UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

“Uso de Aprendizaje de Máquinas para el Estudio de la Obesidad en Chile”

MARCELA BETSABETH AGUIRRE JEREZ

TESIS PARA OPTAR AL GRADO DE

MAGISTER EN INFORMÁTICA MÉDICA

Directora de Tesis: Prof. Dra. Jocelyn Dunstan
Co-director: Prof. Dr. Steffen Härtel

2019

INDICE

RESUMEN

Problema: Cerca del 74,2% de la población chilena tiene sobrepeso u obesidad, considerada un factor de riesgo para variadas enfermedades no transmisibles, las cuales hoy en día representan gran parte de la mortalidad prematura y Años de Vida Potenciales Perdidos. La obesidad se debe al desbalance entre el consumo y el gasto energético de los individuos, el cual se ve afectado tanto por factores fisiológicos individuales como por la interacción con el medio. Dado que las intervenciones realizadas hasta el momento para controlar la obesidad, por medio de estrategias que apuntan a controlar el consumo de alimentos como causa de ésta, no han producido el efecto esperados por los expertos, es necesario utilizar nuevos modelos de estudio que representen el fenómeno con mayor precisión.

Hipótesis: El estado nutricional de la población chilena puede ser predicho a partir del consumo de alimentos individuales y/o los datos sociodemográficos de la población chilena, con un 90% de exactitud.

Objetivo General: Implementar algoritmos de Aprendizaje de Máquinas de clasificación y regresión para predecir el estado nutricional (IMC o clasificación de estado nutricional) de la población chilena adulta, a partir de datos de consumo alimentario, características sociodemográficas individuales o la combinación de éstos.

Material y Métodos: La Encuesta Nacional de Consumo de Alimentos (ENCA) fue aplicada a una muestra representativa de la población chilena el año 2010, que incluye variables de interés como información antropométrica, consumo de alimentos y nivel socioeconómico, entre otros. Con los datos de dieta disponibles de la ENCA, se implementaron algoritmos de Aprendizaje de Máquinas para predecir el estado nutricional de los encuestados (5 algoritmos de clasificación y 5 de regresión), con y sin variables sociodemográficas como parte de las variables predictoras. Los algoritmos se compararon (ANOVA) según métricas de desempeño para clasificación (Exactitud) y regresión (raíz cuadrada del error cuadrático medio (RMSE)). Se realizó

además análisis exploratorio de los datos, por medio de análisis de correlación, análisis de varianza y análisis de componente principal.

Resultados: En la exploración de los datos, las variables de dieta en general se caracterizan por bajas correlaciones y con varianza explicada repartida en una gran cantidad de componentes. En el caso de los algoritmos de clasificación implementados, la exactitud de promedio varía entre 50,8% y 72,2%. El mejor desempeño para la clasificación se reporta para los datos de consumo mensual no agrupado con variables sociodemográficas y el algoritmo *Support Vector Machine* (Exactitud (%): 72,2 ± 2,6). Para la regresión, el promedio de raíz cuadrada del error cuadrático medio (RMSE) varía entre 5,21 a 6,24, y el mejor desempeño se reporta con los datos de consumo mensual agrupado con variables sociodemográficas y el algoritmo de Regresión Lineal (RMSE: 5,2 ± 0,4). Las pruebas estadísticas de comparación de desempeño entre algoritmos son no significativas para la clasificación y la regresión.

Conclusión: Estos resultados deben ser interpretados con cautela. El desempeño regular de algoritmos, sin diferencias significativas entre ellos, llevan al rechazo de la hipótesis de que el estado nutricional de la población chilena puede ser predicho a partir del consumo de alimentos individuales y/o los datos sociodemográficos de la población chilena, con un 90% de exactitud. La capacidad predictiva de los algoritmos implementados depende de las características de los datos utilizados, los cuales por su naturaleza (autorreportados y transversales) demuestran limitaciones para este tipo de análisis. Al comparar con la literatura, los resultados obtenidos son similares a los reportados con población estadounidense encuestada en la *National Health and Nutrition Examination Survey*, incluso tratándose de una menor población y con diferentes características. Si bien, de los resultados obtenidos no se puede establecer causalidad, se obtienen directrices respecto de qué variables son relevantes para predecir la obesidad, que en conjunto pueden proporcionar información para el diseño de futuras investigaciones con métodos que permitan establecer causalidad.

ABSTRACT

Problem: About 74.2% of the Chilean population is overweight or obese, considered a risk factor for various noncommunicable diseases, that represent a large part of premature mortality and Years of Potential Life Lost. Obesity is due to the imbalance between the consumption and energy expenditure of individuals, which is affected by both individual physiological factors and interaction with the environment. Given that the interventions carried out so far to control obesity, through strategies aimed at controlling the consumption of food as a cause of it, have not produced the effect expected by the experts, it is necessary to use new study models that represent the phenomenon with greater precision.

Hypothesis: The nutritional status of the Chilean population can be predicted from the consumption of individual foods and/or sociodemographic data of the Chilean population, with 90% accuracy.

General Objective: Implement machine learning algorithms of classification and regression to predict the nutritional status (BMI or by categories of nutritional) of the Chilean adult population, based on data on food consumption, individual sociodemographic characteristics or the combination of these.

Material and Methods: The National Survey of Food Consumption (ENCA) was applied to a representative sample of the Chilean population in 2010, which includes variables of interest such as anthropometric information, food consumption and socioeconomic status, among others. With the diet data available from the ENCA, Machine Learning algorithms were implemented to predict the nutritional status of the respondents (5 classification algorithms and 5 regression algorithms), with and without sociodemographic variables as part of the predictor variables. The algorithms were compared (ANOVA) according to performance metrics for classification (Accuracy) and regression (square root of the mean square error (RMSE)). An

exploratory analysis of the data was also carried out, through correlation analysis, analysis of variance and principal component analysis.

Results: In the exploration of the data, the variables of diet in general are characterized by low correlations and with explained variance distributed in a large number of components. In the case of the classification algorithms implemented, the average accuracy varies between 50.8% and 72.2%. The best performance for classification is reported for the non-grouped monthly consumption data with sociodemographic variables and Support Vector Machine algorithm (Accuracy (%): 72.2 ± 2.6). For regression, RMSE varies between 5.21 to 6.24, and the best performance is reported with the grouped monthly consumption with sociodemographic data and Linear Regression (RMSE: 5.2 ± 0.4). The statistical tests of performance comparison between algorithms are not significant for classification and regression.

Conclusion: These results should be interpreted with caution. The regular performance of algorithms, without significant differences between them, lead to the rejection of the hypothesis that the nutritional status of the Chilean population can be predicted from the consumption of individual foods and/or the sociodemographic data of the Chilean population, with a 90% accuracy. The predictive capacity of the algorithms implemented is conditioned by the quality of the data used, which by their nature (self-reported and cross-sectional) demonstrate limitations for this type of analysis. When compared with literature, the results obtained are similar to those reported with the US population surveyed in the National Health and Nutrition Examination Survey, even in the case of a smaller population and with different characteristics. Although it is not possible to establish causality from the results obtained, guidelines are obtained regarding which variables are relevant for predicting obesity, which together can provide information for the design of future research with methods that allow establishing causality.

1. INTRODUCCIÓN

1.1 Sobrepeso y Obesidad

La Organización Mundial de la Salud (OMS) define el sobrepeso y la obesidad como *"la acumulación anormal o excesiva de grasa que puede ser perjudicial para la salud"* [1]. Para adultos, el estado nutricional se caracteriza por medio del Índice de Masa Corporal (IMC), definido como el peso (P) de una persona [kg] dividido por su talla (T) [m] al cuadrado ($IMC = P / T^2$), considerándose sobrepeso un IMC entre 25 y 29,9 y obesidad un $IMC \geq 30$[1]. En niños y adolescentes desde los 5 hasta los 19 años, es necesario considerar la edad y el género para ajustar la categorización del estado nutricional, en cuyo caso el sobrepeso es definido como +1 a +1,9 desviaciones estándar (DE) de la referencia OMS, mientras que la obesidad es +2 DE, ambos medidos con respecto a una distribución de referencia[2], al igual que para los niños menores de 5 años[3].

En los últimos 25 años la obesidad se ha más que doblado, y actualmente alrededor de un 52% de la población mundial tiene sobrepeso u obesidad[1]. De acuerdo a la OMS, la causa principal de la obesidad es un desbalance energético entre la energía consumida y gastada, situación propiciada a nivel mundial debido a cambios en los hábitos alimentarios y de actividad física, que en general son consecuencia de cambios ambientales y sociales asociados al desarrollo y de la falta de políticas de apoyo en diversos sectores[1]. Esta definición, en la cual la causa de la obesidad es el desbalance energético, ha sido utilizada para el estudio de la obesidad tanto individual como poblacional. Sin embargo, se ha cuestionado en los últimos años debido a la ineficacia de este modelo causal para controlar el sobrepeso y obesidad a nivel poblacional[4,5]. Las complicaciones médicas de la obesidad son un desafío de salud pública, siendo considerada factor de riesgo para un número importante de enfermedades no transmisibles (ENT)[6]. Estas son caracterizadas por ser enfermedades de larga duración, de lenta progresión, que no se resuelven espontáneamente y que rara vez logran una curación total[6]. Ejemplo de estas enfermedades son las cardiovasculares, diabetes y algunos cánceres, entre otras[1,6]. Estas enfermedades causan alrededor de 15 millones de muertes en personas entre

30 y 69 años, y más de un 85% de estas muertes “prematuras” suceden en los países de bajos y medianos ingresos[6], lo cual ha producido una situación alarmante en la mayor parte del mundo[7]. La transición nutricional desde un mundo afectado por la desnutrición ha dado paso rápidamente en las últimas décadas a una situación de desbalance nutricional dado por un exceso de macronutrientes[8]. Esto se debe en parte al aumento en la disponibilidad de comida alta en calorías y ultra-procesada, y a la reducción del consumo de varios componentes sanos de nuestra dieta, como legumbres, frutas y vegetales[9].

1.2 Sobrepeso y Obesidad: Situación en Chile

La obesidad representa uno de los principales factores de riesgos de diversas enfermedades (enf. isquémica del corazón, enf. cerebrovasculares, hipertensión, cáncer colorrectal, diabetes, cáncer de endometrio, cáncer de mama, osteoartritis, cáncer de vesícula, cáncer renal, disnea, dolor lumbar, entre otras) en la población chilena[10]. Un 59,3% de las muertes son producidas por enfermedades crónicas no transmisibles[11], siendo la obesidad responsable de un 10% de las muertes y un 14,3% de los Años de Vida Potencial Perdidos[10,12]. Según la última Encuesta Nacional de Salud (ENS) 2016-2017, un 39,8% de la población tiene sobrepeso, un 31,2% es obeso y un 3,2% se encuentra en estado de obesidad mórbida, sumando un 74,2% de prevalencia entre las tres categorías[13]. Esta cifra, al ser comparada con las ENS de períodos anteriores, es considerada un alza importante, ya que al año 2003, la prevalencia de obesidad y sobrepeso era de un 61% y el 2009-2010 era de un 67%, mostrando la prevalencia de obesidad una mayor variación entre 2009-2010 y 2016-2017[14,15]. Se observa además en el informe preliminar ENS 2016-2017, que estas prevalencias de obesidad y sobrepeso aumentan según el grupo etario desde los 20 años en adelante. En cuanto al nivel educacional, las personas con menor nivel educacional (menos de 8 años) tienen una prevalencia de obesidad de 46,6% versus un 33,3% en las personas con 8-12 años de educación y un 29,5% en las con más de 12 años[13].

1.3 Obesidad: Un Problema Complejo

De acuerdo al informe *Foresight Tackling Obesities: Future Choices*[16] preparado por el gobierno británico, enfrentar la obesidad corresponde a un problema de política pública altamente complejo dado que existen una gran cantidad de factores que actúan simultáneamente y a diferentes niveles, tanto individuales como poblacionales, todos llevando a una modificación del balance energético de los individuos. Ejemplo de estos factores son la genética, la fisiología individual, la dieta y la actividad física[4], pero también los determinantes de estos últimos, como la condición socioeconómica, el sistema de producción de alimentos y su mercado, las características de los entornos de trabajo, transporte y educación, entre muchos otros[16]. Diferentes autores han reportado que la prevalencia de obesidad en diferentes países muestra patrones dependientes del género (mayor prevalencia y mayor heterogeneidad en mujeres que en hombres)[17–19]. También se ha descrito que a medida que las personas envejecen, la prevalencia de obesidad y sobrepeso aumenta, alcanzando el *peak* de entre los 50 a 65 años, tendiendo a la disminución posteriormente[19,20]. En cuanto a las variables socioeconómicas y el entorno en el que se desenvuelven las personas, se destaca el cuerpo de evidencia[21–27] que soporta la hipótesis de que la calidad de la dieta está directamente asociado al nivel socioeconómico, considerando que limitaciones en el poder adquisitivo llevan a elecciones de comida densa en energía y pobre en nutrientes, más que en el acceso geográfico a fuentes de alimentos sanos[28,29]. Con datos de geolocalización (referente a la ubicación geográfica donde viven las personas, como área rural/urbana o la macrozona norte/metropolitana/sur de un país), se ha determinado asociación (pero no relación causal) entre factores socioeconómicos (valor de las propiedades, educación e ingresos) y el riesgo de obesidad a nivel de barrios[30]. Otro factor a considerar es el rol del sedentarismo y el ejercicio, ya que da cuenta en parte del gasto energético en el desbalance energético que causa la obesidad. La actividad física es reconocida como el componente de gasto energético diario más variable[4]. Sin embargo, intervenciones que sólo aumentan la actividad física no son efectivas para la pérdida de peso[31]. Según Levine *et al*. la energía que se gasta realizando actividades de la vida diaria es un factor de riesgo de obesidad, donde los sujetos

obesos muestran menor *Non-exercise Activity Thermogenesis* (NEAT) que los de peso normal, el cual es invariante a cambios de peso[32]. En los últimos 50 años, se estima que la actividad física que se realiza en la ocupación de las personas ha disminuido en más de 100 calorías por día[33]. La complejidad de factores explicaría, al menos en parte, las dificultades que en todo el mundo se han encontrado para poder enfrentar el problema de la obesidad de manera efectiva[5,34,35].

