaluation of a Smart Medication Recommendation System for the Electronic Prescription. **Stud Health Technol Inform**, v. 260, p. 128–135, 2019.

GIABBANELLI, P. J.; ADAMS, J. Identifying small groups of foods that can predict achievement of key dietary recommendations: data mining of the UK National Diet and Nutrition Survey, 2008-12. **Public health nutrition**, v. 19, n. 9, p. 1543–51, jun. 2016.

GLYMOUR, M. M.; BIBBINS-DOMINGO, K. The Future of Observational Epidemiology: Improving Data and Design to Align With Population Health. **American Journal of Epidemiology**, v. 188, n. 5, p. 836–839, 1 maio 2019.

GUENTHER, P. M.; KOTT, P. S.; CARRIQUIRY, A. L. Development of an Approach for Estimating Usual Nutrient Intake Distributions at the Population Level. **The Journal of Nutrition**, v. 127, n. 6, p. 1106–1112, 1 jun. 1997.

GURINOVIĆ, M. et al. EURRECA nutritional planning and dietary assessment software tool: NutPlan. **European Journal of Clinical Nutrition**, v. 64, p. S38–S42, 2010.

GURINOVIĆ, M. et al. Development, features and application of DIET ASSESS & PLAN (DAP) software in supporting public health nutrition research in Central Eastern European Countries (CEEC). **Food Chemistry**, v. 238, p. 186–194, 2018.

HAHSLER, M. recommenderlab: A Framework for Developing and Testing Recommendation Algorithms. **R package version 0.2-2.**, n. <http://lyle.smu.edu/IDA/recommenderlab/>, 2017.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: Concepts and Techniques**. 3rd. ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

HAND, D. J.; MANNILA, H.; SMITH, P. **Principles of data mining**. 1° ed. Cambridge, Massachusetts: Press, Cambridge Mass MIT, 2001.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, v. 143, n. 1, p. 29–36, abr. 1982.

HARTTIG, U. et al. The MSM program: web-based statistics package for estimating usual dietary intake using the Multiple Source Method. **European Journal of Clinical Nutrition**, v. 65, n. S1, p. S87–S91, 6 jul. 2011.

HEARTY, A. P.; GIBNEY, M. J. Analysis of meal patterns with the use of supervised data mining techniques--artificial neural networks and decision trees. **The American journal of clinical nutrition**, v. 88, n. 6, p. 1632–1642, 2008a.

HEARTY, Á. P.; GIBNEY, M. J. Comparison of cluster and principal component analysis techniques to derive dietary patterns in Irish adults. **British Journal of Nutrition**, v. 101, n. 4, p. 598–608, 25 jun. 2008b.

HÉBERT, J. R. et al. Considering the value of dietary assessment data in informing nutrition-related health policy. **Advances in nutrition (Bethesda, Md.)**, v. 5, n. 4, p. 447–55, jul. 2014.

HELANDER, E. et al. Factors Related to Sustained Use of a Free Mobile App for Dietary Self-Monitoring With Photography and Peer Feedback: Retrospective Cohort Study. **Journal of Medical Internet Research**, v. 16, n. 4, p. 1–13, 15 abr. 2014.

HILL, J. O. Can a small-changes approach help address the obesity epidemic? A report of the Joint Task Force of the American Society for Nutrition, Institute of Food Technologists, and International Food Information Council. **The American Journal of Clinical Nutrition**, v. 89, n. 2, p. 477–484, 1 fev. 2009.

HOFFMANN, K. et al. Application of a New Statistical Method to Derive Dietary Patterns in Nutritional Epidemiology. **American Journal of Epidemiology**, v. 159, n. 10, p. 935–944, 15 maio 2004.

HORS-FRAILE, S. et al. Analyzing recommender systems for health promotion using a multidisciplinary taxonomy: A scoping review. **International Journal of Medical Informatics**, v. 114, n. December 2017, p. 143–155, 2018a.

HORS-FRAILE, S. et al. A recommender system to quit smoking with mobile motivational messages: study protocol for a randomized controlled trial. **Trials**, v. 19, n. 618, p. 1–12, 2018b.

HU, F. B. Dietary pattern analysis: a new direction in nutritional epidemiology. **Curr Opin Lipidol**, v. 13, n. 1, p. 3–9, 2002.

ILLNER, A.-K. et al. Review and evaluation of innovative technologies for measuring diet in nutritional epidemiology. **International Journal of Epidemiology**, v. 41, n. 4, p. 1187–1203, ago. 2012.

JUNG, Y. G.; KANG, M. S.; HEO, J. Clustering performance comparison using K-means and expectation maximization algorithms. **Biotechnology and Biotechnological Equipment**, v. 28, n. 1, p. S44–S48, 2014.

KASTORINI, C. M. et al. Comparative analysis of a-priori and a-posteriori dietary patterns using state-of-the-art classification algorithms: A case/case-control study. **Artificial Intelligence in Medicine**, v. 59, n. 3, p. 175–183, 2013.

KELDERS, S. M. et al. Comparing human and automated support for depression: Fractional factorial randomized controlled trial. **Behaviour Research and Therapy**, v. 72, p. 72–80, 2015.

KELLY, J. T. et al. Telehealth methods to deliver dietary interventions in adults with chronic disease: a systematic review and meta-analysis. **The American Journal of Clinical Nutrition**, v. 104, n. 6, p. 1693–1702, 1 dez. 2016.

KESSE-GUYOT, E. et al. Dietary patterns and their sociodemographic and behavioural correlates in French middle-aged adults from the SU.VI.MAX cohort. **European Journal of Clinical Nutrition**, v. 63, n. 4, p. 521–528, 23 abr. 2009.

KHAN, A. S.; HOFFMANN, A. An advanced artificial intelligence tool for menu design. **Nutr Health**, v. 17, n. 1, p. 43–53, 2003.

KRAVCHYCHYN, A. et al. Uso de Recursos de Informática e Mídias Sociais para Avaliação do Consumo Alimentar. In: RIBEIRO, S. M. L.; MELO, C. M. DE; TIRAPÉGUI, J. (Eds.). . **Avaliação nutricional: teoria e prática**. 2º ed. Rio de Janeiro, RJ: Guanabara Koogan, 2018. p. 33–37.

KREBS-SMITH, S. M. et al. Update of the Healthy Eating Index: HEI-2015. **Journal of the Academy of Nutrition and Dietetics**, v. 118, n. 9, p. 1591–1602, set. 2018.

KRIEGER, J. P. et al. Dietary patterns and their sociodemographic and lifestyle determinants in switzerland: Results from the national nutrition survey menuCH. **Nutrients**, v. 11, n. 1, p. 1–16, 2019.

LANDIS, J.; GG, K. The measurement of observer agreement for categorical data.

**Biometrics**, v. 33, n. 1, p. 159–174, 1977.

LIBERATO, S. C.; BRESSAN, J.; HILLS, A. P. Assessment of energy and macronutrient intake in young men: a comparison of 4-day food record and 24-hour dietary recall. **Revista De Nutricao-Brazilian Journal of Nutrition**, v. 22, n. 5, p. 621–630, 2009.

LOVEGROVE, J. A.; GITAU, R. Personalized nutrition for the prevention of cardiovascular disease: a future perspective. **Journal of Human Nutrition and Dietetics**, v. 21, n. 4, p. 306–316, ago. 2008.

MALTA, D. C. et al. Fatores de risco e proteção para doenças crônicas por inquérito telefônico nas capitais brasileiras, Vigitel 2014. **Revista Brasileira de Epidemiologia**, v. 18, n. suppl 2, p. 238–255, dez. 2015.

MAYÉN, A.-L. et al. Socioeconomic determinants of dietary patterns in low- and middle-income countries: a systematic review. **Am J Clin Nutr**, v. 100, p. 1520–31, 2014.

MAYÉN, A.-L. et al. Socioeconomic Differences in Dietary Patterns in an East African Country: Evidence from the Republic of Seychelles. **PLOS ONE**, v. 11, n. 5, p. e0155617, 23 maio 2016.

MCCULLOUGH, M. L. Dietary assessment in the digital age: the ongoing quest for better methods. **The American Journal of Clinical Nutrition**, v. 107, n. 1, p. 1–2, 1 jan. 2018.

MDS. **Ministério do Desenvolvimento Social e Combate à Fome. Marco de Referência de Educação Alimentar e Nutricional para as Políticas Públicas.** Brasília, 2012.

MEHTA, N.; PANDIT, A. Concurrence of big data analytics and healthcare: A systematic review. **International Journal of Medical Informatics**, v. 114, n. January, p. 57–65, 2018.

MOLINA, M. D. C. B. et al. Reprodutibilidade e validade relativa do Questionário de Frequência Alimentar do ELSA-Brasil. **Cadernos de Saúde Pública**, v. 29, n. 2, p. 379–389, fev. 2013a.

MOLINA, M. DEL C. B. et al. Diet assessment in the Brazilian Longitudinal Study of

Adult Health ( ELSA- Brasil ): Development of a food frequency questionnaire  
Avaliação da dieta no Estudo Longitudinal desenvolvimento do Questionário de  
Frequência Alimentar. v. 26, n. 2, p. 167–176, 2013b.

MOLINA, M. DEL C. B. et al. Diet assessment in the Brazilian Longitudinal Study of  
Adult Health (ELSA-Brasil): Development of a food frequency questionnaire. **Revista  
de Nutrição**, v. 26, n. 2, p. 167–176, abr. 2013c.

