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

**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 predizer 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, predizer 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.