En los últimos años, se han propuesto nuevos paradigmas para el estudio de la obesidad. Uno de estos plantea que la patogénesis de la obesidad no es únicamente explicada por el desbalance del gasto y consumo energético, y los factores subyacentes que propician el desequilibrio energético, sino que depende también del llamado *reseteo* del *set point* del peso de los individuos a un valor aumentado, es decir, una vez que se establece la cantidad de adiposidad elevada en el cuerpo, esta es *defendida* biológicamente en los individuos obesos. Se cree que el *reseteo* está relacionado con los sistemas de homeostasis de la energía y la regulación del comportamiento de alimentación y/o el gasto energético, sin embargo, la relación descrita aún se encuentra en estudio[4]. Por otro lado, existe el *Carbohydrate-Insulin Model* (CIM), otro paradigma para el estudio de la obesidad, el cual postula que el aumento en el consumo de carbohidratos procesados con alto contenido glicémico produce cambios hormonales que promueven depósitos de calorías en el tejido adiposo, exacerban el hambre y disminuyen el gasto de energía[5]. Debido a que estos modelos se encuentran en etapas tempranas de estudios, aún requieren un mayor cuerpo de evidencia para ser validados.

1.4 Estudio de la Obesidad y el Consumo de Alimentos

El estudio de la obesidad, debido a su complejidad, se lleva a cabo desde distintos aspectos, como, por ejemplo, mediciones antropométricas y composición corporal, la caracterización de la dieta y el nivel de actividad de las personas, además de los factores sociodemográficos y psicológicos relacionados. Las formas de evaluación del consumo de alimentos son variadas, y responden a diferentes propósitos, como, por ejemplo, estudios clínicos o poblacionales de la obesidad, evaluación de

programas de intervención, o para ser utilizado por los profesionales del área de la salud en terapias individuales; en general, la objetivación de la dieta empuja políticas, cambios organizacionales (e institucionales) y tratamientos para la obesidad[36].

La evidencia del efecto de la dieta sobre la obesidad es controversial, debido a la complejidad de los patrones de comportamiento y consumo de ésta, y la variada disponibilidad de alimentos que existe[37–40]. La investigación de la dieta y la obesidad se ha concentrado en el estudio de nutrientes o comidas individuales; sin embargo, dada su complejidad, en la actualidad la dieta se estudia como una entidad en sí, evaluando su calidad en general en lugar de hacerlo por nutrientes y alimentos por separado[41–43]. Existen estudios que evalúan la calidad de la dieta y encuentran asociación negativa con la obesidad (a mayor calidad de dieta, hay menor riesgo de ser obeso o sobrepeso)[44–48]. Al mismo tiempo, existen otros estudios que no encuentran una asociación significativa[49–51]. Esta variabilidad de resultados se puede deber en parte a la influencia de los instrumentos que se utilizan para objetivar la dieta, a la calidad de esta, o a diferencias metodológicas entre los estudios[52,53].

Existen diferentes instrumentos por medio de los cuales se describe la dieta de las personas, con diferentes propósitos y limitaciones[36,54].

La encuesta recordatoria de 24 horas es una entrevista estructurada realizada por un entrevistador, la que se utiliza para obtener información detallada (tamaño de porciones, preparaciones, horarios, etc.) de los alimentos y bebidas consumidas en las últimas 24 horas. Los datos recopilados por medio de este instrumento pueden ser utilizados para: medir la ingesta total diaria y aspectos particulares de la dieta de una persona, para describir el consumo de una población y para examinar la relación de la dieta con el estado de salud u otras variables. En general este instrumento implica baja carga para los sujetos. Tiene la limitación de no capturar información sobre la variabilidad día a día de la alimentación de las personas, por lo que se requieren múltiples mediciones que representen la variabilidad de los días de la semana, además de depender directamente de la memoria específica de los sujetos[36,54].

El registro de alimentación estimado tiene el propósito de obtener información detallada de las comidas (nombres de marcas, métodos de preparación, tamaño de porciones, horarios y lugares de comida) consumidas en uno o más días, recabada a partir del registro en tiempo real lo que consumen los sujetos durante el periodo que se determine. En general se solicita que se reporten varios días para dar cuenta de la variabilidad diaria de la dieta, buscando representar los días de la semana de forma proporcional, la cantidad de días que se reportan depende de los objetivos del estudio[54]. Al igual que la encuesta recordatoria de 24 horas, los datos recopilados por medio del registro de alimentos pueden ser utilizados para: medir la ingesta total diaria y aspectos particulares de la dieta de una persona, para describir el consumo de una población y para examinar la relación de la dieta con el estado de salud u otras variables. Una de las limitaciones de este método son los costos para digitar y procesar este tipo de información, además de los posibles errores de cuantificación de porciones que pueden cometer los sujetos[54]. Existen también los registros de alimentación con registro de masa y son el método más preciso para la estimación del consumo de nutrientes y alimentos a nivel individual. También deben ser medidos más de un día, con el fin de caracterizar la variabilidad del consumo de alimentos usual[54]. Los sujetos pueden tender a modificar su patrón de consumo típico para simplificar el proceso de medición o para impresionar al investigador. En este tipo de instrumentos puede haber *underreporting* (reportar menos consumo que en la realidad)[54].

Los cuestionarios de frecuencia de consumo tienen el propósito de obtener la frecuencia de consumo (diaria, semanal, mensual o anual, dependiendo del objetivo del estudio) y, en algunos casos, las porciones que consumen, durante un periodo de tiempo establecido de un número finito de alimentos y bebidas (un cuestionario completo puede tener más de 100 ítems). Pueden demorar entre 15 a 30 minutos en completarse e implican menor carga a los sujetos. Los datos que proveen los cuestionarios de frecuencia son utilizados para: proveer información sobre el consumo de los alimentos consultados; si es combinada con bases de datos suplementarias, pueden servir para dar información sobre la ingesta dietética total; también se utilizan en estudios poblacionales, con el fin de calcular el riesgo relativo

de algunas enfermedades en relación al consumo de ciertos alimentos, grupos de alimentos o nutrientes[54]. Debido a que se cuestiona por un número limitado de alimentos, una encuesta de frecuencia para cierta población puede no ser útil para otra población con perfil de consumo diferente. Tiene la desventaja de no incluir información sobre el contexto del consumo de alimentos. Depende de la memoria general (y no específica) de los sujetos[36,54].

En cualquier instrumento de objetivación de la dieta existe error sistemático o sesgo, que es definido como una condición que causa que un resultado se desvié del valor real en una dirección consistente. Existen varios tipos de sesgo, los principales son el sesgo de selección y el sesgo de medición. El sesgo de selección ocurre cuando hay diferencias sistemáticas en la selección de los sujetos y cualquier tipo de medición de dieta puede ser sujeto a sesgo de selección. El sesgo de medición puede ser introducido de varias formas, entre ellas: sesgo por conveniencia social, en la cual los sujetos responden lo que creen se espera de ellos; sesgo de entrevistador, que ocurre cuando los entrevistadores difieren en la forma en la cual obtienen, procesan e interpretan la información; sesgo de recuerdo, ocurre cuando sujetos no recuerdan eventos o experiencias previas de forma precisa, u omiten detalles, influenciados por eventos o experiencias subsecuentes[54].

La selección de qué instrumento utilizar para evaluar la dieta de las personas debe ser en base a un balance entre la mejor validez y factibilidad, considerando que ningún instrumento cuenta con validez de constructo perfecto, y que además el problema del sesgo de la fuente, las personas, es prevalente y sigue siendo investigado[36].

1.5 Métodos Estadísticos y Aprendizaje de Máquinas para la Predicción de la Obesidad y Dieta

Como ha sido descrito previamente, la obesidad se caracteriza por ser multifactorial en sus causas y asociaciones, dentro de las cuales se encuentra la dieta, y los instrumentos para evaluarla tienden a generar datos sesgados. Por otro lado,

métodos estadísticos como la Regresión Lineal y Logística son útiles en la identificación factores de riesgo de algunos problemas de salud, sin embargo, al modelar enfermedades relacionadas los estilos de vida, donde las causas son multifactoriales, estos métodos no son tan exitosos. Esto se debe a que no todas las relaciones entre variables son lineales[35], y en los métodos mencionados (Regresión Logística y Regresión Lineal), la correlación entre las variables predictoras impide la interpretación aislada de los componentes[55].

Debido a los avances tecnológicos desde fines de la década de 1990, que permiten mayor capacidad de cómputo y mayor almacenamiento de datos, se han masificado y adoptado métodos agrupados bajo el término Aprendizaje de Máquinas (AM). Los métodos de AM pueden ser considerados extensiones a modelos estadísticos más tradicionales[56], los cuales incorporan la complejidad de diversos fenómenos. Un ejemplo lo anterior es su aplicación en problemas del área de la salud, ya que son más flexibles para *aprender* propiedades a partir de los datos, prescindiendo de la intervención humana para la selección de variables, y con beneficios adicionales, como incorporar estrategias más robustas para el manejo de datos perdidos o mal registrados[57,58]. El Aprendizaje de Maquinas (AM) - o *Machine Learning* en inglés - consiste en un conjunto de métodos estadísticos paramétricos y no paramétricos en los que se realizan predicciones a partir de los datos, “aprendiendo” de ellos[59,60]. Estos métodos se caracterizan por considerar relaciones complejas entre variables, lo cual les permite prescindir de decisiones subjetivas o que dependen de la intervención humana, a diferencia de la Regresión Logística y la Regresión Lineal, que asumen una relación únicamente lineal o logística entre las variables, y, por lo tanto, condicionan *a priori* el desempeño de los modelos generados. Son capaces además, de detectar patrones en grandes *sets* de datos, que son desconocidos *a priori*[57,58]. Beam y Kohane destacan el llamado *Machine Learning Spectrum*, descrito como el intercambio que va desde las especificaciones humanas de las propiedades de un algoritmo predictivo versus el aprendizaje de esas propiedades a partir de los datos. A medida que los humanos imponen menos suposiciones en los modelos, los algoritmos se acercan más al AM. Sin embargo, destacan también que no existe un umbral determinado dentro de este espectro para definir cuándo un algoritmo se

transforma en AM[56]. Los algoritmos más cercanos al AM en el espectro destacan por ser más flexibles (modelan de mejor forma relaciones complejas entre variables), pero son frecuentemente menos interpretables y funcionan como “caja negra”, y además requieren de una gran cantidad de datos y recursos computacionales. En contraste, los algoritmos ubicados más lejos del AM en el espectro son fácilmente interpretables por humanos, pero menos flexibles, es decir, que imponen ajustar los datos a relaciones establecidas *a priori*.

Los métodos de AM pueden ser clasificados en supervisados, no supervisados y reforzados[60,61]. En el aprendizaje supervisado, para cada *set* de variables predictoras (*inputs*), existe una variable respuesta conocida (*outcome*), y se trata de aprender de los datos de modo de relacionar los predictores con la respuesta[59,60]. En general los datos disponibles se dividen en un conjunto de entrenamiento (*training set)* y un conjunto de prueba (*testing set)*, donde el primero es usado para aprender la relación entre las variables predictoras y la respuesta, mientras que el conjunto de prueba sirve para calcular el poder predictivo del algoritmo[35]. Según el tipo de variable respuesta, el AM puede ser considerado un problema de clasificación o regresión. En la clasificación, la variable respuesta es discreta, es decir, existen clases conocidas. Por otro lado, en la regresión, la variable respuesta es continua[60]. En adelante, cuando se utilice el término “regresión”, se refiere al conjunto de métodos que resuelven problemas donde la variable respuesta es continua, al referirse a la “regresión lineal” como método estadístico, se utilizará el término “Regresión Lineal” (con mayúsculas) y cuando se refiera a “regresión logística”, se utilizará el término “Regresión Logística” (con mayúsculas).

En el aprendizaje no supervisado no existe un *outcome* definido, sino que se busca encontrar patrones en los datos a partir de sus propias características[59,60]. Tal como es explicado en el libro *“An Introduction to Statistical Learning”*, escrito por James *et al*., el aprendizaje no supervisado es como pedirles a niños que separen grupos de objetos sin decirles la clase a la que pertenecen. En ese caso es necesario reconocer características comunes entre ellos con el fin de agruparlos. El ejemplo más clásico de aprendizaje no supervisado es el *clustering*[60,62].