MOZAFFARIAN, D.; ROSENBERG, I.; UAUY, R. History of modern nutrition science—  
implications for current research, dietary guidelines, and food policy. **BMJ**, v. 361, p.  
k2392, 13 jun. 2018.

MS. **Ministério da Saúde. Portaria nº710/GM/MS, de 10 de junho de 1999. Institui  
a Política Nacional de Alimentação e Nutrição.**Brasília, 1999.

MS. **Ministério da Saúde. Portaria nº2.715/GM/MS, de 17 de novembro de 2011.  
Atualiza a Política Nacional de Alimentação e Nutrição.**Brasília, 2011.

MS. **Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de  
Atenção Básica. Guia Alimentar para a população brasileira.**Brasília, 2014.

MS. **Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde.  
Departamento de Análise em Saúde e Vigilância de Doenças Não  
Transmissíveis. Vigitel Brasil 2019: vigilância de fatores de risco e proteção para  
doenças crônicas por inquérito telefônico.** 1º ed. Brasília, DF: Brasil. Ministério da  
Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise em Saúde e  
Vigilância de Doenças Não Transmissíveis., 2020.

MUGA, M. A. et al. Association between Dietary Patterns and Cardiovascular Risk  
Factors among Middle-Aged and Elderly Adults in Taiwan: A Population-Based Study  
from 2003 to 2012. **PLOS ONE**, v. 11, n. 7, p. 1–18, 1 jul. 2016.

MUSEN, M. A.; MIDDLETON, B.; GREENES, R. A. Clinical Decision-Support  
Systems. In: **Biomedical Informatics**. London: Springer London, 2014. p. 643–674.

MUSTAQEEM, A.; ANWAR, S. M.; MAJID, M. A modular cluster based collaborative  
recommender system for cardiac patients. **Artificial Intelligence in Medicine**, v. 102,  
n. September 2018, p. 101761, 2020.

NASCIMENTO, F. A. DO; SILVA, S. A. DA; JAIME, P. C. Cobertura da avaliação do

consumo alimentar no Sistema de Vigilância Alimentar e Nutricional Brasileiro: 2008 a 2013. **Revista Brasileira de Epidemiologia**, v. 22, p. e190028, 1 abr. 2019.

NAVARRO SILVERA, S. A. et al. Diet and lifestyle factors and risk of subtypes of esophageal and gastric cancers: classification tree analysis. **Annals of epidemiology**, v. 24, n. 1, p. 50–7, jan. 2014.

NCC. **National Cancer Center. Nutrition data system for research software.** Minneapolis, 2010.

NEALE, E. P.; TAPSELL, L. C. Perspective: The Evidence-Based Framework in Nutrition and Dietetics: Implementation, Challenges, and Future Directions. **Adv Nutr**, v. 10, n. 5, p. 1–8, 2019.

NEAPOLITAN, R. **Learning Bayesian Networks**. 1° ed. [s.l.] River, Upper Saddle, 2004.

NEUHOUSER, M. L. et al. A brief dietary assessment instrument for assessing target foods, nutrients and eating patterns. **Public Health Nutrition**, v. 4, n. 1, p. 73–78, 2 fev. 2001.

NEWBY, P. K.; TUCKER, K. L. Empirically Derived Eating Patterns Using Factor or Cluster Analysis: A Review. **Nutr Rev.**, v. 62, n. 5, p. 177–203, 2004.

NOROUZI, S. et al. A mobile application for managing diabetic patients' nutrition: A food recommender system. **Archives of Iranian Medicine**, v. 21, n. 10, p. 466–472, 2018.

NOUH, R. M. et al. A Smart Recommender Based on Hybrid Learning Methods for Personal Well-Being Services. **Sensors**, v. 19, n. 431, p. 1–21, 2019.

OLSON, D. L.; DELEN, D. **Advanced Data Mining Techniques**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

PANARETOS, D. et al. **Repeatability of dietary patterns extracted through multivariate statistical methods: a literature review in methodological issues** **International Journal of Food Sciences and Nutrition**, 2017.

PANARETOS, D. et al. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk



(2002-2012): The ATTICA study. **British Journal of Nutrition**, v. 120, n. 3, p. 326–334, 2018.

PARR, C. L. et al. Comparing methods for handling missing values in food-frequency questionnaires and proposing k nearest neighbours imputation: Effects on dietary intake in the Norwegian Women and Cancer study (NOWAC). **Public Health Nutrition**, v. 11, n. 4, p. 361–370, 2008.

PASHAZADEH, A.; NAVIMIPOUR, N. J. Big data handling mechanisms in the healthcare applications: A comprehensive and systematic literature review. **Journal of Biomedical Informatics**, v. 82, n. April, p. 47–62, jun. 2018.

PENCINA, M. J. et al. Performance of a Method for identifying the Unique Dietary Patterns of Adult Women and Men: The Framingham Nutrition Studies. **J Am Diet Assoc**, v. 108, n. 9, p. 1453–1460, 2009.

PEREIRA, R. A.; SICHIERI, R. Métodos de Avaliação do Consumo de Alimentos. In: KAC, G.; SICHIERI, R.; GIGANTE, D. P. (Eds.). . **Epidemiologia nutricional**. 1° ed. ed. Rio de Janeiro, RJ: Editora Fiocruz / Atheneu, 2007. p. 181–200.

PÉREZ RODRIGO, C. et al. Screeners and brief assessment methods. **Nutricion hospitalaria**, v. 31 Suppl 3, p. 91–8, 26 fev. 2015.

PORTUGAL, I.; ALENCAR, P.; COWAN, D. **The use of machine learning algorithms in recommender systems: A systematic review** *Expert Systems with Applications*, 2018.

PRENTICE, R. L.; HUANG, Y. Nutritional epidemiology methods and related statistical challenges and opportunities. **Statistical Theory and Related Fields**, v. 2, n. 1, p. 2–10, 2 jan. 2018.

RAJKOMAR, A.; DEAN, J.; KOHANE, I. Machine Learning in Medicine. **The New England journal of medicine**, v. 380, n. 14, p. 1347–1358, 2019.

REDDY, C. K.; VINZAMURI, B. A Survey of Partitional and Hierarchical Clustering Algorithms. In: AGGARWAL, C. C.; REDDY, C. K. (Eds.). . **Data Clustering. Algorithms and Applications**. 1° ed ed. [s.l.] CRC Press Taylor & Francis Group, 2014. p. 8–106.

REDDY, K. S.; KATAN, M. B. Diet, nutrition and the prevention of hypertension and

cardiovascular diseases. **Public Health Nutrition**, v. 7, n. 1a, p. 167–186, 2 fev. 2004.

RICCI, F. et al. **Recommender Systems Handbook**. Boston, MA: Springer US, 2011.

RICHTER, A. N.; KHOSHGOFTAAR, T. M. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. **Artificial Intelligence in Medicine**, n. January 2017, p. 0–1, 2018.

SABERIOON, M. et al. Comparative performance analysis of support vector machine, random forest, logistic regression and k-nearest neighbours in rainbow trout (*Oncorhynchus mykiss*) classification using image-based features. **Sensors (Switzerland)**, v. 18, n. 4, p. 1–15, 2018.

SÁNCHEZ-VILLEGAS, A. et al. Gender, age, socio-demographic and lifestyle factors associated with major dietary patterns in the Spanish Project SUN (Seguimiento Universidad de Navarra). **European Journal of Clinical Nutrition**, v. 57, n. 2, p. 285–292, 6 fev. 2003.

SASE, A. et al. A Proposed Book Recommender System. **IJARCCCE**, v. 4, n. 2, p. 481–483, 2015.

SATIJA, A. et al. Understanding nutritional epidemiology and its role in policy. **Advances in nutrition (Bethesda, Md.)**, v. 6, n. 1, p. 5–18, jan. 2015.

SATIJA, A.; HU, F. B. Big data and systematic reviews in nutritional epidemiology. **Nutrition Reviews**, v. 72, n. 12, p. 737–740, 2014.

SCAGLIUSI, F. B. Elementos para anamnese: abordagem de aspectos objetivos e subjetivos. In: RIBEIRO, S. M. L.; MELO, C. M. DE; TIRAPÉGUI, J. (Eds.). . **Avaliação nutricional: teoria e prática**. 2º ed. Rio de Janeiro, RJ: Guanabara Koogan, 2018. p. 340.

SCAGLIUSI, F. B.; LANCHI, A. H. **Subnotificação da ingestão energética na avaliação do consumo alimentar** *Revista de Nutricao*, 2003.

SCHMIDT, M. I. et al. Cohort profile: Longitudinal study of adult health (ELSA-Brasil). **International Journal of Epidemiology**, v. 44, n. 1, p. 68–75, 2015.

SHIM, J.-S.; OH, K.; KIM, H. C. Dietary assessment methods in epidemiologic studies. **Epidemiology and health**, v. 36, p. e2014009, 2014.

SHIOKAWA, Y. et al. Application of Market Basket Analysis for the Visualization of Transaction Data Based on Human Lifestyle and Spectroscopic Measurements. **Analytical Chemistry**, v. 88, n. 5, p. 2714–2719, 2016.

SIDEY-GIBBONS, J. A. M.; SIDEY-GIBBONS, C. J. Machine learning in medicine: a practical introduction. **BMC medical research methodology**, v. 19, n. 1, p. 64, 19 mar. 2019.