Finalmente, los métodos de reforzamiento son aquellos en los que se aprende constantemente de los datos por medio de la optimización de una función recompensa[63]. Un ejemplo de este método es el utilizado en sistemas de interfaz cerebro-máquina, donde el *input* neuronal no es estacionario[64].

A continuación, se presentan ejemplos de métodos de AM aplicados al estudio de la obesidad.

- Selya y Anshutz, en el libro *Advanced Data Analytics in Health*, clasifican el estado nutricional de las personas en base a información de dieta (recordatorio 24 horas) y actividad física de la encuesta nacional de alimentación estadounidense (*National Health and Nutrition Examination Survey*), utilizando algoritmos de Support Vector Machine y Redes Neuronales, entre otros, donde demuestran que los métodos de clasificación permiten predecir el riesgo de obesidad. Los autores destacan que los métodos de AM pueden mejorar la predicción del riesgo de resultados en salud, por sobre los enfoques epidemiológicos convencionales[35].
- Nau *et al.* utilizan *Random Forest* (método que consiste en un conjunto de árboles de decisión[60]) para encontrar los mejores predictores de riesgo obesidad infantil dentro de una serie de variables, que incluyeron datos del entorno económico, social, alimentario y de actividad física. Es importante mencionar que este método no solo entrega un valor para el error con el cual predice el resultado, sino también un *ranking* de las variables predictoras que modifican el poder de predicción. Destacan que el método utilizado es una herramienta innovadora y flexible para operacionalizar entornos de riesgo de forma ecológica y proveen estrategias de análisis que reconocen la complejidad, en vez de reducirla imponiendo supuestos que pueden ser no razonables, prescindiendo así de los enfoques de una variables a la vez, que terminan en resultados inconsistentes[65].
- Seyednasrollah *et al*. evalúan, usando métodos de AM con *boosting* (método general que mejora la exactitud de los algoritmos de AM[66], determinando reglas que disminuyen el error de predicción), los factores genéticos y clínicos

de la infancia con el fin de predecir riesgo de obesidad en la adultez, resultando en un modelo útil para el *screening* de niños con alto riesgo de desarrollar obesidad[67].

- Dugan *et al*. logran predecir obesidad en niños basados en información rescatada de fichas clínicas electrónicas, comparando diferentes modelos de clasificación (*Random Tree*, *Naive Bayes,* entre otros). Destacan el uso de técnicas de AM supervisado para descubrir nuevas asociaciones que previamente no se habían identificado con técnicas estadísticas tradicionales[57].
- Otros ejemplos del uso de métodos de AM en el estudio de la obesidad se pueden encontrar en Sze y Schloss[68], Thaiss *et al*.[69] y Lee *et al*.[70].

Por otro lado, métodos de AM han sido utilizados en el estudio de la dieta, enfocados principalmente en análisis de patrones, explorando componentes de la dieta que predicen diferentes estados de salud:

- Giabbanelli y Adams utilizaron árboles de decisión (métodos de AM de clasificación) para determinar el número de alimentos necesarios para predecir de forma precisa el cumplimiento o no de recomendaciones de consumo de frutas y vegetales, azúcares, sodio grasas y grasas saturadas, destacando la eficiencia para capturar relaciones múltiples no lineales y efectos de interacción entre las variables[71].
- Lazarou *et al*. utilizaron árboles de decisión, análisis de componentes principales y Regresión Logística para determinar que variables de hábitos alimenticios se relacionan con la obesidad en niños, destacando que el uso de estos métodos permite definir patrones que pueden no haber sido determinados con estadística tradicional[72].
- Kastorini *et al*. realizaron una comparación entre patrones dietarios determinados *a priori* (utiliza índices dietarios para capturar patrones pre-definidos) y *a posteriori* (determinado por el análisis multivariado de datos dietarios de los sujetos en estudio) de la exactitud de predicción para

síndrome coronario agudo, utilizando algoritmos de clasificación como *Naive Bayes*, árboles de decisión, redes neuronales, *Support Vector Machine*, entre otros[73].

- Thangamani y Sudha utilizan datos de dieta y variable sociodemográficas para clasificar a el estado nutricional de niños, implementando algoritmos como árboles de decisión, *random forest* y redes neuronales. Los autores describen que las técnicas utlizadas proveen métodos flexibles y apropiados para grandes cantidades de datos con el objetivo de detectar malnutrición[74].
- Otros ejemplos del uso de AM en el estudio de la dieta se pueden encontrar en Einsel *et al*.[75], Weber y Achananuparp[76] y Hossain *et al*.[77].

Basado en lo expuesto, los métodos de AM surgen como herramientas adoptadas en el contexto del estudio de la obesidad y la dieta; y de utilidad, ya que cuentan con el potencial para considerar una gran cantidad de variables predictoras, con el fin de generar un modelo lo suficientemente flexible para representar un fenómeno dado en diversas áreas como: genética y genómica[78], estudio del cáncer[79], astronomía[80] y finanzas[81]; cuyas propiedades predictivas se generen en base a los datos.

En el caso del presente trabajo, se propone utilizar métodos de AM para determinar si algunos de los factores asociados a o causantes de la obesidad (consumo de alimentos y variables sociodemográficas) pueden predecir el sobrepeso o la obesidad de cierta población.

2. HIPÓTESIS

El estado nutricional de la población chilena puede ser predicho a partir del consumo de alimentos individuales y/o los datos sociodemográficos de la población chilena, con un 90% de exactitud.

3. OBJETIVO GENERAL

Implementar algoritmos de Aprendizaje de Máquinas de clasificación y regresión para predecir el estado nutricional (IMC o clasificación de estado nutricional) de la población chilena adulta, a partir de datos de consumo alimentario, características sociodemográficas individuales o la combinación de estos.

4. OBJETIVOS ESPECÍFICOS

1. Explorar los datos disponibles.
2. Implementar algoritmos de Aprendizaje de Máquinas de clasificación y regresión que predigan el estado nutricional a partir del consumo de alimentos de las personas.
3. Implementar algoritmos de Aprendizaje de Máquinas de clasificación y regresión que predigan el estado nutricional a partir de características sociodemográficas de las personas.
4. Implementar algoritmos de Aprendizaje de Máquinas de clasificación y regresión que predicen el estado nutricional a partir de la combinación de consumo de alimentos y características sociodemográficas de las personas.
5. Comparar el desempeño de los algoritmos implementados que predigan el estado nutricional.
6. Identificar grupos de alimentos relevantes en la predicción de obesidad.
7. Explorar el efecto de la actividad física y sedentarismo sobre la predicción de obesidad.

5. MATERIAL Y MÉTODOS

La presente tesis utilizó como fuente de datos la Encuesta Nacional del Consumo Alimentario 2010 – 2011 (ENCA)[82]. Esta encuesta se llevó a cabo en Chile con el objetivo de "… conocer los patrones de consumo, en el contexto de conductas y hábitos alimentarios de la población chilena, y así entregar antecedentes para el diseño de políticas públicas e intervenciones específicas para mejorar el estado nutricional de los chilenos y chilenas."[82]. La ENCA cuenta con una muestra representativa por macrozonas (norte, centro norte, centro sur, sur y metropolitana) y por área (urbana o rural), asegurando así la cobertura nacional. Fue realizada por la Escuela de Salud Pública, el Departamento de Nutrición de la Facultad de Medicina y el Centro de Microdatos de la Facultad de Economía y Negocios de la Universidad de Chile, en temporada primavera-verano (noviembre 2010 y enero 2011), y cuenta con datos de 4.920 sujetos válidos disponibles en bases de datos de libre acceso (disponibles en http://web.minsal.cl/enca/ [83]) .

La ENCA cuenta con datos referente a[82]: variables socioeconómicas, de salud y estilos de vida; antropometría, que provee la clasificación de estado nutricional de los encuestados (peso y altura de los entrevistados); frecuencia de consumo en un mes para 457 alimentos, obtenidos desde la Encuesta de Tendencia de Consumo Cuantificado (ETCC); nutrientes consumidos en las últimas 24 horas, obtenidos a partir de la Encuesta de Recordatorio de 24 horas (R24h). La variabilidad intraindividual de estos datos fue estimada comparando dos mediciones, separadas por 2 - 3 días, para el 20% de los sujetos de la muestra.

Estos datos están organizados en varios archivos disponibles en la página web[83], donde cada archivo corresponde a diferentes agrupaciones de datos, según la descripción que se entrega en el Manual de Usuario[84]. En las bases de datos de la ENCA, cada sujeto es identificado por un número identificador único (folio), permitiendo el cruce de variables de interés, descritas a continuación, entre bases de datos. Esta tesis considera el pre-procesamiento de estas bases de datos para obtener las variables que serán utilizadas.

En el Anexo 1, Figura 12, extraída del manual de usuario ENCA[84], muestra la secuencia de preguntas realizadas en la Encuesta de Tendencia de Consumo Cuantificado (ETCC) para obtener la cantidad y frecuencia mensual de consumo de 457 alimentos distintos. Estos alimentos fueron agrupados por los autores de la encuesta en 32 grupos. En el presente trabajo se utilizaron las bases de datos con reporte de consumo mensual de alimentos no agrupado y agrupado.

En el Anexo 1, Figura 13, se muestra la secuencia de preguntas realizadas para la captación de la cantidad de alimentos consumida el día anterior a la toma de datos en la Encuesta de Recordatorio de 24 horas (R24h). Al relacionar estos datos con la Tabla de Composición de Alimentos (elaborada por autores de la ENCA basados en Tablas de Composición de Alimentos Chilenos, información de la USDA (*Nutrient Database for Standard Reference*) y, para alimentos procesados, información de etiquetado[84]), se genera la base de datos de nutrientes consumidos por los individuos el día anterior a la encuesta[84]. Los datos demográficos que fueron recolectados como parte de la ENCA[84] son: género (masculino/femenino), edad (años al momento de la encuesta), nivel socioeconómico (obtenido de percentiles de puntaje calculado en base a bienes del hogar y nivel educacional del jefe de hogar; bajo/medio bajo/medio/medio alto/alto), área de residencia (urbano/rural) y macrozona del país (norte/centro norte/metropolitana/centro sur/sur).

Se obtuvieron tablas organizadas tal que las filas representan a cada sujeto y las columnas sus características demográficas o de consumo de alimentos según lo descrito anteriormente (consumo mensual de alimentos no agrupado, consumo mensual de alimentos agrupados o nutrientes, ver Tabla 1. En Anexo 2 se encuentran las variables seleccionadas para esta tesis).

Las variables predictoras (inputs) utilizadas son: A. Variables de dieta: consumo mensual no agrupado (457 categorías de alimentos), consumo mensual agrupado (32 categorías de alimentos agrupados) o nutrientes (8 variables); B. variables sociodemográficas (5 variables); y la combinación de A y B (cuando las variables de consumo de alimentos A se combinan con variables sociodemográficas B, el set de

datos en conjunto se le asignará la sigla “cSOC”, cuando no, se utilizará la sigla “sSOC”). La Figura 1 representa la esquematización del modelo de análisis descrito.