SILVA, J. K. et al. Alimentação e cultura como campo científico no Brasil. **PHYSIS: Rev. Saúde Coletiva**, v. 20, n. 2, p. 413–442, 2010.

SOUZA, A. D. M. et al. Avaliação dos marcadores de consumo alimentar do VIGITEL (2007-2009). **Revista Brasileira de Epidemiologia**, v. 14, n. 1, p. 44–52, 2011.

SPICER, J.; SANBORN, A. N. What does the mind learn? A comparison of human and machine learning representations. **Current Opinion in Neurobiology**, v. 55, p. 97–102, 2019.

SUBAR, A. F. et al. Addressing Current Criticism Regarding the Value of Self-Report Dietary Data. **The Journal of Nutrition**, v. 145, n. 12, p. 2639–2645, 1 dez. 2015.

THOMAS, A.; SUJATHA, A. . Comparative Study of Recommender systems. **2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)**, v. March, p. 1–6, 2016.

THOMPSON, F. E. et al. Comparison of Interviewer-Administered and Automated Self-Administered 24-Hour Dietary Recalls in 3 Diverse Integrated Health Systems. **American Journal of Epidemiology**, v. 181, n. 12, p. 970–978, 15 jun. 2015.

THORPE, M. G. et al. A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians. **International Journal of Behavioral Nutrition and Physical Activity**, v. 13, n. 1, 2016.

TIMON, C. M. et al. A review of the design and validation of web- and computer-based 24-h dietary recall tools. **Nutrition Research Reviews**, v. 29, n. 2, p. 268–280, 13 dez. 2016.

TOOZE, J. A. et al. A New Statistical Method for Estimating the Usual Intake of Episodically Consumed Foods with Application to Their Distribution. **Journal of the American Dietetic Association**, v. 106, n. 10, p. 1575–1587, out. 2006.

TRAILL, W. B.; CHAMBERS, S. A.; BUTLER, L. Attitudinal and demographic determinants of diet quality and implications for policy targeting. **Journal of Human Nutrition and Dietetics**, v. 25, n. 1, p. 87–94, fev. 2012.

TSOUKALAS, A.; ALBERTSON, T.; TAGKOPOULOS, I. From Data to Optimal Decision Making: A Data-Driven, Probabilistic Machine Learning Approach to Decision Support for Patients With Sepsis. **JMIR Med Inform**, v. 3, n. 1, p. 1–15, 2015.

UPPU, S.; KRISHNA, A.; GOPALAN, R. A review on methods for detecting SNP interactions in high-dimensional genomic data. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v. 5963, n. c, p. 1–1, 2017.

VASCONCELOS, F. DE A. G. DE. Tendências históricas dos estudos dietéticos no Brasil. **História, Ciências, Saúde-Manguinhos**, v. 14, n. 1, p. 197–219, 2007.

VERLY-JR, E. et al. Precision of Usual Food Intake Estimates According to the Percentage of Individuals with a Second Dietary Measurement. **Journal of the Academy of Nutrition and Dietetics**, v. 112, n. 7, p. 1015–1020, jul. 2012.

VILLEGAS, R. et al. Dietary patterns in middle-aged Irish men and women defined by cluster analysis. **Public Health Nutrition**, v. 7, n. 8, p. 1017–1024, 2 dez. 2004.

WEBER, I.; ACHANANUPARP, P. Insights from Machine-Learned Diet Success Prediction. **Pacific Symposium on Biocomputing 2016**, v. 2016, p. 540–551, 2016.

WHO. **Adherence to long-term therapies: evidence for action**. Geneva, Switzerland: [s.n.].

WHO. WORLD HEALTH STATISTICS - MONITORING HEALTH FOR THE SDGs. **World Health Organization**, v. 58, n. 12, p. 1.121, 7 nov. 2016.

WHO. **Sustainable healthy diets. Guiding Principles**. Rome: FAO and WHO, 2019.

WIESNER, M.; PFEIFER, D. Health Recommender Systems: Concepts, Requirements, Technical Basics and Challenges. **International Journal of Environmental Research and Public Health**, v. 11, n. 3, p. 2580–2607, 3 mar. 2014.

WILLET, W. **Nutritional epidemiology**. 2 ed ed. New York: Oxford University Press, 1998a.

WILLET, W. Correction for the effects of measurement error. In: WILLET, W. (Ed.). .

**Nutritional Epidemiology**. 2nd. ed. New York, NY, USA: Oxford University Press, 1998b. p. 74–147.

WINKVIST, A. et al. More distinct food intake patterns among women than men in northern Sweden: a population-based survey. **Nutrition Journal**, v. 8, n. 1, p. 12, 19 dez. 2009.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining practical machine learning tools and techniques**. 3° ed. United States: Elsevier, 2011.

WONG, A. et al. **Natural Language Processing and Its Implications for the Future of Medication Safety: A Narrative Review of Recent Advances and Challenges** *Pharmacotherapy*, 9 jun. 2018.

XU, R.; WUNSCH, D. C. Clustering algorithms in biomedical research: A review. **IEEE Reviews in Biomedical Engineering**, v. 3, p. 120–154, 2010.

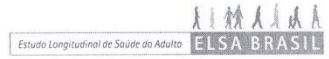
ZHANG, W. et al. “snap-n-Eat”: Food recognition and nutrition estimation on a smartphone. **Journal of Diabetes Science and Technology**, v. 9, n. 3, p. 525–533, 2015.

ZHU, F. et al. Segmentation assisted food classification for dietary assessment. **Proc SPIE Int Soc Opt Eng**, p. 78730B, 10 fev. 2011.

## **ANEXO 1 – Questionário de Frequência Alimentar**

ID NUMERO:

--	--	--	--	--	--	--	--	--	--

Código Formulário: QF2  
Versão: 09/07/2009

# Questionário ELSA

## Fase 2

(09.07.2009)

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

Código Formulário: DIE  
 Versão: 09/07/2009



Informações Administrativas: 0a. Data da entrevista: / /  Ob. Nº Entrevistador(a):

**DIETA (DIE)**

<p><b>"Agora vamos falar sobre a sua alimentação habitual dos últimos 12 meses. Gostaríamos de saber o que o(a) Sr(a) come e bebe por dia, por semana ou por mês, como está nesse cartão. [Apresente o cartão DIE 01]</b></p> <p><b>Vou ler alimento por alimento. Diga quais o(a) Sr(a) come ou bebe e em que quantidade.</b></p> <p><b>Para auxiliar na quantificação dos alimentos e bebidas, vamos utilizar esses utensílios. [Apresente os utensílios].</b></p> <p><b>Podemos começar?"</b></p>											
<p><b>"Vou iniciar listando os alimentos do GRUPO dos PÃES, CEREAIS E TUBÉRCULOS. Por favor, refira sobre seu consumo habitual dos últimos 12 meses"</b></p>											
<p><b>"Com que frequência o(a) Sr(a) come ou bebe [diga o nome do alimento]?"</b>. Se não especificar frequência, pergunte: <b>"Quantas vezes por dia, semana ou mês?"</b>. <b>"E quantas [diga a medida caseira correspondente, mostrando o utensílio] o(a) Sr(a) come ou bebe?"</b>. Repita essas instruções para todos os alimentos.</p>											
	Alimento	Quantidade consumida por vez	Mais de 3x/dia	2 a 3x/dia	1x/dia	5 a 6x semana	2 a 4x semana	1x semana	1 a 3x/mês	Nunca/quase nunca	Referiu consumo sazonal
1.	Arroz ( ) Integral ( ) Branco	Colher de servir									
2.	Aveia/Granola/Farelos/Outros cereais	Colher sopa cheia									
3.	Farofa/Cuscuz salgado/Cuscuz paulista	Colher sopa cheia									
4.	Farinha de Mandioca/Farinha de Milho	Colher sopa cheia									
5.	Pão light (branco ou integral)	Fatia (25g)									



ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

	Alimento	Quantidade consumida por vez	Mais de 3x/dia	2 a 3x/dia	1x/dia	5 a 6x semana	2 a 4x semana	1x semana	1 a 3x/mês	Nunca/quase nunca	Referiu consumo sazonal
6.	Pão francês/pão de Forma/ Pão sírio/Pão torrado	_____ Unidade (50g)									
7.	Pão doce/Pão Caseiro	_____ Unidade média									
8.	Pão Integral/ Centeio	_____ Fatia (30g)									
9.	Pão de queijo	_____ Unidade média									
10.	Bolo simples (sem recheio)	_____ Fatia média									
11.	Bolo recheado/ Torta/Pavê [Cuca]	_____ Fatia média									
12.	Biscoito salgado (tipo água e sal e outros)	_____ Unidade									
13.	Biscoito doce ( ) com recheio ( ) sem recheio	_____ Unidade									
14.	Polenta/Angu/Pirão	_____ Colher de servir									
15.	Batata inglesa cozida/ Batata ensopada/purê	_____ Colher sopa cheia									
16.	Mandioca [Aipim] /Inhame/Cará, Banana da terra cozida/Batata doce cozida	_____ Pedação médio									
17.	Batata frita/Mandioca frita/Banana frita/Polenta frita/batata doce frita	_____ Tigela									

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

**"Agora vou listar os alimentos do GRUPO das FRUTAS. Por favor, refira sobre seu consumo habitual dos últimos 12 meses, excluindo suco de frutas, frutas secas e em calda."**

	Alimento	Quantidade consumida por vez	Mais de 3x/dia	2 a 3x/dia	1x/dia	5 a 6x semana	2 a 4x semana	1x semana	1 a 3x/mês	Nunca/quase nunca	Referiu consumo sazonal
18.	Laranja/Mexerica/Tangerina/Pokan [Bergamota]	_____ Unidade média									
19.	Banana	_____ Unidade média									
20.	Mamão/Papaia	_____ Unidade média									
21.	Maçã/Pêra	_____ Unidade média									
22.	Melancia	_____ Fatia média									
23.	Melão	_____ Fatia média									
24.	Abacaxi	_____ Fatia média									
25.	Abacate	_____ Unidade média									
26.	Manga	_____ Fatia média									
27.	Uva	_____ Unidade									
28.	Goiaba	_____ Unidade média									

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

	Alimento	Quantidade consumida por vez	Mais de 3x/dia	2 a 3x/dia	1x/dia	5 a 6x semana	2 a 4x semana	1x semana	1 a 3x/mês	Nunca/quase nunca	Referiu consumo sazonal
29.	Morango	_____ Unidade									
30.	Pêssego/Ameixa/Kiwi/Caju/Cajá/Nectarina	_____ Unidade média									
31.	Caqui/Jaca/Pinha/Fruta do conde	_____ Unidade média									
32.	Salada de frutas ( ) com açúcar ou complementos ( ) sem açúcar ou complementos	_____ Tigela									
<b>"Agora vou listar os alimentos do GRUPO das VERDURAS, LEGUMES e LEGUMINOSAS. Por favor, refira sobre seu consumo habitual dos últimos 12 meses"</b>											
33.	Alface	_____ Pegador cheio									
34.	Couve/espinafre refogado	_____ Colher sopa cheia									
35.	Repolho	_____ Pegador cheio									
36.	Chicória/Agrião/Rúcula/Couve crua/Almeirão/Escarola/Acelga crua/Espinafre cru	_____ Pegador cheio									
37.	Tomate	_____ Rodela média									
38.	Abóbora [moranga]	_____ Colher sopa cheia									

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

Código Formulário: DIE  
Versão: 09/07/2009



	Alimento	Quantidade consumida por vez	Mais de 3x/dia	2 a 3x/dia	1x/dia	5 a 6x semana	2 a 4x semana	1x semana	1 a 3x /mês	Nunca/quase nunca	Referiu consumo sazonal
39.	Abobrinha (italiana)/Chuchu/Berinjela	Colher sopa cheia									
40.	Vagem	Colher sopa cheia									
41.	Quiabo	Colher sopa cheia									
42.	Cebola										
43.	Alho	<b>Anote só a frequência</b>									
44.	Cenoura	Colher sopa cheia									
45.	Beterraba	Rodela média									
46.	Couve-flor	Ramo médio									
47.	Brócolis	Ramo médio									
48.	Milho Verde	Colher sopa cheia									
49.	Feijão (preto, vermelho, branco, de corda, etc)	Concha Cheia									
50.	Feijoada/Feijão tropeiro	Concha Cheia									
51.	Lentilha/Grão de bico/Ervilha	Concha Cheia									

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

	Alimento	Quantidade consumida por vez	Mais de 3x/dia	2 a 3x/dia	1x/dia	5 a 6x semana	2 a 4x semana	1x semana	1 a 3x/mês	Nunca/quase nunca	Referiu consumo sazonal
52.	Nozes/castanha de caju/castanha do Pará/Amendoim/Amêndoas/Pistache	_____									
		Punhado									
<b>“Agora vou listar os alimentos do GRUPO dos OVOS, CARNES, LEITE E DERIVADOS. Por favor, refira sobre seu consumo habitual dos últimos 12 meses”</b>											
53.	Ovo cozido/Pochê	_____									
		Unidade									
54.	Ovo frito/ omelete / mexido	_____									
		Unidade									
55.	Maionese	_____									
		( ) light ( ) comum									
56.	Leite	_____									
		( ) desnatado ( ) semi-desnatado ( ) integral ( ) de soja									
57.	Iogurte	_____									
		( ) light ( ) normal									
58.	Queijos Brancos (Minas fresca/Ricota/Cottage/muçarela de búfala)	_____									
		Fatia média									
59.	Queijos Amarelos (Minas padrão/Muçarela/Prato/Cheddar/Canastra processado tipo polenghi, etc.)	_____									
		Fatia média									
60.	Requeijão	_____									
		( ) light ( ) normal									
		Colher sopa cheia									

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

	Alimento	Quantidade consumida por vez	Mais de 3x/dia	2 a 3x/dia	1x/dia	5 a 6x semana	2 a 4x semana	1x semana	1 a 3x/mês	Nunca/quase nunca	Referiu consumo sazonal
61.	Margarina/creme vegetal	_____ Ponta de faca									
62.	Manteiga	_____ Ponta de faca									
63.	Fígado/Miúdos	_____ bife médio									
64.	Bucho/dobradinha	_____ Concha cheia									
65.	Carne de boi com osso (Mocotó/Costela/Rabo)	_____ pedaço médio									
66.	Carne de boi sem osso (bife, carne moída, carne ensopada)	_____ Bife médio									
67.	Carne de porco	_____ pedaço médio									
68.	Peito de frango/Chester/Peru/etc	_____ Filé de peito médio									
69.	Frango Frito (Outras partes)	_____ pedaço médio									
70.	Frango cozido (Outras partes)	_____ pedaço médio									
71.	Lingüiça/ Chouriço [Salsichão]	_____ Unidade									
72.	Hambúrguer (bife)	_____ Unidade média									

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

	Alimento	Quantidade consumida por vez	Mais de 3x/dia	2 a 3x/dia	1x/dia	5 a 6x semana	2 a 4x semana	1x semana	1 a 3x/mês	Nunca/quase nunca	Referiu consumo sazonal
73.	Frios light (blanquet/Peito de peru/Peito de chester)	_____ Fatia média									
74.	Presunto/Mortadela/Copa/Salame/Patê/etc	_____ Fatia média									
75.	Bacon/Toucinho/Torresmo	_____ Fatia média									
76.	Peixe cozido [moqueca capixaba]/peixe assado/ensopado/grelhado	_____ Posta média									
77.	Peixe frito	_____ Filé médio									
78.	Sardinha/Atum	_____ Lata									
79.	Camarão/mariscos	_____ Colher sopa cheia									
80.	Caranguejo/Siri	_____ Unidade média									
<b>“Agora vou listar os alimentos do GRUPO das MASSAS e OUTRAS PREPARAÇÕES. Por favor, refira sobre seu consumo habitual dos últimos 12 meses”</b>											
81.	Pizza	_____ Fatia									
82.	Macarrão (caneloni, lasanha, ravioli, [tortei])	_____ Escumadeira cheia									
83.	Macarrão instantâneo	_____ Pacote									

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

	Alimento	Quantidade consumida por vez	Mais de 3x/dia	1x/dia	5 a 6x semana	2 a 4x semana	1x semana	1 a 3x/mês	Nunca/quase nunca	Referiu consumo sazonal
84.	Salgados assados (Esfira,/Empada)/Empanada/Pastel de forno/etc)	Unidade média								
85.	Salgados fritos (quibe/pastel/coxinha)	Unidade média								
86.	Acarajé	Unidade média								
87.	Cachorro-quente	Unidade média								
88.	Pipoca	Saco médio								
89.	Estrogonofe	Colher de servir								
90.	( ) Vatapá ( ) Caruru ( ) moqueca de peixe	Colher de servir								
91.	( ) sushi, ( ) sashimi, tofu ( ) Yakisoba	Tigela cheia								
92.	Sopa de Legumes	Concha cheia								
93.	Sopa instantânea	Concha cheia								
<b>"Agora vou listar os DOCES. Por favor, refira sobre seu consumo habitual dos últimos 12 meses"</b>										



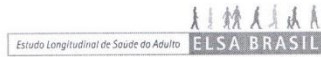
ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

	Alimento	Quantidade consumida por vez	Mais de 3x/dia	2 a 3x/dia	1x/dia	5 a 6x semana	2 a 4x semana	1x semana	1 a 3x/mês	Nunca/quase nunca	Referiu consumo sazonal
94.	Sorvete cremoso	_____ Bola média									
95.	Picolé de frutas	_____ Unidade									
96.	Caramelo/Bala	_____ Unidade									
97.	Gelatina	_____ Tigela									
98.	Chocolate em pó/ Achocolatado em pó/Capuccino	_____ Colher de sobremesa cheia									
99.	Chocolate em barra/Bombom, Brigadeiro [Negrinho], Doce de leite/ Docinho de festa	_____ Bombom (20g)									
100.	Pudim/Doce à base de leite/Mousse	_____ Colher sopa cheia									
101.	Doce de Fruta [chimia], Geléia	_____ Colher sopa cheia *									
102.	Mel/Melado	_____ Colher sopa cheia									
103.	Barra de cereais	_____ Unidade									
<b>"Agora vou listar as BEBIDAS. Por favor, refira sobre seu consumo habitual dos últimos 12 meses"</b>											



ID NUMERO:

--	--	--	--	--	--	--	--	--	--

Código Formulário: DIE  
Versão: 09/07/2009**As próximas perguntas se referem aos seus hábitos alimentares****Entrevistador(a):** Continue a utilizar o CARTÃO DIE 01 para referir a frequência

115. Com que frequência o(a) Sr(a) costuma comer no refeitório/restaurante (universitário ou não) no horário de trabalho?