Tabla 1. Descripción de tablas obtenidas de las bases de datos ENCA

Nombre Tabla	Fuente	Dimensión (filas x columnas)	Observaciones
Consumo Mensual Alimentos No Agrupado	ETCC	4.920 x 471	Incluye folio
Consumo Mensual Alimentos Agrupado	ETCC	4.920 x 33	Incluye folio
Nutrientes	R24h	4.920 x 9	Incluye folio
Sociodemográfico	Encuesta Sociodemográfica	4.920 x 8	Incluye folio, estado nutricional, IMC

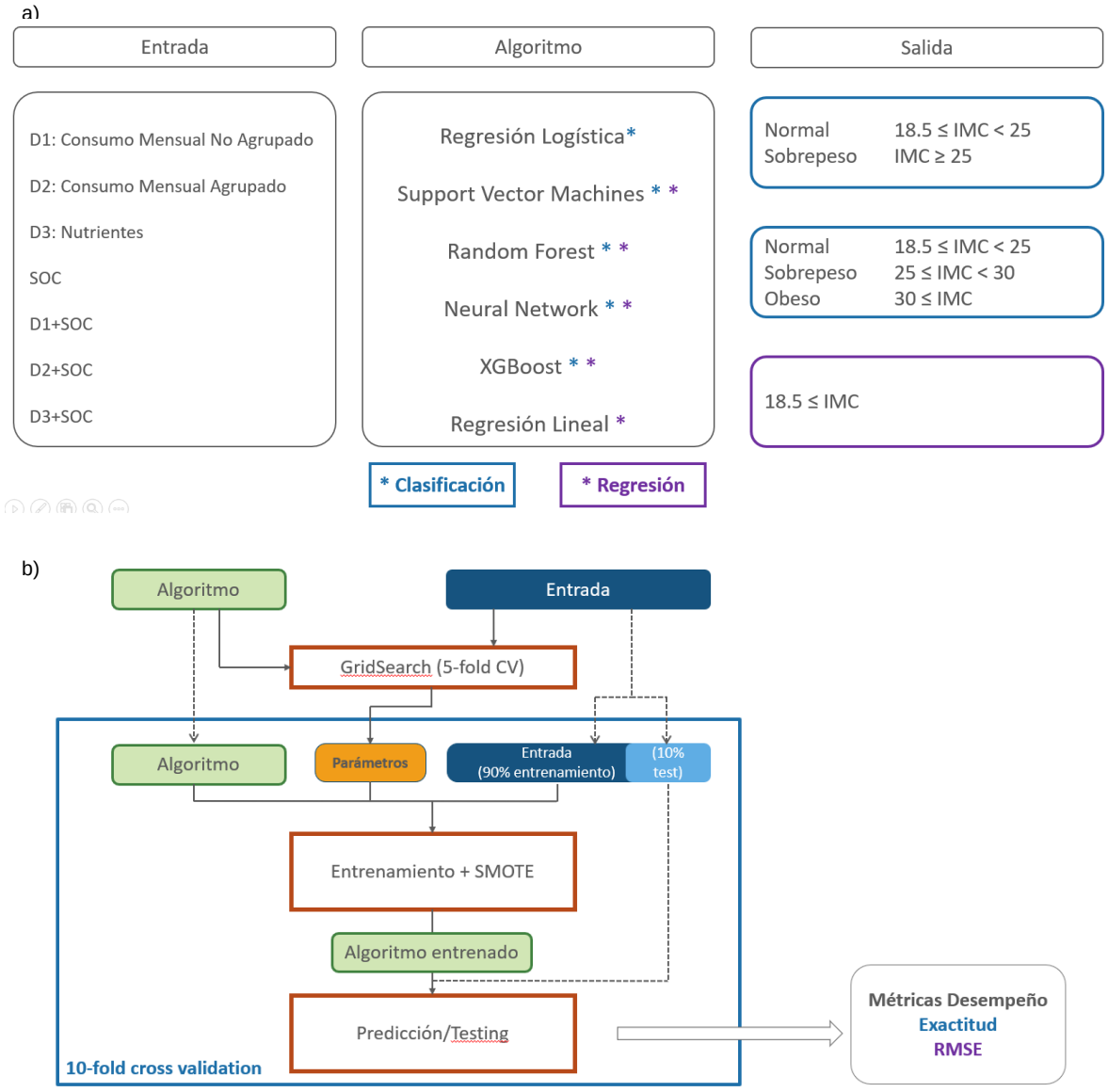


Figura 1. Modelo Análisis. a) Esquema de modelo de análisis para predicción de estado nutricional o IMC. SOC: Sociodemográfico. Entrada: variables predictoras del modelo y las combinaciones. Algoritmos: Algoritmos utilizados para construir modelos. Salida: Azul: Variable a predecir discreta; Morado: Variable a predecir continua. b) Esquema de análisis predictivo utilizado.

Para el caso presentado en esta tesis, los algoritmos de Aprendizaje de Máquinas implementados son de aprendizaje supervisado, ya que se conocen las variables predictoras (dieta y sociodemográfico) y la variable respuesta (estado nutricional o

IMC). La variable respuesta (IMC) es continua, y se discretiza con las siguientes categorías: normal (18.5 ≤ IMC < 25), sobrepeso (25 ≤ IMC < 30) y obeso (IMC ≥ 30). Se realizó la clasificación con variable respuesta binaria, donde las clases son “Normal” (18.5 ≤ IMC < 25) y “Sobrepeso” (IMC ≥ 25), agrupando en esta última categoría a las personas con sobrepeso u obesidad. Además, se realizó la clasificación multiclase, usando la variable respuesta en las tres categorías originales (normal, sobrepeso y obeso). Cuando la variable respuesta es discreta, se utilizaron métodos de clasificación, y en el caso de que la variable respuesta sea continua, se utilizaron métodos de regresión.

Los casos elegibles para esta tesis cumplen con los siguientes criterios de inclusión: edad mayor a 18 años y menor a 65 años; estado nutricional Normal, Sobrepeso u Obeso (IMC ≥ 19, excluyendo a los casos con IMC < 19); contar con toda la información de las variables en estudio (solo casos completos). Se excluyeron a mujeres embarazadas y mujeres lactantes.

Se realizó análisis exploratorio de los datos, que consistió en:

- Descripción de muestra, a través de gráficos (histogramas y barras de distribución porcentual) y medidas de tendencia central para cada variable sociodemográfica, separados por estado nutricional.
- Prueba ANOVA/Kruskall-Wallis (según resultado de la prueba de normalidad Shapiro-Wilk, *p-value* < 0,05 de significancia) para cada variable en estudio, considerando la separación por grupo de estado nutricional.
- Análisis de correlación de Pearson entre variables en estudio de cada set de datos, de la cual se obtiene una matriz de coeficientes de correlación, donde -1 equivale a correlación inversa entre variables y 1 corresponde a correlación directa entre variables. La matriz de correlación es reportada a través de un mapa de calor que representa las variables más y menos correlacionadas. Además, se calculó la media y desviación estándar de los coeficientes de correlación y se reportaron las parejas de variables con mayor coeficiente de correlación para cada set de datos.

- Análisis de Componentes Principales (ACP) de cada set de datos de dieta. El ACP corresponde a una forma de resumir datos en un set de variables, llamados componentes, que colectivamente representan la mayor parte de la variabilidad de los datos originales. Son una representación de los datos originales en una dimensión menor, con la mayor variabilidad posible. Cada uno de los componentes corresponde a la combinación lineal de las variables originales[60]. Se considera que se deben usar una cantidad de componentes para resumir los datos, tal que se alcance un 80% de la variabilidad explicada por esos componentes. Esta información se expresa en el gráfico de sedimentación. Se reporta también el gráfico de las tres primeras componentes, con el fin de visualizar de forma resumida los datos, coloreando cada punto según el estado nutricional.

Para obtener datos aptos para la implementación de los algoritmos de Aprendizaje de Máquinas, se realizó: 1) *Winzorisacion*[85], para tratamiento de *outliers*, donde los valores mayores a cierto umbral determinado por el intervalo de confianza de 95% son reemplazados por el valor umbral; 2) eliminación de variables no informativas (en las cuales no existe consumo de alimento reportado para ningún sujeto), con el criterio de media y desviación estándar por variable con valor igual a 0; 3) eliminación de variables predictoras (dieta y sociodemográficas) donde el test de ANOVA/Kruskall-Wallis resulte sin diferencias significativas (*p-value* > 0,05) para la variable, entre las categorías de estado nutricional; estandarización (centrar y reducir) de cada variable, con el fin de permitir comparaciones en escala y unidad de medida.

Los algoritmos implementados en esta tesis son caracterizados a continuación:

- Regresión Logística (RegLog): consiste en un modelo estadístico de clasificación de categorías, en el cual se modela la probabilidad de una clase o evento en función de otros factores. Es parte de los Modelos Lineales Generalizados, donde la función de enlace corresponde a una función logística de la probabilidad (*odds*) de que ocurra el evento que se quiere predecir. Alguna de sus ventajas son que cuenta con varias extensiones para poder

predecir variables con más de una categoría (Regresión Logística Multinominal) o para predecir variables categóricas ordinales (Regresión Logística Ordinal). Una desventaja de interpretación es que el modelo en sí no predice la clase, sino que predice la probabilidad de ésta (entre 0 y 1), teniendo que decidir algún punto de corte para realizar la clasificación[59].

- Regresión Lineal (RegLin[2]): es un modelo estadístico que modela la relación entre una respuesta escalar y una o más variables independientes, asumiendo una relación lineal (intercepto y coeficientes) entre las variables predictoras y las variables predichas. Tiene como ventaja la fácil implementación y bajo costo computacional. Para objetivar el ajuste del modelo se utilizan mediciones como el error estándar residual o el R^2 (proporción de varianza explicada de la variable predicha al usar las variables predictoras). Es uno de los métodos más utilizados y sirve como base estadística para otros métodos de AM más avanzados[59,60]. Tiene la desventaja de asumir *a priori* relaciones lineares entre la variable predicha y variables predictoras, y modelando de forma deficiente cuando existe colinearidad entre variables.
- *Support Vector Machine* (SVM): es un modelo basado en la generalización del clasificador *máximo marginal*, donde los vectores de entrada (datos) están mapeados de una forma no lineal a un espacio de alta dimensionalidad, agrandado usando *kernels*, con el fin de acomodar límites no-lineales entre las clases, que es una de sus ventajas principales. Las clases son separados por un hiperplano en el espacio de alta dimensionalidad. Una de sus desventajas es que requiere que las clases que se predicen estén separadas linealmente y las variables predictoras se encuentren normalizadas. SVM está destinado principalmente a problemas de clasificación binaria, aunque existen extensiones para utilizarse en problemas de multiclase[59,60,86].
- *Random Forest* (RF): consiste en un conjunto de árboles de decisión con baja correlación, donde las variables predictoras de cada árbol se eligen de forma aleatoria. Los árboles de decisiones son métodos que segmentan el espacio predictor en un número de regiones más simples, donde se utiliza la media o la

moda de las observaciones de entrenamiento de ese espacio para predecir sobre nuevas observaciones. Estos métodos se caracterizan por ser fáciles de interpretar, pero en general no son competitivos con otros algoritmos, es por esto los bosques aleatorios (*Random Forest*) tienen mejores desempeños ya que consisten en una gran cantidad de árboles que toman decisiones en consenso, que es una de sus ventajas, junto con la capacidad de evaluar las variables predictoras por si solas, no requiriendo normalización[60]. Las desventajas de estos métodos son principalmente la alta complejidad y, al consistir en un conjunto de árboles, pierden intuitividad.

- *Neural Networks* (NN): es modelo de aprendizaje que simula el funcionamiento de las redes neuronales biológicas. Se basa en nodos (neuronas), las cuales computan una salida (conexiones), basados en una entrada y una función no lineal. La salida tiene un peso, el cual se utiliza como función de aprendizaje con el fin de fortalecer o debilitar una "*conexión*". La red neuronal consiste en varias capas de neuronas interconectadas. Una de sus ventajas es que estos modelos "aprenden" en base a ejemplos sin tener información *a priori* de las clases que debe diferenciar, generando automáticamente características a partir de los datos de entrenamiento. Se utilizan comúnmente para tareas de visión computacional, reconocimiento de voz y diagnóstico médico, entre otros. Una desventaja es el alto costo computacional ya que requiere una gran cantidad de datos para desempeñarse bien que requiere su implementación y la baja interpretabilidad de los modelos generados[59].
- *xGBoost* (XGB): el *Extreme Gradient Boosting* es un algoritmo variante del *Gradient Boosting*, el cual consiste en generar árboles de decisión con parámetros optimizados de forma consecutiva, los cuales se van ajustando en base a una función de pérdida definida. Este método es relativamente nuevo, pero se ha utilizado exitosamente en distintos problemas de clasificación[87]. La principal desventaja es el tiempo que demora en generar el modelo, ya que construye árboles secuenciales, sin embargo, ha demostrado mejor desempeño que otros algoritmos.

Según corresponda de acuerdo al algoritmo, se realizó búsqueda y ajuste de hiper-parámetros utilizando el método de búsqueda por grilla (*grid-search*)[88], en el cual se realiza una búsqueda exhaustiva en un set de valores manualmente asignado para los hiper-parámetros de un algoritmo. La búsqueda por grilla es guiada por una métrica de desempeño, la escogida para este trabajo fue la exactitud. Cuando en los datos disponibles la variable respuesta está desbalanceada (es decir, exista una clase minoritaria versus una mayoritaria), se implementó la técnica *SMOTE* (*Synthetic Minority Over-sampling Technique)*[89], la cual consiste en crear nuevos casos sintéticos de la clase minoritaria, estadísticamente similares a los casos existentes, con el fin de que las clases estén balanceadas (mismo número de casos para ambas clases)[90].

Para cada algoritmo, y con el fin de garantizar la reproducibilidad de los métodos, se implementó *k-fold cross validation* (k=10), método que divide el *set* de observaciones en k grupos (*folds*) del mismo tamaño. Se realiza el entrenamiento del modelo con 9 de los 10 grupos, dejando el décimo para validar el modelo (conjunto de prueba o *test set)*, iterando para que cada grupo cumpla la función de conjunto prueba[60].