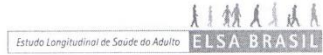
 Mais de 3x/dia 2 a 3x/dia 1x/dia 5 a 6x /semana 2 a 4x/semana 1x/semana 1 a 3x/mês Nunca/quase nunca NÃO SE APLICA NÃO SABE/NÃO QUER RESPONDER

116. Com que frequência o(a) Sr(a) costuma comer em restaurantes fora do horário de trabalho?

 Mais de 3x/dia 2 a 3x/dia 1x/dia 5 a 6x /semana 2 a 4x/semana 1x/semana 1 a 3x/mês Nunca/quase nunca NÃO SE APLICA NÃO SABE/NÃO QUER RESPONDER

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

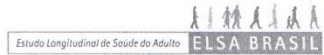
Código Formulário: DIE  
Versão: 09/07/2009



117. Com que frequência o(a) Sr(a) costuma comer <i>fast food</i> (pizza, esfirra, quibe, coxinha, hambúrguer, acarajé)?
<input type="checkbox"/> Mais de 3x/dia
<input type="checkbox"/> 2 a 3x/dia
<input type="checkbox"/> 1x/dia
<input type="checkbox"/> 5 a 6x /semana
<input type="checkbox"/> 2 a 4x/semana
<input type="checkbox"/> 1x/semana
<input type="checkbox"/> 1 a 3x/mês
<input type="checkbox"/> Nunca/quase nunca
<input type="checkbox"/> NÃO SABE/NÃO QUER RESPONDER
118. Com que frequência o(a) Sr(a) costuma comer churrasco?
<input type="checkbox"/> Mais de 3x/dia
<input type="checkbox"/> 2 a 3x/dia
<input type="checkbox"/> 1x/dia
<input type="checkbox"/> 5 a 6x /semana
<input type="checkbox"/> 2 a 4x/semana
<input type="checkbox"/> 1x/semana
<input type="checkbox"/> 1 a 3x/mês
<input type="checkbox"/> Nunca/quase nunca
<input type="checkbox"/> NÃO SABE/NÃO QUER RESPONDER

ID NUMERO:

--	--	--	--	--	--	--	--	--	--

Código Formulário: DIE  
Versão: 09/07/2009

119. Com que frequência o(a) Sr(a) costuma comer VERDURAS e LEGUMES crus, cozidos ou refogados, sem incluir batatas, mandioca/aipim, inhame e cará?

 Mais de 3x/dia 2 a 3x/dia 1x/dia 5 a 6x /semana 2 a 4x/semana 1x/semana 1 a 3x/mês Nunca/quase nunca NÃO SABE/NÃO QUER RESPONDER

120. Com que frequência o(a) Sr(a) costuma comer FRUTAS, sem incluir sucos de frutas?

 Mais de 3x/dia 2 a 3x/dia 1x/dia 5 a 6x /semana 2 a 4x/semana 1x/semana 1 a 3x/mês Nunca/quase nunca NÃO SABE/NÃO QUER RESPONDER**Entrevistador(a):** RECOLHA O CARTÃO DIE01.

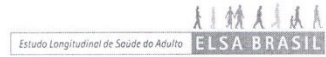
121. No total, quantas refeições o(a) Sr(a) faz por dia, incluindo os lanches e excluindo cafezinho entre os intervalos?

|\_|\_|\_|\_| refeições

 NÃO SABE/NÃO QUER RESPONDER

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

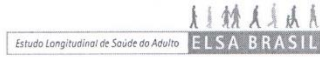
Código Formulário: DIE  
Versão: 09/07/2009



122. Quando o(a) Sr(a) come doces (de todos os tipos citados), qual a forma mais consumida? LEIA AS ALTERNATIVAS	
<input type="checkbox"/>	Normal
<input type="checkbox"/>	Light/diet
<input type="checkbox"/>	Ambas
<input type="checkbox"/>	NÃO COME DOCES
123. O(a) Sr(a) usa adoçante artificial?	
<input type="checkbox"/>	NÃO SABE/NÃO QUER RESPONDER
<input type="checkbox"/>	Não
<input type="checkbox"/>	Sim ----->
	124. Qual o tipo? LEIA AS ALTERNATIVAS. Se for o caso, marque mais de uma resposta.
<input type="checkbox"/>	Líquido (embalagem transparente)
<input type="checkbox"/>	Líquido (embalagem opaca)
<input type="checkbox"/>	Pó
<input type="checkbox"/>	Stévia
<input type="checkbox"/>	Forno e Fogão
<input type="checkbox"/>	NÃO SABE
125. Em sua casa que tipo de ÓLEO/GORDURA é usado no COZIMENTO/PREPARO de refeições? [Marque o óleo de uso mais freqüente].	
<input type="checkbox"/>	óleo de soja
<input type="checkbox"/>	óleo de milho/girassol/canola
<input type="checkbox"/>	azeite de oliva
<input type="checkbox"/>	óleo composto (azeite de oliva + óleo de soja)
<input type="checkbox"/>	azeite de dendê
<input type="checkbox"/>	margarina
<input type="checkbox"/>	manteiga
<input type="checkbox"/>	banha
<input type="checkbox"/>	não se usa óleo/gordura
<input type="checkbox"/>	NÃO SABE

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

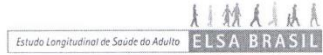
Código Formulário: DIE  
Versão: 09/07/2009



126. Que tipo de ÓLEO/AZEITE o(a) Sr(a) costuma usar em saladas? [Marque o óleo de uso mais freqüente].
<input type="checkbox"/> óleo de soja
<input type="checkbox"/> óleo de milho/girassol/canola
<input type="checkbox"/> azeite de oliva
<input type="checkbox"/> óleo composto (azeite de oliva + óleo de soja)
<input type="checkbox"/> não se usa óleo/azeite
<input type="checkbox"/> NÃO SABE
127. Quando o(a) Sr(a) come CARNE DE BOI/VACA, PORCO ou CARNEIRO, costuma tirar a gordura visível? LEIA AS ALTERNATIVAS.
<input type="checkbox"/> nunca/raramente
<input type="checkbox"/> algumas vezes
<input type="checkbox"/> sempre
<input type="checkbox"/> NÃO COME CARNE
128. Quando o(a) Sr(a) come CARNE DE FRANGO ou OUTRO TIPO DE AVE, costuma tirar a pele? LEIA AS ALTERNATIVAS.
<input type="checkbox"/> nunca/raramente
<input type="checkbox"/> algumas vezes
<input type="checkbox"/> sempre
<input type="checkbox"/> NÃO COME AVES, EM GERAL

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

Código Formulário: DIE  
Versão: 09/07/2009

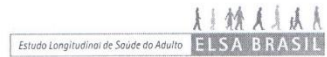


129. NOS ÚLTIMOS SEIS MESES o(a) Sr(a) mudou seus hábitos alimentares ou está fazendo dieta por algum motivo?	
<input type="checkbox"/> NÃO SABE/NÃO QUER RESPONDER	
<input type="checkbox"/> Não	
<input type="checkbox"/> Sim ----->	130. Qual foi o motivo? Se for o caso, escolha mais de uma resposta. LEIA AS ALTERNATIVAS.
	<input type="checkbox"/> para perda de peso
	<input type="checkbox"/> para redução de colesterol
	<input type="checkbox"/> para redução de sal
	<input type="checkbox"/> dieta vegetariana/redução de carnes
	<input type="checkbox"/> para ganho de peso
	<input type="checkbox"/> para diabetes
	<input type="checkbox"/> por outro motivo. Especifique:
<input type="checkbox"/> NÃO SABE/NÃO QUER RESPONDER	



ID NUMERO:

--	--	--	--	--	--	--	--	--	--

Código Formulário: DIE  
Versão: 09/07/2009

131. NOS ÚLTIMOS DOZE MESES o(a) Sr(a) fez uso de algum suplemento de vitaminas e/ou de minerais?

NÃO SABE/NÃO QUER RESPONDER

Não

Sim, regularmente ----->

Sim, mas não regularmente -->

132. Qual? LEIA AS ALTERNATIVAS. Se for o caso marque mais de uma opção

Polivitamínico

Polimineral

Vitamina A

Vitamina B6

Vitamina C

Vitamina D

Vitamina E

Ácido Fólico

Beta Caroteno

Cálcio

Ferro

Selênio

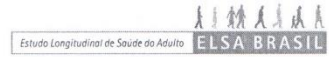
Zinco

Outro:

NÃO SABE/NÃO QUER RESPONDER

ID NUMERO:									
------------	--	--	--	--	--	--	--	--	--

Código Formulário: DIE  
Versão: 09/07/2009



133. O Sr(a) consome café? LEIA AS ALTERNATIVAS.	
<input type="checkbox"/> Não	<b>(PULE PARA A QUESTÃO 01 DO BLOCO COA)</b>
<input type="checkbox"/> NÃO SABE/NÃO QUER RESPONDER	
<input type="checkbox"/> Sim, com cafeína	
<input type="checkbox"/> Sim, descafeinado	
134. O café que o(a) Sr(a) usualmente consome é de que tipo? LEIA AS ALTERNATIVAS.	
<input type="checkbox"/> Passado em filtro ou coador	
<input type="checkbox"/> Espresso	
<input type="checkbox"/> Cafeteira Italiana	
<input type="checkbox"/> Solúvel (instantâneo)	
<input type="checkbox"/> Outro tipo. Especifique: _____	
<input type="checkbox"/> NÃO SABE/NÃO QUER RESPONDER	

**ANEXO 2 – Carta CONEP**

Fls. nº 109  
Rubrica f

**MINISTÉRIO DA SAÚDE**  
Conselho Nacional de Saúde  
Comissão Nacional de Ética em Pesquisa

**CARTA Nº 976 CONEP/CNS/MS**

Brasília, 04 de agosto de 2006.

Senhora Coordenadora,

Tendo a CONEP recebido desse CEP o projeto de pesquisa "*Estudo Longitudinal de Saúde do Adulto – ELSA*" Registro CEP-HU/USP 659/06 - CAAE 0016.1.198.000-06, Registro Sipar MS: nº 25000.083729/2006-38, Registro CONEP nº 13065, verifica-se que:

Trata-se de protocolo a ser desenvolvido por consórcio vencedor da Chamada Pública DECIT/MS/FINEP/CNPq que foi constituído por sete instituições de ensino superior e pesquisa de seis estados, das regiões Nordeste (Universidade Federal da Bahia), Sudeste (FIOCRUZ/RJ, USP, UERJ, UFMG e UFES) e Sul (UFRS). Será um estudo de coorte de 15 mil funcionários de instituições públicas com idade igual ou superior a 35 anos. A coorte será acompanhada anualmente para verificação do estado geral e, a cada três anos, será chamada para avaliações mais detalhadas que incluem exames clínicos. Os sujeitos de pesquisa serão entrevistados por pessoas treinadas e certificadas e os exames serão realizados por profissionais de saúde. O estudo tem como objetivos principais: estimar a incidência do diabetes e das doenças cardiovasculares e estudar sua história natural; investigar associações entre fatores biológicos, comportamentais, ambientais, ocupacionais, psicológicos e sociais relacionados a essas doenças e complicações decorrentes, buscando compor modelo causal que contemple suas inter-relações; descrever a evolução temporal desses fatores e os determinantes dessa evolução; identificar modificadores de efeito das associações observadas; identificar diferenciais nos padrões de risco entre os centros participantes que possam expressar variações regionais relacionadas a essas doenças no país. Dentre os objetivos secundários consta "*estocar material biológico, para estudos futuros com diversos tipos de marcadores relacionados à inflamação, coagulação, disfunção endotelial, resistência à insulina, obesidade central, estresse e fatores de risco tradicionais, bem como prover a extração de DNA para exames genéticos futuros*". De acordo com informação da pág. 11 do protocolo, item "coleta de sangue", as amostras de sangue serão estocadas para

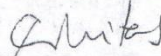
Fls. n.º 110  
 Rubrica F

Cont. Carta CONEP nº 976/2006

exames adicionais e formação de banco de DNA. Haverá um laboratório central que fará as "determinações básicas do estudo em amostras encaminhadas pelos centros de investigação", as "determinações simples" serão feitas nos próprios laboratórios. O banco de material biológico está em fase de planejamento com local e coordenador a serem definidos.

Diante do exposto, embora nos objetivos do estudo verifica-se que haverá também pesquisa genética, pelas informações do protocolo tal pesquisa não será realizada no momento, não estando descrito ainda (nem no protocolo, nem no Termo de Consentimento Livre e Esclarecido-TCLE) os procedimentos para tal. Portanto, nesse primeiro momento do estudo não se trata de projeto da área temática especial "genética humana" (Grupo I), conforme registrado na folha de rosto, mas sim, do grupo III. Nesse caso, a aprovação ética é delegada ao Comitê de Ética em Pesquisa da instituição, devendo ser seguido o procedimento para projetos do grupo III, conforme o fluxograma disponível no site : <http://conselho.saude.gov.br> e no Manual Operacional para CEP. Não cabe, portanto, a referência a CONEP no 3º parágrafo da pág. 1 e no 6º parágrafo da pág.2 do TCLE. Evidenciamos, entretanto, que o armazenamento e utilização de materiais biológicos humanos no âmbito de projetos de pesquisa está regulamentado pela Resolução CNS 347/2005 e que o projeto em questão deve incluir as determinações dessa resolução. Quando for elaborado o protocolo para os estudos genéticos, deverá também ser cumprida a Resolução CNS 340/04 incluindo obtenção de TCLE específico. Em se tratando de pesquisa com funcionários de instituições públicas, cabe ressaltar o disposto no item IV.3 "b" da Res. 196/96.

Atenciosamente ,



**CORINA BONTEMPO DUCA DE FREITAS**  
 Secretária Executiva da  
 COMISSÃO NACIONAL DE ÉTICA EM PESQUISA

À Sua Senhoria

Sr(a) Maria Teresa Zulini da Costa  
 Coordenadora Comitê de Ética em Pesquisas  
 Hospital Universitário da Universidade de São Paulo - HU/USP  
 Av. Profº Lineu Prestes, 2565  
 Cidade Universitária São Paulo  
 Cep:05.508-900

C/ cópia para os CEPs: UFBA, FIOCRUZ/RJ, UERJ, UFMG, UFES e UFRS

**ANEXO 3 – Aprovação CEP HU**



Fls. nº 99  
 Rubrica [assinatura]

São Paulo, 19 de maio de 2006.

Il<sup>mo</sup>(a). Sr<sup>a</sup>(a).

**Prof. Dr. Paulo Andrade Lotufo**  
 Superintendência  
 Hospital Universitário da USP

**Referente:** Projeto de Pesquisa “*Estudo Longitudinal de Saúde do Adulto - ELSA*” –  
 Cadastro CEP-HU: 669/06 - Cadastro SISNEP: FR – 93920 – CAAE – 0016.1.198.000-  
 06 - Área temática especial: Grupo I – I.1. Genética Humana

Prezado(a) Senhor(a)

O Comitê de Ética em Pesquisa do Hospital Universitário da Universidade de São Paulo, em reunião realizada no dia 19 de maio de 2006, analisou o projeto de pesquisa acima citado, considerando-o como **APROVADO**, bem como, seu Termo de Consentimento Livre e Esclarecido. Informamos que **o projeto estará sendo encaminhado para apreciação da Comissão Nacional de Ética em Pesquisa – CONEP- Brasília, devendo ser iniciado o estudo somente após a aprovação da referida Comissão.**

Lembramos que cabe ao pesquisador elaborar e apresentar a este Comitê, relatórios semestrais (e relatório final ao término do trabalho), de acordo com a Resolução do Conselho Nacional de Saúde 251/97, item V.1.c. **O primeiro relatório está previsto para 19 de novembro de 2006.**

Atenciosamente,

**Dra. Maria Teresa Zulini da Costa**  
 Coordenadora  
 Comitê de Ética em Pesquisa – CEP



São Paulo, 2 de julho de 2009.

*II<sup>no</sup>. S.*

*Prof. Dr. Paulo Andrade Lotufo*

Superintendência do Hospital Universitário  
UNIVERSIDADE DE SÃO PAULO

**Nota de Correção:**

Na carta de aprovação do Protocolo de Pesquisa “Estudo Longitudinal de Saúde do Adulto – ELSA”, expedida por este Comitê de Ética em Pesquisa, em 19 de maio de 2006, foi citado, erroneamente, o número de registro CEP-HU/USP 669/06, que deverá ser corrigido para 659/06.

Também havia sido classificado na folha de rosto como projeto da área temática especial – genética humana (Grupo I), **classificação modificada para Grupo III, projeto não pertencente à área temática especial**, de acordo com a carta N<sup>o</sup> 976 CONEP/CNS/MS de 4 de agosto de 2006.

*Dr. Maurício Seckler*  
*Coordenador do Comitê de Ética em Pesquisa*  
*Hospital Universitário da USP*



**ANEXO 4 – Aprovação CEP FSP**

USP - FACULDADE DE SAÚDE  
PÚBLICA DA UNIVERSIDADE  
DE SÃO PAULO - FSP/USP



### PARECER CONSUBSTANCIADO DO CEP

#### DADOS DO PROJETO DE PESQUISA

**Título da Pesquisa:** Predição de qualidade da dieta com machine learning

**Pesquisador:** VANDERLEI CARNEIRO DA SILVA

**Área Temática:**

**Versão:** 1

**CAAE:** 80274517.0.0000.5421

**Instituição Proponente:** Faculdade de Saúde Pública USP/SP

**Patrocinador Principal:** Financiamento Próprio

#### DADOS DO PARECER

**Número do Parecer:** 2.566.286

#### Apresentação do Projeto:

A avaliação do consumo alimentar é uma atividade importante para a saúde pública, porém desafiadora por sua complexidade. O uso de técnicas como análise de cluster e componentes principais são bastante utilizadas para criar padrões, mas não são técnicas que possam ser utilizadas na predição de desfechos. A análise de regressão utilizada em estudos epidemiológicos, permite a identificação de um conjunto de preditores com a probabilidade de um evento ocorrer. No entanto, é limitada ao lidar com relações não lineares frequentemente observadas em dados sobre o consumo alimentar. As técnicas de machine learning, por outro lado, oferecem a oportunidade de modelar problemas complexos com centenas de variáveis preditoras, inclusive na presença de interações entre elas, tal como ocorre quando é avaliado o consumo de alimentos e nutrientes. Machine Learning (ou aprendizado de máquina) estuda a construção de algoritmos com a capacidade de aprender a partir de um conjunto de dados conhecidos gerando modelos preditivos acurados que podem ser aplicados a outros conjuntos de dados não conhecidos

#### Objetivo da Pesquisa:

Geral:

Predição de qualidade da dieta com machine learning em servidores públicos participantes do Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil).