La evaluación del desempeño de los métodos de clasificación se realizó a partir de las métricas obtenidas desde la matriz de confusión, la cual se construye en base a los valores reales del conjunto de prueba y los predichos obtenidos de la implementación de los algoritmos:

Matriz de Confusión

		Condición Real	
		Negativo	Positivo
Condición Predicha	Negativo	Verdadero Negativo (VN)	Falso Negativo (FN)
	Positivo	Falso Positivo (FP)	Verdadero Positivo (VP)

Para el caso presentado en esta tesis, el positivo representará el estado "Sobrepeso". Las métricas que se utilizaron fueron[60]:

- Exactitud: corresponde a la proporción de todos casos (positivos y negativos) identificados correctamente por el clasificador con respecto al total de casos. Su fórmula es:

$$\frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{VN}}$$

- Precisión: corresponde al porcentaje de casos positivos identificados correctamente con respecto a todos los casos clasificados positivos del clasificador. Su fórmula es:

$$\frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}}$$

- Sensibilidad: corresponde al porcentaje de casos positivos identificados por el clasificador con respecto a los casos positivos reales. Su fórmula es:

$$\frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}}$$

- Especificidad: Corresponde a el porcentaje de casos negativos identificados por el clasificador. Su fórmula es:

$$\frac{\mathrm{VN}}{\mathrm{VN} + \mathrm{FP}}$$

- F1: Corresponde a la media harmónica de la precisión y la sensibilidad, se utiliza como indicador de relación entre estas dos métricas. Su fórmula es:

$$2 \cdot \frac{\mathrm{precision} \cdot \mathrm{sensibilidad}}{\mathrm{precision} + \mathrm{precision}} = \frac{2\mathrm{VP}}{2\mathrm{VP} + \mathrm{FP} + FN}$$

Ya que los resultados obtenidos en cada iteración o *fold* de las implementaciones de cada combinación de algoritmos con set de datos son variables (similares a

experimentos repetidos), se aplicaron pruebas de normalidad (Shapiro Wilk). Considerando una distribución normal, se reportó la media y desviación estándar de cada métrica, y se visualizaron con gráficos de caja. Para determinar qué algoritmo tiene mejor rendimiento, se utilizó reporte (media y desviación estándar) y pruebas estadísticas paramétricas para determinar si las diferencias son significativas (ANOVA con corrección Bonferroni) en la métrica de exactitud, así como en Precisión, Sensibilidad, Especificidad y F1. Para el caso de la clasificación con 3 clases, se utilizaron las métricas: exactitud, precisión, especificidad y F1, considerando la media macro de cada una, la cual considera la métrica para cada clase por separado, y luego calcula la media[91].

Para los métodos de regresión, la evaluación se realizó obteniendo la raíz cuadrada del error cuadrático medio (*RMSE: Root Mean Square Error*), determinado por la comparación entre los valores reales y predichos del IMC. En este caso también se utilizó ANOVA con corrección Bonferroni para determinar el mejor rendimiento entre algoritmos y *set* de datos.

Una vez obtenidos los resultados de desempeño de la implementación de algoritmos para la clasificación de 2 clases y regresión, considerando las mejores combinaciones (algoritmo y combinación de *sets* de datos), se realizaron los siguientes ejercicios de exploración:

- Se implementaron predicciones para clasificación de 2 clases y regresión, estratificando a la población en rangos etarios (19-28, 29-38, 39-48, 49-58, 59-64) y género.
- Se implementaron modelos predictivos considerando la actividad física como variable predictora. Se realizaron comparaciones con la prueba estadística Wilcoxon.
- Considerando el consumo mensual agrupado, se obtuvieron las variables de mayor importancia al realizar el entrenamiento y predicción del estado nutricional o del IMC.

- Se escogieron 14 variables del consumo mensual agrupado (Anexo 3) las cuales pertenecen a grupos de alimentos que *a priori* se asocian a un alto consumo de calorías. Se implementaron modelos predictivos considerando solo estas variables y se compararon con la referencia, que considera todas las variables de esa base de datos (prueba estadística Wilcoxon).

Para todas las pruebas estadísticas se considera *p-value* < 0,05 de significancia.

Se escogieron dos variables a explorar como representativas de la actividad física. La EER (*Estimated Energy Requirement*) se define como el promedio de energía necesaria para mantener un balance energético saludable en un sujeto de cierta edad, género, peso, talla y nivel de actividad física[92]. Por otro lado, considerando la referencia de los autores de la ENCA, el nivel de actividad física se obtuvo a partir de la variable "p50" de la ENCA[82,84], la cual corresponde a la pregunta: ¿con qué frecuencia hace ejercicio hasta quedar transpirando o sin aire?, agrupando en 4 grupos: Sedentario (nunca realiza actividad física), Ligeramente Activo (realiza actividad física menos de una vez al mes o una vez al mes), Activo (realiza actividad física 1 vez por semana o 2-3 veces por semana) y Muy Activo (realiza actividad física 4-6 veces por semana o todos los días).

Para llevar a cabo el preprocesamiento, análisis, visualización e implementación de los algoritmos de Aprendizaje de Máquinas, se utilizó el lenguaje de programación Python y las librerías *scikit-learn*[93], orientada al análisis de datos y *data mining*; *scipy*[94]*,* librería dedicada a análisis estadístico de datos en ciencias, matemáticas e ingenieria; *xgboost*[87]*,* librería diseñada especíicamente para realizar *gradient boosting* distribuido de forma óptima; *imblearn*[89], librería de herramientas para set de datos desbalanceados utilizados en AM.

6. RESULTADOS

6.1 Descripción de la Muestra

De los datos disponibles (N = 4.920) se utilizaron n = 2.284 que cumplen con los criterios de inclusión descritos. La Tabla 2 muestra el resumen de las características sociodemográficas de la muestra. En la Figura 2 se observan los gráficos representativos de las variables sociodemográficas, agrupando según el estado nutricional en 2 ("Normal" / "Sobrepeso") y 3 clases ("Normal" / "Sobrepeso" / "Obeso").

Tabla 2. Descripción de muestra y resumen de variables sociodemográficas.

		Normal		**Sobrepeso[1]**		**Población Total**	
N (% del total)		624 (27,3)		1.660 (72,7)		2.284 (100)	
		media	**DE[2]**	**media**	**DE**	**Media**	**DE**
Edad		37,58	13,85	44,19	12,84	42,38	13,45
IMC		22,86	1,56	30,78	4,76	28,61	5,44
		%	**IC[3]**	**%**	**IC**	**%**	**IC**
Genero	Hombre	39,1	35,3 – 42,9	32,0	29,7 – 34,2	33,9	32,0 – 35,9
	Mujer	60,9	57,1 – 64,7	68,0	65,8 – 70,3	66,1	64,1 – 68,0
Área	Urbano	87,8	85,3 – 90,4	87,7	86,1 – 89,2	87,7	86,3 – 89,0
	Rural	12,2	9,6 – 14,7	12,3	10,8 – 13,9	12,3	11,0 – 13,7
Macrozona	Norte	11,4	8,9 – 13,9	10,2	8,8 – 11,7	10,6	9,3 – 11,8
	Centro-Norte	22,1	18,9 – 25,4	19,5	17,6 – 21,4	20,2	18,6 – 21,9
	Metropolitana	41,2	37,3 – 45,0	39,2	36,9 – 41,6	39,8	37,7 – 41,8
	Centro-Sur	15,4	12,6 – 18,2	18,3	16,5 – 20,2	17,5	16,0 – 19,1
	Sur	9,9	7,6 – 12,3	12,7	11,1 – 14,3	12,0	10,6 – 13,3
Nivel Socio-económico	Alto	12,2	9,6 – 14,7	8,7	7,4 – 10,1	9,7	8,5 – 10,9
	Medio-Alto	22,6	19,3 – 25,9	21,4	19,4 – 23,4	21,7	20,0 – 23,4
	Medio	20,7	17,5 – 23,9	24,2	22,1 – 26,2	23,2	21,5 – 24,9
	Medio-Bajo	34,3	30,6 – 38,0	34,6	32,3 – 36,9	34,5	32,6 – 36,5
	Bajo	10,3	7,9 – 12,6	11,1	9,6 – 12,7	10,9	9,6 – 12,2

[1]Sobrepeso representa la suma de casos con sobrepeso y obesos
[2]DE: Desviación Estándar; [3]IC: Intervalo de 95% confianza de proporciones

6.2 Análisis Exploratorio

Luego de aplicar la prueba de normalidad Shapiro-Wilk a las variables de consumo de alimentos y sociodemográficas, más del 95% de las variables presentan una distribución no normal (*p-value* < 0,05), por lo que se utilizaron pruebas no paramétricas para la comparación de variables según grupo. Al comparar la varianza de las variables explicativas (prueba Kruskal-Wallis), separando por clases según estado nutricional ("Normal" / "Sobrepeso"), los porcentajes de variables que muestran diferencias significativas (*p-value* < 0,05) entre las varianzas interclase, al compararlas con las varianzas intraclase, son: consumo mensual no agrupado: 12%; consumo mensual agrupado: 50%; nutrientes: 0%. Para las variables sociodemográficas, solo existen diferencias entre grupos para la edad, género y la macrozona (Tabla 3).

Tabla 3. Resumen de pruebas de normalidad y Kruskal-Wallis por base de datos.

	N° Total de Variables	% de variables con distribución normal	% de variables con diferencias significativas (Kruskal-Wallis)
Consumo Mensual No Agrupado	457	5,5	12,0
Consumo Mensual Agrupado	40	0	50,0
Nutrientes	8	0	0
Sociodemográficas	5	0	60,0

[1] Test Shapiro Wilk, *p-value* > 0,05; [2] Test Kruskal Wallis, *p-value* < 0,05

En cuanto a las correlaciones (Pearson), a continuación se reporta la media, DE y las 3 variables con mayores coeficientes de correlación para cada base de datos[1]. En consumo mensual (media = 0,02; DE = 0,05), las parejas con mayor correlación fueron: Fideos Blancos de Acompañamiento con Arroz Blanco de Acompañamiento (coef. correlación = 0,4), Perejil con Cilantro (coef. correlación = 0,3) y Vienesas con Fideos Blancos de Acompañamiento (coef. correlación = 0,2); para el consumo mensual agrupado (media = 0,04; DE = 0,09), las parejas fueron: Lácteo Años Dorados con Crema Años Dorados (coef. correlación = 0,5), Leguminosas Frescas con Cereales y Pastas (coef. correlación = 0,3) y Aceites Saturados con Cereales y

[1] Cálculos realizados sin considerar las autocorrelaciones.

Pastas (coef. correlación = 0,3); y para los nutrientes (media = 0,6; DE = 0,1), las parejas fueron: Grasa Saturadas con Lípidos (coef. correlación = 0,9), Grasas Monosaturadas con Lípidos (coef. correlación = 0,9) y Grasas Polisaturadas con Lípidos (coef. correlación = 0,8). En la Figura 3, se observa el mapa de calor de la correlación entre variables en estudio, según base de datos.

Al realizar el análisis de componentes principales, en los tres grupos de variables en estudio se observa que las primeras componentes no logran explicar con suficiencia las varianzas. Se reporta a continuación el número de componentes requeridas para lograr un 80% de varianza explicada para cada base de datos. Consumo mensual: 191 componentes; consumo mensual agrupado: 24 componentes; nutrientes: 3 componentes. Por otro lado, al visualizar las primeras tres componentes y las clases en estudio, no se observan agrupaciones naturales de estas clases. En la Figura 4 se muestran los gráficos de sedimentación y la visualización de las primeras 3 componentes para cada base de datos general, a través de la exploración se observa que, si bien existe una gran cantidad de datos con respecto a la dieta y variables sociodemográficas, estos no demuestran diferencias entre los estados nutricionales en estudio.

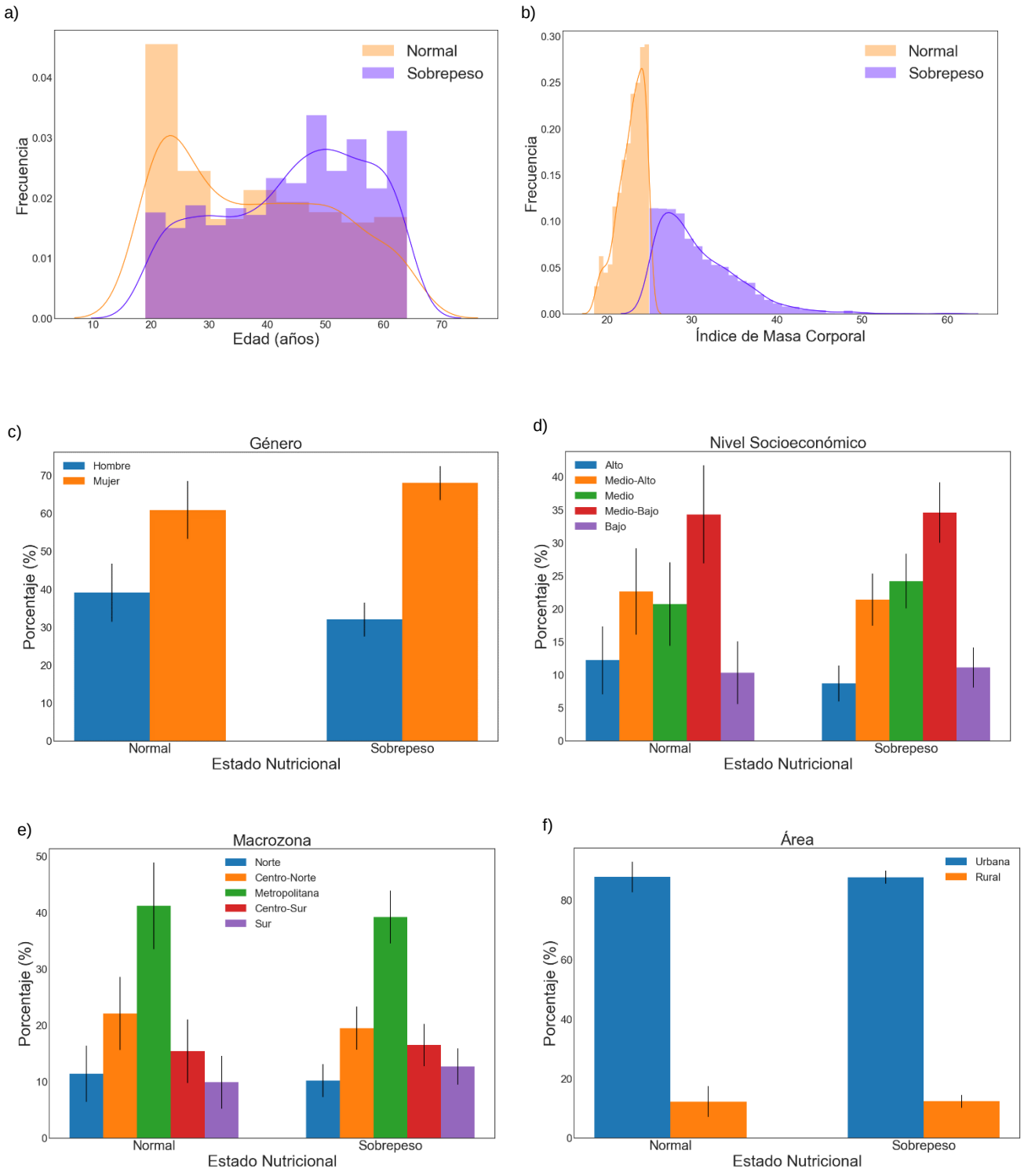


Figura 2. Descripción de variables sociodemográficas. Según estado nutricional (Normal, Sobrepeso). (a) Histograma de edad. (b) Histograma de Índice de Masa Corporal. (c) Porcentaje de población por género. (d) Porcentaje de población por nivel socioeconómico. (e) Porcentaje de población por macrozonas geográficas. (f) Porcentaje de población por área (urbana y rural).

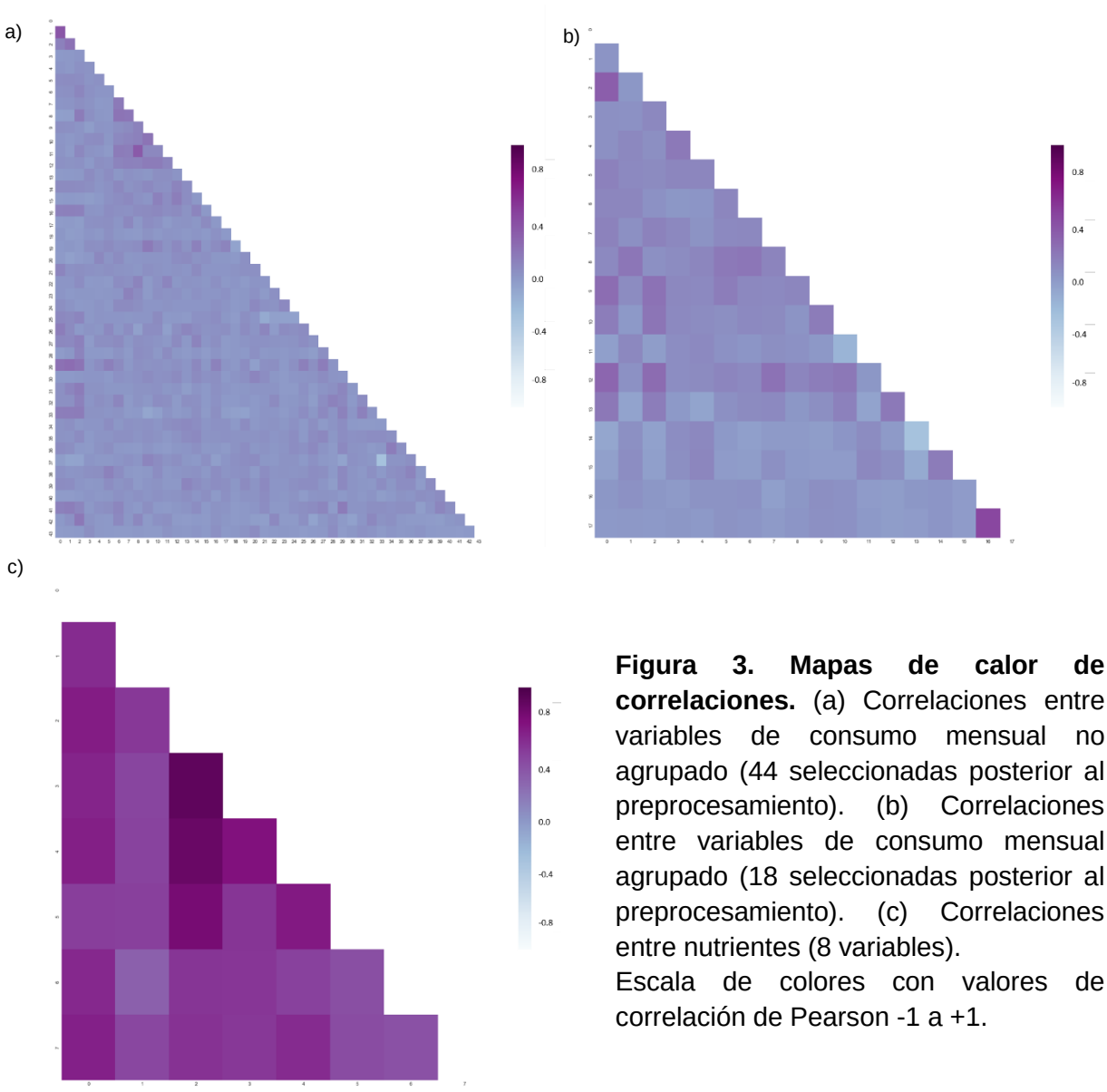


Figura 3. Mapas de calor de correlaciones. (a) Correlaciones entre variables de consumo mensual no agrupado (44 seleccionadas posterior al preprocesamiento). (b) Correlaciones entre variables de consumo mensual agrupado (18 seleccionadas posterior al preprocesamiento). (c) Correlaciones entre nutrientes (8 variables).
Escala de colores con valores de correlación de Pearson -1 a +1.

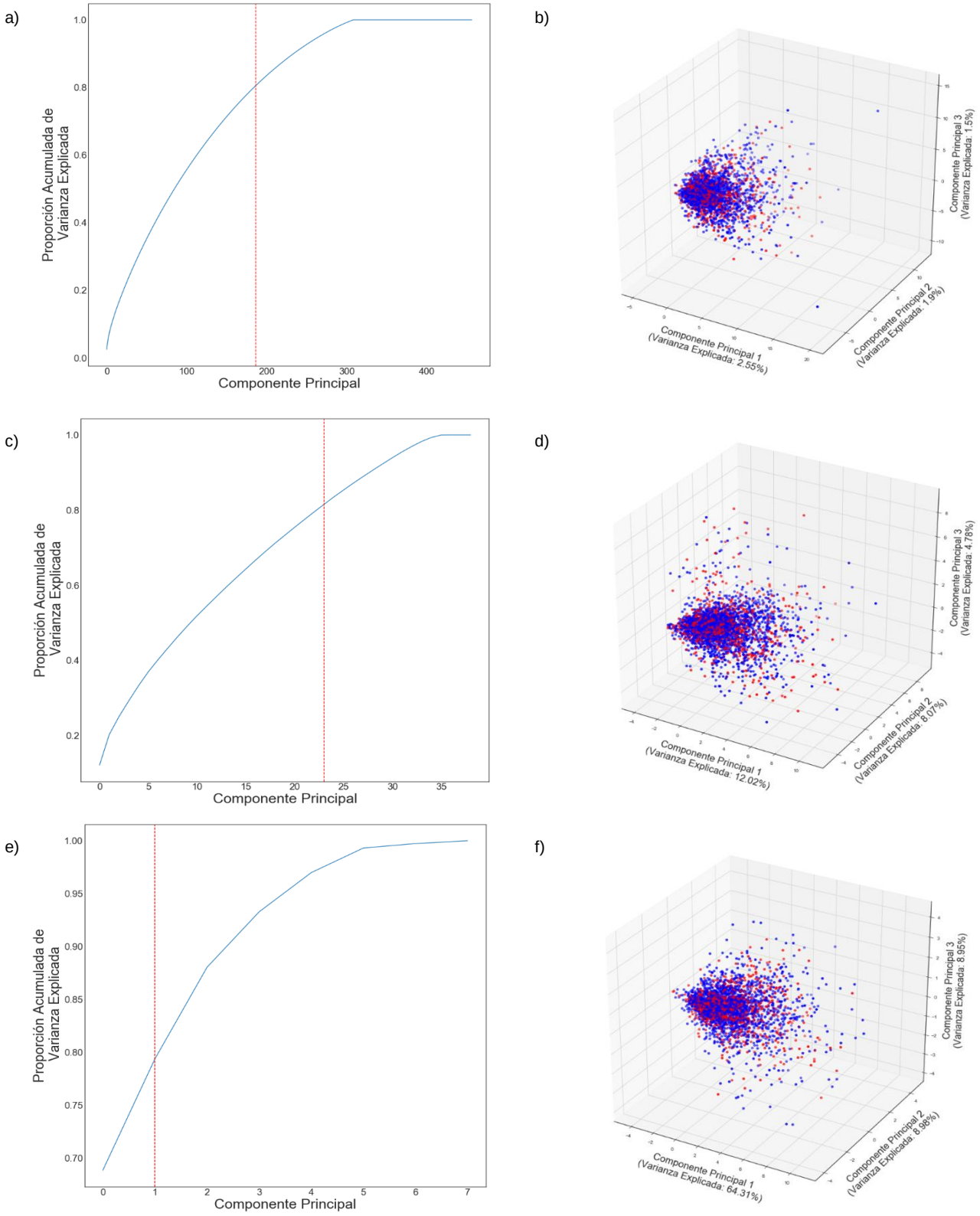


Figura 4. Análisis de Componentes Principales. (a) Gráfico sedimentación consumo mensual no agrupado*. (b) Gráfico 3 primeras componentes consumo mensual no agrupado (Rojo: Normal; Azul: Sobrepeso). (c) Gráfico sedimentación consumo mensual agrupado*. (d) Gráfico 3 primeras componentes consumo mensual agrupado. (e) Gráfico sedimentación de nutrientes*. (f) Gráfico 3 primeras componentes nutrientes. *Línea roja: Componente donde se alcanza 80% varianza explicada.

6.3 Predicción de Estado Nutricional

Se describen los resultados de la clasificación considerando el estado nutricional binario (Normal-Sobrepeso). En la Figura 5 se observan gráficos de cajas que representan la exactitud como métrica de comparación para cada una de las seis combinaciones posibles, 3 sets de datos de dieta (consumo mensual no agrupado, consumo mensual agrupado y nutrientes, con y sin datos demográficos) y para las variables sociodemográficas. Se observa que la mejor combinación son los datos de consumo mensual no agrupado con datos sociodemográficos, usando el algoritmo SVM (Exactitud (%): 72,2 ± 2,6).

Para los datos de consumo mensual agrupado, el mejor algoritmo es SVM, combinado con datos sociodemográficos (Exactitud (%): 70,6 ± 1,0). Para nutrientes, el mejor algoritmo también es SVM, con datos sociodemográficos (Exactitud (%): 68,8 ± 2,3). En el caso de las variables sociodemográficas, el XGB es el mejor algoritmo, con Exactitud (%): 66,7 ± 3,3.

En el caso del set de datos de nutrientes, este tiene un peor rendimiento que las variables sociodemográficas por sí solas, y al combinar ambas, el desempeño mejora. En general, cuando se incluyen las variables sociodemográficas como predictoras, hay mejor desempeño. No se alcanza significancia estadística entre los mejores algoritmos para ninguna comparación realizada (Anexo 4).

En el Anexo 5 se muestran las matrices de confusión que muestran la sobrestimación en el desempeño de los algoritmos. A modo de ejemplo, al comparar los resultados de SVM con XGB, los primeros muestran un significativo desbalance al asignar una gran cantidad de casos a la clase mayoritaria, lo que no ocurre de forma tan marcada en XGB. Finalmente, al comparar los resultados utilizando como métrica el F1, se identifican las mismas combinaciones de datos y algoritmos con mejor desempeño (Anexo 4).