Específicos:

Comparar acurácia dos algoritmos na predição de qualidade da dieta utilizando como referência

**Endereço:** Av. Doutor Arnaldo, 715

**Bairro:** Cerqueira Cesar

**CEP:** 01.246-904

**UF:** SP

**Município:** SAO PAULO

**Telefone:** (11)3061-7779

**Fax:** (11)3061-7779

**E-mail:** coep@fsp.usp.br

USP - FACULDADE DE SAÚDE  
PÚBLICA DA UNIVERSIDADE  
DE SÃO PAULO - FSP/USP



Continuação do Parecer: 2.566.286

o Índice de Qualidade da Dieta Revisado e marcadores do consumo alimentar (consumo de frutas, verduras e legumes [400g/dia]; ingestão de sódio, proporção de calorias na dieta a partir da ingestão de gorduras totais, gordura saturada e bebidas artificiais.

Identificar o número mínimo de variáveis necessárias para predizer com o máximo de eficiência a qualidade da dieta com classificações pelo IQD ou por marcadores do consumo alimentar [alcance ou não das recomendações];

Comparar preditores identificados por machine learning e regressão logística;

**Avaliação dos Riscos e Benefícios:**

No corpo do projeto não há descrição de riscos/benefícios.

Nas informações básicas fornecidas para a plataforma Brasil o autor relata:

1. Nível de risco: baixo; 2. Identificação dos riscos: Vazamento de informações e os indivíduos sentirem algum desconforto com a classificação de uma dieta saudável ou não; 3. Gerenciamento dos riscos: os dados utilizados para análise e publicação dos resultados não irão incluir informações que permitam a identificação dos participantes, tampouco a predição do nível de aderência às recomendações de uma dieta saudável serão utilizadas para estigmatizar os sujeitos ou o grupo de pesquisa. Os dados serão analisados apenas em ambiente restrito (computador pessoal).

**Comentários e Considerações sobre a Pesquisa:**

Devido à ausência de informação sobre os riscos no projeto de pesquisa, não fica evidenciado se o autor reconhece ou não a presença de riscos.

**Considerações sobre os Termos de apresentação obrigatória:**

O pesquisador apresentou a carta de anuência que autoriza a utilização dos dados do Projeto ELSA.

**Recomendações:**

Por se tratar de um projeto que fará a análise de dados secundários, sugiro apenas incluir a questão do risco no projeto de pesquisa.

**Conclusões ou Pendências e Lista de Inadequações:**

Projeto aprovado

**Considerações Finais a critério do CEP:**

Endereço: Av. Doutor Arnaldo, 715  
Bairro: Cerqueira Cesar CEP: 01.246-904  
UF: SP Município: SAO PAULO  
Telefone: (11)3061-7779 Fax: (11)3061-7779 E-mail: coep@fsp.usp.br

USP - FACULDADE DE SAÚDE  
PÚBLICA DA UNIVERSIDADE  
DE SÃO PAULO - FSP/USP



Continuação do Parecer: 2.566.286

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1004373.pdf	10/11/2017 14:57:54		Aceito
Outros	anuencia.jpg	10/11/2017 14:54:34	VANDERLEI CARNEIRO DA SILVA	Aceito
Folha de Rosto	folha.docx	20/10/2017 14:08:39	VANDERLEI CARNEIRO DA SILVA	Aceito
Projeto Detalhado / Brochura Investigador	projetodoutoradoVCS120617.pdf	26/09/2017 16:37:58	VANDERLEI CARNEIRO DA SILVA	Aceito

**Situação do Parecer:**

Aprovado

**Necessita Apreciação da CONEP:**

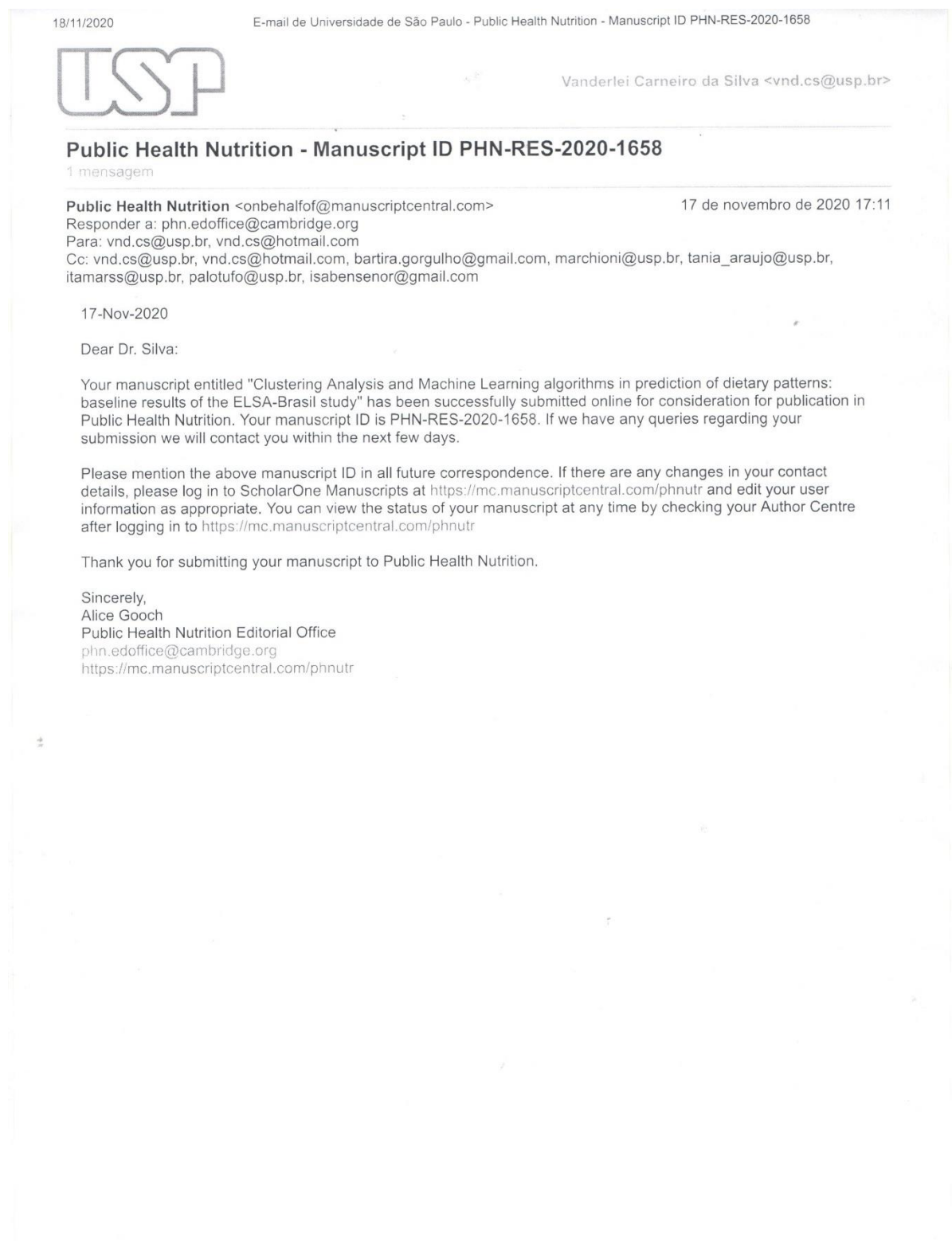
Não

SAO PAULO, 27 de Março de 2018

---


Assinado por:  
Maria Regina Alves Cardoso  
(Coordenador)

**Endereço:** Av. Doutor Arnaldo, 715  
**Bairro:** Cerqueira Cesar **CEP:** 01.246-904  
**UF:** SP **Município:** SAO PAULO  
**Telefone:** (11)3061-7779 **Fax:** (11)3061-7779 **E-mail:** coep@fsp.usp.br

**COMPROVANTE DE SUBMISSÃO – ARTIGO 1**

## COMPROVANTE DE SUBMISSÃO – ARTIGO 2

18/11/2020 E-mail de Universidade de São Paulo - Preventive Medicine: Submission Confirmation



Vanderlei Carneiro da Silva <vnd.cs@usp.br>

---

**Preventive Medicine: Submission Confirmation**  
1 mensagem

---

**PM (ELS)** <em@editorialmanager.com> 17 de novembro de 2020 14:20  
 Responder a: "PM (ELS)" <pm@elsevier.com>  
 Para: Vanderlei Silva <vnd.cs@usp.br>

Title: Recommender System Based on Collaborative Filtering for Personalized Dietary Advice: A Cross-sectional Analysis of the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil)  
 Corresponding Author: Dr. Vanderlei Silva  
 Authors: Bartira Gorgulho; Dirce Marchioni; Tânia Araujo; Itamar Santos; Paulo Lotufo; Isabela Bensenor  
 Research paper

Dear Dr. Silva,

This is to confirm that the above-mentioned manuscript has been received for consideration in Preventive Medicine.

Once your paper is assigned a manuscript number, you will receive an e-mail with this number for your reference.

You will then be able to check on the progress of your manuscript by logging on to the Editorial Manager for Preventive Medicine as an author:  
<https://www.editorialmanager.com/ypmed/>  
 Your username is: vnd.cs

If you need to retrieve password details, please go to: <http://www.editorialmanager.com/ypmed>.