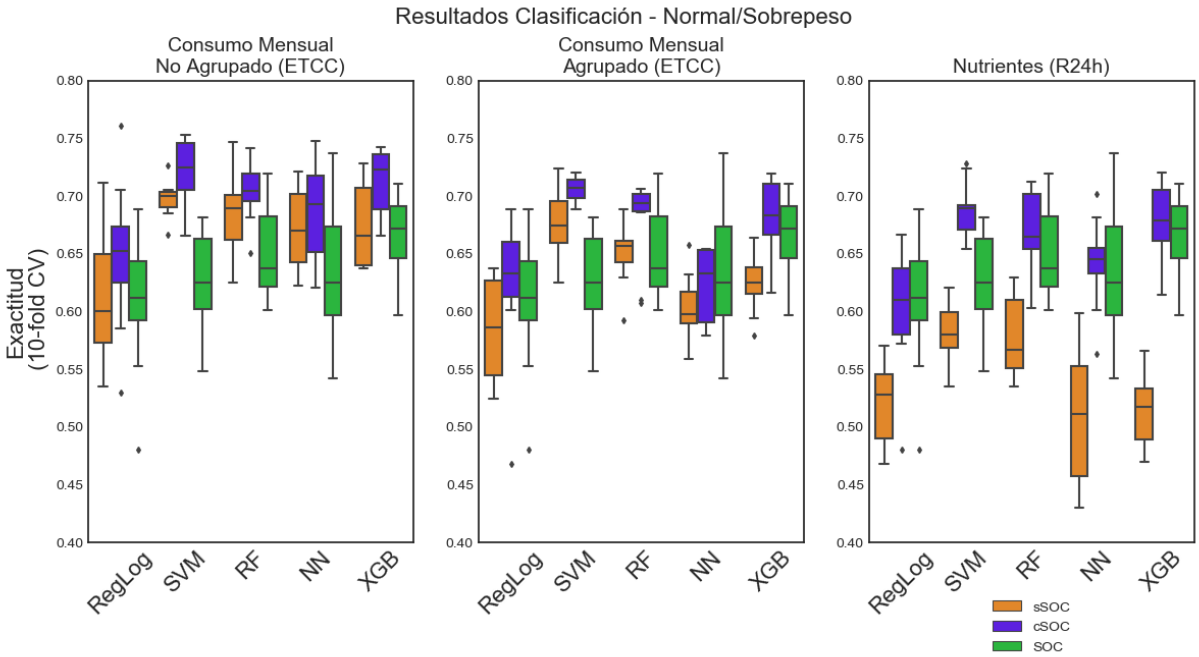


Figura 5. Clasificación 2 Clases. Se muestran gráficos de caja de la exactitud (%) de la implementación de algoritmos de clasificación (2 clases) separado por consumo mensual no agrupado, consumo mensual agrupado y nutrientes. sSOC: sin variables sociodemográficas; cSOC: con variables sociodemográficas; SOC: variables sociodemográficas. RegLog: Regresión Logística; SVM: *Support Vector Machine*; RF: *Random Forest*; NN: *Neural Networks*; XGB: *XGBoost*.

En el caso de la clasificación considerando tres clases para el estado nutricional (Normal, Sobrepeso y Obeso) como la variable a predecir, se observan en la Figura 6 gráficos de cajas que representan la exactitud como métrica de comparación para cada una de las seis combinaciones posibles, 3 sets de datos de dieta (consumo mensual no agrupado, consumo mensual agrupado y nutrientes, con y sin datos demográficos) y para las variables sociodemográficas. Se observa que la mejor combinación son los datos de consumo mensual no agrupado con datos sociodemográficos, usando el método de RF (Exactitud (%): 46,7 ± 3,0).

Para los datos de consumo mensual agrupado, el mejor algoritmo es Regresión Logística, combinado con datos sociodemográficos (Exactitud (%): 45,3 ± 4,5). Para

nutrientes, el mejor algoritmo es SVM con datos sociodemográficos (Exactitud (%): 44,7 ± 2,5). Finalmente, para los datos socioeconómicos el mejor algoritmo también es SVM, Exactitud (%): 44,0 ± 2,6.

Nuevamente, los algoritmos en los que se incluyen las variables sociodemográficas como predictoras tienen mejor desempeño. No se alcanza significancia estadística entre los mejores modelos para todas las comparaciones realizadas (Anexo 6).

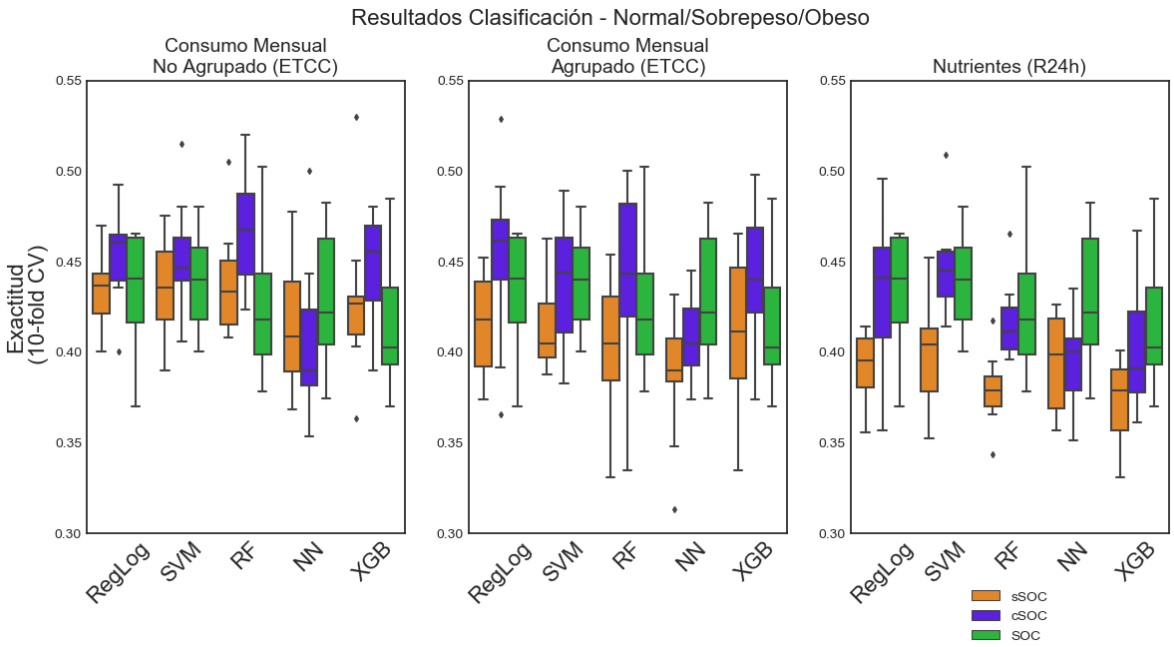


Figura 6. Clasificación Multiclase. Se muestran gráficos de caja de la exactitud (%) de implementación de algoritmos de clasificación (3 clases) separado por consumo mensual no agrupado, consumo mensual agrupado y nutrientes. sSOC: sin variables sociodemográficas; cSOC: con variables sociodemográficas; SOC: variables sociodemográficas. RegLog: Regresión Logística; SVM: *Support Vector Machine*; RF: *Random Forest*; NN: *Neural Networks*; XGB: *XGBoost*.

6.4 Predicción de IMC

En el caso de los algoritmos de regresión para la predicción del IMC (variable continua), se obtuvieron los siguientes resultados: para consumo mensual, el mejor

algoritmo de regresión es XGB, incluyendo datos sociodemográficos (RMSE: 5,2 ± 0,4). Para consumo mensual agrupado, el mejor algoritmo es Regresión Lineal con datos sociodemográficos (RMSE: 5,3 ± 0,4). Para nutrientes, el mejor algoritmo nuevamente es Regresión Lineal, con datos de sociodemográficos (RMSE: 5,3 ± 0,4). Finalmente, para los datos sociodemográficos, la Regresión Lineal tiene mejor desempeño, RMSE: 5,2 ± 0,4.

En la Figura 7 se muestran gráficos de cajas correspondientes a las diferentes combinaciones de datos y los algoritmos implementados. Se observa marcadamente que el peor desempeño es de la red neuronal (NN) en el set de datos de consumo mensual no agrupado.

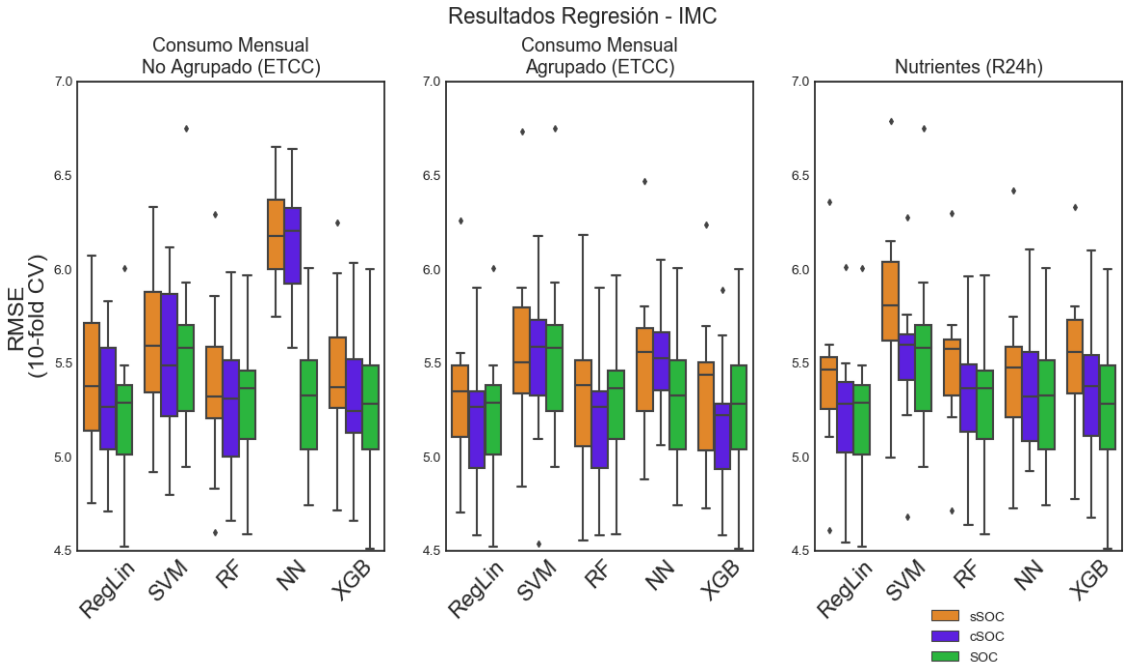


Figura 7. Predicción de IMC. Se muestran gráficos de caja del RMSE, separado por consumo mensual no agrupado, consumo mensual agrupado y nutrientes. sSOC: sin variables sociodemográficas; cSOC: con variables sociodemográficas; SOC: variables sociodemográficas. RegLog: Regresión Logística; SVM: *Support Vector Machine*; RF: *Random Forest*; NN: *Neural Networks*; XGB: *XGBoost*.

La Tabla 4 muestra el resumen de los mejores resultados por base de datos y método implementado (clasificación de 2 y 3 clases y regresión) y la comparación con ANOVA de estos resultados.

Tabla 4. Resumen de mejores resultados de clasificación y regresión.

	1. Consumo Mensual No Agrupado (ETCC)*	2. Consumo Mensual Agrupado (ETCC)*	3. Nutrientes (R24h)*	ANOVA (p-value < 0.05)
Clasificación – 2 clases (Exactitud %)	72,2 ± 2,6	70,6 ± 1,0	68,8 ± 2,3	1:2/1-3/3:2
Clasificación – 3 clases (Exactitud %)	46,7 ± 3,0	45,3 ± 4,5	44,7 ± 2,5	1:2:3
Regresión (RMSE)	5,3 ± 0,4	5,2 ± 0,4	5,3 ± 0,4	1:2:3

*Media ± Desviación Estándar; " – " representa diferencias significativas en esta comparación, " : " representa que no existen diferencias significativas en esta comparación.

6.5 Análisis Estratificados

Se realizó la estratificación de la población seleccionada en rangos etarios de 10 años y el género. Se implementaron los algoritmos de clasificación y regresión con el mejor desempeño para cada base de datos con la población estratificada por edad y género. Se reportan los resultados en la Tabla 5 y 6. En la Figura 8 se observan los gráficos de distribución de consumo de energía y gasto de energía estimado (EER) según los rangos etarios utilizados. Al realizar prueba estadística de comparación entre los grupos etarios, en ambos casos, consumo de energía y gasto de energía estimado, hay diferencias significativas (prueba ANOVA, p-value < 0.05) entre los grupos.