Thank you for submitting your manuscript to Preventive Medicine. Should you have any questions, please feel free to contact our office.

Kind regards,

Editorial Manager  
 Preventive Medicine

\*\*\*\*\*

For further assistance, please visit our customer support site at <http://help.elsevier.com/app/answers/list/p/7923>. Here you can search for solutions on a range of topics, find answers to frequently asked questions and learn more about EM via interactive tutorials. You will also find our 24/7 support contact details should you need any further assistance from one of our customer support representatives.

---

In compliance with data protection regulations, you may request that we remove your personal registration details at any time. (Use the following URL: <https://www.editorialmanager.com/ypmed/login.asp?a=r>). Please contact the publication office if you have any questions.

## CURRÍCULO LATTES DO ALUNO



### Vanderlei Carneiro da Silva

Endereço para acessar este CV: <http://lattes.cnpq.br/3723751735538126>

ID Lattes: 3723751735538126

Última atualização do currículo em 18/10/2019

Graduação em Nutrição (2008), Mestre em Saúde Pública (2017) e Doutorando em Epidemiologia pela USP. Atua como nutricionista da Secretaria de Saúde de Guarulhos desde 2009. Em pesquisa trabalha com avaliação do consumo alimentar, epidemiologia nutricional e modelos preditivos. Possui interesse em temas como: envelhecimento, qualidade da dieta e o uso de machine learning em nutrição. **(Texto informado pelo autor)**

#### Identificação

Nome	Vanderlei Carneiro da Silva
Nome em citações bibliográficas	SILVA, V. C.
Lattes ID	<a href="http://lattes.cnpq.br/3723751735538126">http://lattes.cnpq.br/3723751735538126</a>

#### Endereço

Endereço Profissional	Prefeitura Municipal de Guarulhos, Divisão Técnica de Vigilância Sanitária. Rua Írsis, 300 Gopoúva 07051080 - Guarulhos, SP - Brasil Telefone: (11) 24725087 URL da Homepage: <a href="http://www.guarulhos.sp.gov.br/">http://www.guarulhos.sp.gov.br/</a>
-----------------------	--

#### Formação acadêmica/titulação

2017	Doutorado em andamento em Epidemiologia (Conceito CAPES 5). Universidade de São Paulo, USP, Brasil. Título: Predição de qualidade da dieta com Machine Learning. Orientador:  Isabela Judith Martins Bensenor. Grande área: Ciências da Saúde
2015 - 2017	Grande Área: Ciências da Saúde / Área: Saúde Coletiva / Subárea: Epidemiologia. Mestrado em Saúde Pública (Conceito CAPES 6). Universidade de São Paulo, USP, Brasil. Título: Aposentadoria, alimentação e fatores de risco à saúde no Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil), Ano de Obtenção: 2017. Orientador:  Alexandre Dias Porto Chiavegatto Filho. Grande área: Ciências da Saúde
2013 - 2014	Grande Área: Ciências da Saúde / Área: Saúde Coletiva / Subárea: Epidemiologia. Especialização em Gestão da Política Nacional de Alimentação. (Carga Horária: 420h). Fundação Oswaldo Cruz, FIOCRUZ, Brasil. Título: Qualificação das equipes de saúde para enfrentamento da obesidade infantil. Orientador: Mariana Carvalho Pinheiro.
2009 - 2010	Especialização em Fisiologia do Exercício. (Carga Horária: 400h). Universidade de São Paulo, USP, Brasil. Título: Nutrição, estresse oxidativo e imunidade no paciente com HIV. Orientador: Felipe Fedrizzi Donatto.
2005 - 2008	Graduação em Nutrição. Universidade Nove de Julho, UNINOVE, Brasil. Título: Defeitos do Fechamento do Tubo Neural: evolução do coeficiente de incidência no Brasil. Orientador: Julicristie Machado de Oliveira.

## CURRÍCULO LATTES DA ORIENTADORA



## Isabela Judith Martins Bensor

Bolsista de Produtividade em Pesquisa do CNPq - Nível 1D

Endereço para acessar este CV: <http://lattes.cnpq.br/1661431070098079>

ID Lattes: **1661431070098079**

Última atualização do currículo em 09/11/2020

Possui graduação em Medicina pela Faculdade de Medicina da Universidade de São Paulo (1982), especialização em saúde pública pela Faculdade de Saúde Pública da USP (1994), doutorado pela Faculdade de Medicina da USP (1997), pós-doutorado em Epidemiologia no Brigham and Women's Hospital, Harvard Medical School (1997-99) e livre-docência pela Faculdade de Medicina da USP (2003). Atualmente é professora associada nível 3 da Universidade de São Paulo (USP) no Departamento de Clínica Médica em regime de dedicação exclusiva. É bolsista de produtividade em pesquisa do CNPq nível 1D. Atua em duas linhas de pesquisa: epidemiologia das doenças crônicas e dos sintomas. Vice-coordenadora do "Estudo Longitudinal de Saúde do Adulto" no Centro de Investigação de São Paulo desde 2005. Coordenadora do Centro de Pesquisa Clínica e Epidemiológica do Hospital Universitário da USP desde 2008. Organizou junto com outros professores e pesquisadores cinco livros: "Semiologia Clínica", "Epidemiologia: uma abordagem prática", "Orientação Nutricional", "Medicina em Ambulatório" e "Clínica Médica" junto com com diversos coautores, todos editados pela Sarvier Editora de Livros Médicos. Editora do Brazilian Journal of Medical and Biological Research de 2015 a 2018, Membro da Associação Brasileira de Divulgação Científica (ABDC) 2016-2020; Resumo da produção; 375 artigos publicados (352 indexados na base ISI) sendo 216 nos últimos 4 anos; 9 artigos aceitos para publicação (todos em revistas indexadas na base ISI); 27 orientações concluídas sendo 12 doutorados, 6 mestrados e uma co-orientação de doutorado nos últimos 10 anos; 2 orientações de pós-doutorado, 9 orientações de iniciação científica com bolsa (7 FAPESP e 2 PIBIC); supervisão de 35 bolsistas de longa duração (DTI ou EXP) CNPq nos últimos 10 anos; 2 orientações de mestrado em andamento, 5 orientações de doutorado em andamento, 2 orientações de pós-doutorados em andamento, participações em bancas de trabalhos de conclusão sendo 62 nos último 10 anos; 30 participações em bancas de comissões julgadoras; participou da editoria científica de 3 revistas indexadas na base ISI nos últimos 10 anos; 10 projetos em andamento como pesquisadora principal; parecerista de agências de fomento nacionais (CNPq, FAPESP, UFRN) e internacionais (CYTED); parecerista de várias revistas nacionais e internacionais; 108 resumos publicados sendo 51 em revistas indexadas na base ISI; 62 capítulos de livros escritos, 30 nos últimos 10 anos; É co-responsável por 6 disciplinas de pós-graduação. Participa do Global Burden of Diseases desde 2012. Índice H: 48, 20.109 citações (**Texto informado pelo autor**)

### Identificação

Nome	Isabela Judith Martins Bensor
Nome em citações bibliográficas	BENSENOR, I. J. M.; Benseñor, Isabela M.; Benseñor, Isabela M.; BENSENOR, I.; Benseñor, Isabela Martins; BENSEÑOR, I M.; BENSEÑOR, ISABELA M.; BENSENOR, ISABELA M.; BENSENOR, I. M.; BENSEÑOR, I. M.; Isabela Benseñor; BENSEÑOR, ISABELA; BENSEÑOR, ISABELA MARTINS; BENSEÑOR, I.M.; BENSEÑOR, I.M.; Isabela M. Benseñor; BENSENOR, ISABELA; BENSENOR, IM; BENSEÑOR, I.; BENSEÑOR, ISABELA JUDITH MARTINS; BENSENOR, ISABELA MARTINS; BENSEÑOR, ISABELA M.; BENSEÑOR, ISABELA M.; BENSEÑOR, I.; MARTINS BENSEÑOR, ISABELA; BENSEÑOR, ISABELA J. M.; BENSEÑOR, ISABELA M.; BENSEÑOR, ISABELA J.; BENSEÑOR, I.M.; BENSENOR, ISABELA JUDITH MARTINS; BENSENOR, ISABELA M.; BENSEÑOR, ISABELA JUDITH MARTINS; Benseñor, IJ; BENSENOR, I. J.; Benseñor, I.; Benseñor, Isabella; BENSEÑOR, ISABELA M.J.; BENSEÑOR, ISABELA J. MARTINS; BENSENOR, I.M.; BENSENOR, ISABELA J.M.; M BENSEÑOR ISABELA J. BENSEÑOR, ISABELA; BENSENOR, ISABELA J.; Isabela Judith Bensor; ISABELA JUDITH BENSENOR; BENSENOR, ISABELA JUDITH; BENSEÑOR, ISABELA JUDITH MARTINS; BENSENOR, ISABELA J. MARTINS; BENSEÑOR, ISABELA JUDITH; BENSEÑOR, ISABELA J.M.; Isabela M Bensor
Lattes iD	<a href="http://lattes.cnpq.br/1661431070098079">http://lattes.cnpq.br/1661431070098079</a>
Orcid iD	<a href="https://orcid.org/0000-0002-6723-5678">https://orcid.org/0000-0002-6723-5678</a>