Tabla 5. Resultados de implementación de clasificación para población estratificada

		Consumo mensual no agrupado*	Consumo mensual agrupado*	Nutrientes*	Sociodemográfico*
Algoritmo		SVM	SVM	SVM	XGB
Población Total		72,2 ± 2,6	70,6 ± 1,0	68,8 ± 2,3	66,7 ± 3,3
19 – 28 años	Hombres	50,7 ± 14,0	52,1 ± 13,1	48,4 ± 10,4	46,9 ± 11,7
	Mujeres	60,3 ± 9,3	55,5 ± 7,6	53,7 ± 12,2	55,4 ± 9,9
29 – 38 años	Hombres	64,3 ± 9,5	71,9 ± 14,4	56,7 ± 19,0	59,4 ± 17,4
	Mujeres	52,9 ± 7,9	53,7 ± 11,2	46,6 ± 11,5	52,7 ± 16,8
39 – 48 años	Hombres	67,6 ± 8,4	53,1 ± 12,8	51,8 ± 11,5	61,3 ± 16,1
	Mujeres	66,6 ± 6,3	67,5 ± 9,4	61,1 ± 8,8	69,3 ± 7,4
49 – 58 años	Hombres	67,7 ± 10,9	55,9 ± 8,2	51,3 ± 8,7	59,4 ± 12,9
	Mujeres	66,3 ± 5,9	59,0 ± 10,0	59,6 ± 14,9	73,8 ± 9,0
59 – 64 años	Hombres	71,6 ± 15,0	65,8 ± 14,8	67,9 ± 20,3	58,8 ± 19,0
	Mujeres	70,6 ± 8,7	68,2 ± 10,0	65,0 ± 10,8	66,1 ± 18,6

*Exactitud (%), Media ± Desviación Estándar

Tabla 6. Resultados de implementación de regresión para población estratificada

		Consumo mensual no agrupado*	Consumo mensual agrupado*	Nutrientes*	Sociodemográfico*
Algoritmo		RegLin	XGB	RegLin	RegLin
Población Total		5,3 ± 0,4	5,2 ± 0,4	5,3 ± 0,4	5,2 ± 0,4
19 – 28 años	Hombres	5,1 ± 1,1	4,8 ± 0,5	4,6 ± 0,6	4,5 ± 1,0
	Mujeres	6,2 ± 0,6	6,1 ± 1,0	6,1 ± 0,9	5,9 ± 4,6
29 – 38 años	Hombres	4,9 ± 1,1	4,5 ± 1,0	4,5 ± 0,7	6,1 ± 4,6
	Mujeres	6,2 ± 0,8	5,9 ± 0,9	5,8 ± 1,07	5,8 ± 1,0
39 – 48 años	Hombres	4,7 ± 1,1	4,8 ± 1,1	4,8 ± 0,9	4,6 ± 0,8
	Mujeres	5,1 ± 0,5	5,4 ± 0,8	5,3 ± 0,8	5,2 ± 0,8
49 – 58 años	Hombres	4,7 ± 0,9	4,6 ± 1,1	4,4 ± 0,9	4,2 ± 0,7
	Mujeres	6,0 ± 1,3	5,7 ± 0,8	5,8 ± 0,9	5,7 ± 1,3
59 – 64 años	Hombres	5,2 ± 1,4	4,9 ± 1,2	4,7 ± 1,4	4,6 ± 1,4
	Mujeres	5,9 ± 1,6	6,1 ± 1,7	5,8 ± 1,7	6,1 ± 1,9

*RMSE, Media ± Desviación Estándar

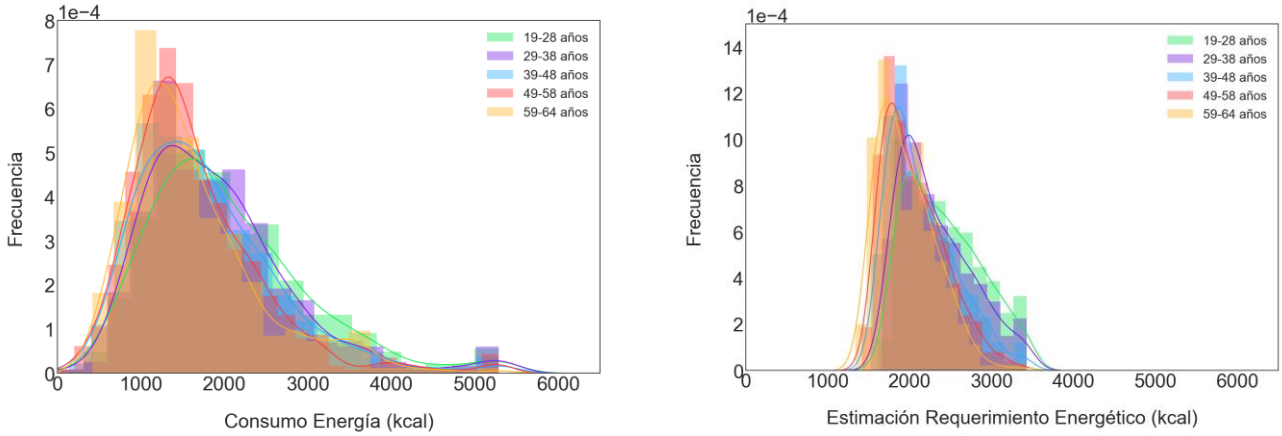


Figura 8. Consumo y Estimación de Gasto Energético. Gráfico distribuciones según rangos etarios para estimación del consumo de energía y la estimación del requerimiento energético.

6.6 Selección de Variables de Interés

6.6.1 Selección de Variables

El mejor algoritmo para clasificar en dos clases según el consumo mensual agrupado es SVM, sin embargo, la implementación realizada no permite obtener los coeficientes de importancia de cada variable. Considerando que no existen diferencias significativas con el algoritmo que le sigue en desempeño, XGB, y que para el caso de la regresión XGB es el algoritmo con mejor desempeño para el consumo mensual agrupado, se utilizó XGB para comparar las variables que tienen mayor importancia al momento de entrenar y predecir con este *set* de datos. En la Figura 9 se observa el grafico de barras de los 20 grupos de alimentos más importantes.

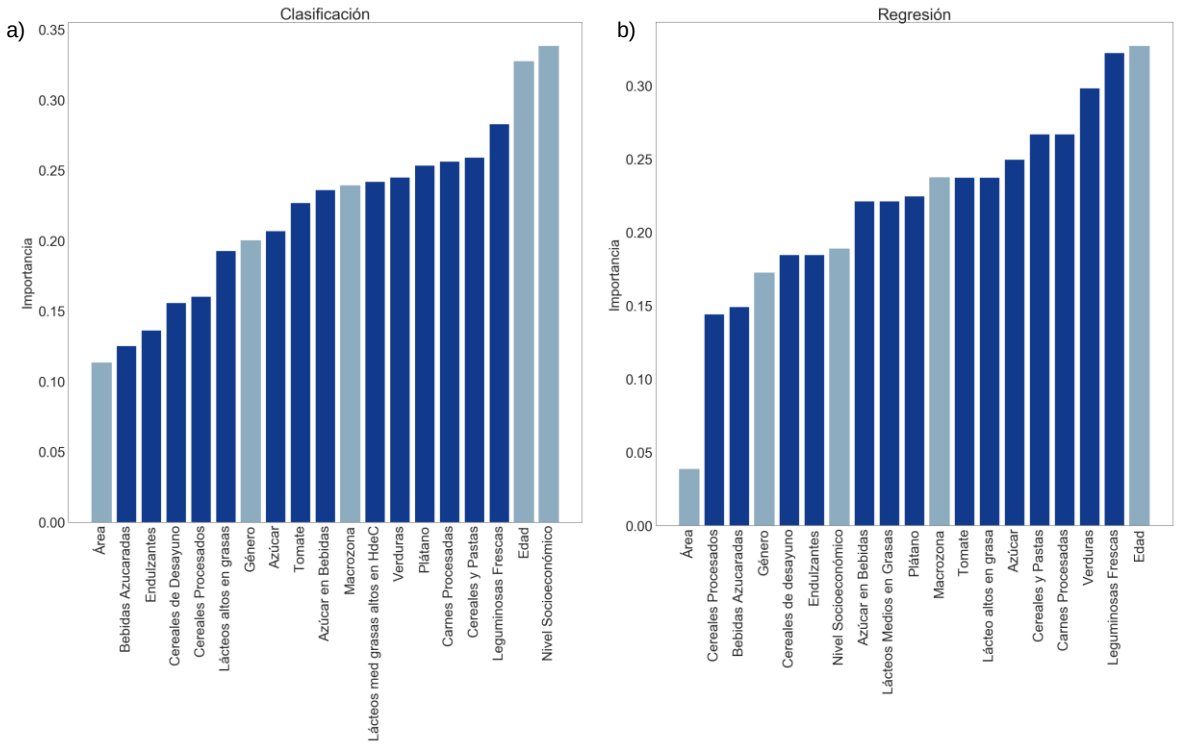


Figura 9. Variables de mayor importancia (XGB). Selección de 20 variables de mayor importancia determinadas por el algoritmo XGB, utilizando como predictores el consumo mensual agrupado y variables sociodemográficas. a) Método clasificación (2 clases). b) Método regresión.

6.6.2 Predictor con Variables Seleccionadas

Con las variables de consumo mensual agrupado seleccionadas (Anexo 3), se realizaron modelos de clasificación de 2 clases y regresión. En la Tabla 7 se observan los resultados de estas implementaciones comparados con la predicción realizada con todas las variables. Se observan diferencias significativas solo para el predictor entrenado con las variables de consumo mensual agrupado seleccionadas sin variables sociodemográficas, con el cual se obtiene un peor desempeño que al usar todas las variables. La Figura 10 muestra el gráfico de cajas de la comparación descrita, separando por los sets de datos con y sin datos sociodemográficos incluidos en el predictor. Se observa que en general, un predictor con variables seleccionadas, es decir menos información, tiene peor desempeño que con los sets de datos completos. Sin embargo, estas diferencias son significativas solo en el caso de la clasificación, al comparar el desempeño con los sets de datos sin las variables sociodemográficas. Se observa también que la inclusión de datos sociodemográficos en la predicción mejora el desempeño en el caso de la clasificación para la predicción con variables seleccionadas y la referencia (set de datos completo). Lo anterior no aplica en el caso de la regresión.

Tabla 7. Resultados de predicción con variables seleccionadas de consumo mensual agrupado comparado con predicción con todas las variables disponibles.

		Todas las Variables	Variables Seleccionadas	*p-value*
	Algoritmo		SVM	
Clasificación – 2 clases (Exactitud %)	Consumo Mensual Agrupado sSOC	69,7 ± 1,5	67,1 ± 1,5	0,009
	Consumo Mensual Agrupado cSOC	72,2 ± 2,6	70,5 ± 1,7	0,169
	Algoritmo		XGB	
Regresión (RMSE)	Consumo Mensual Agrupado sSOC	5,3 ± 0,4	5,4 ± 0,4	0,878
	Consumo Mensual Agrupado cSOC	5,2 ± 0,4	5,2 ± 0,4	0,386

*Media ± Desviación Estándar. sSOC: sin variables sociodemográficas; cSOC: con variables sociodemográficas.

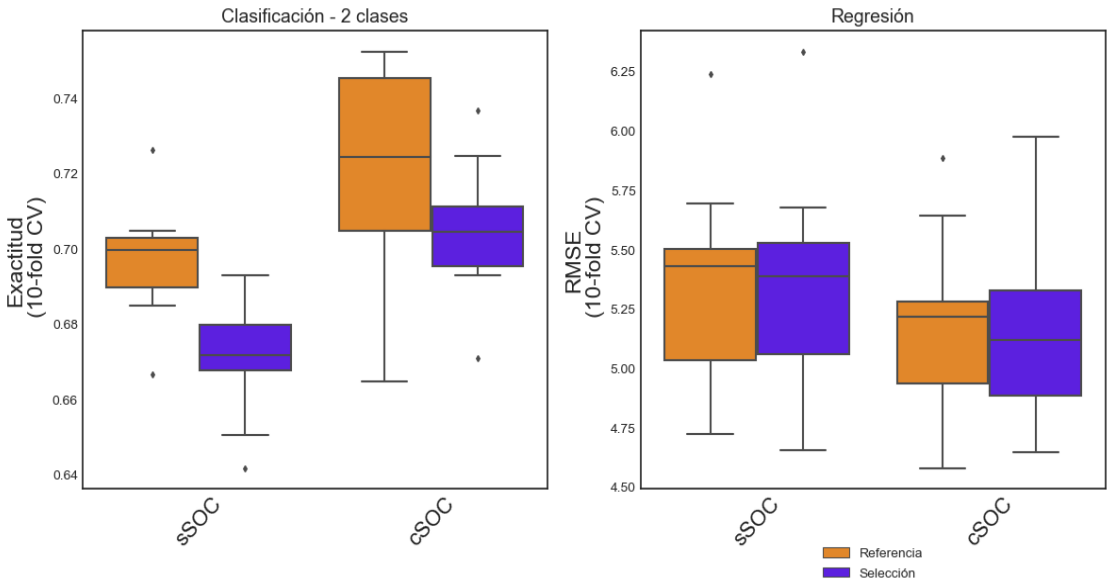


Figura 10. Comparación mejor desempeño con variables seleccionadas. Gráfico de cajas de comparación entre implementación de algoritmos con mejor desempeño con todas las variables de consumo mensual agrupado (Referencia) versus variables de consumo mensual agrupado seleccionadas (Selección). Variables seleccionadas se encuentran en Anexo 3. sSOC: sin variables sociodemográficas; cSOC: con variables sociodemográficas

6.7 Efecto de la Actividad Física en Modelos de Predicción

En la Figura 11 se observan las distribuciones de esta variable, considerando las categorías de estado nutricional en dos clases (Normal y Sobrepeso) y en tres clases (Normal, Sobrepeso y Obeso) y los porcentajes de actividad física reportada según la agrupación de estado nutricional descrita anteriormente.

En la Figura 12 se muestra la comparación entre la mejor combinación de algoritmo y *set* de datos para la clasificación de dos clases y para la regresión, utilizando el mismo *set* de datos, incluyendo la variable de actividad física. Las diferencias entre estas comparaciones no son significativas para la clasificación (test Wilcoxon, *p-value* = 0,823) ni para la regresión (test Wilcoxon, *p-value* = 0,333).

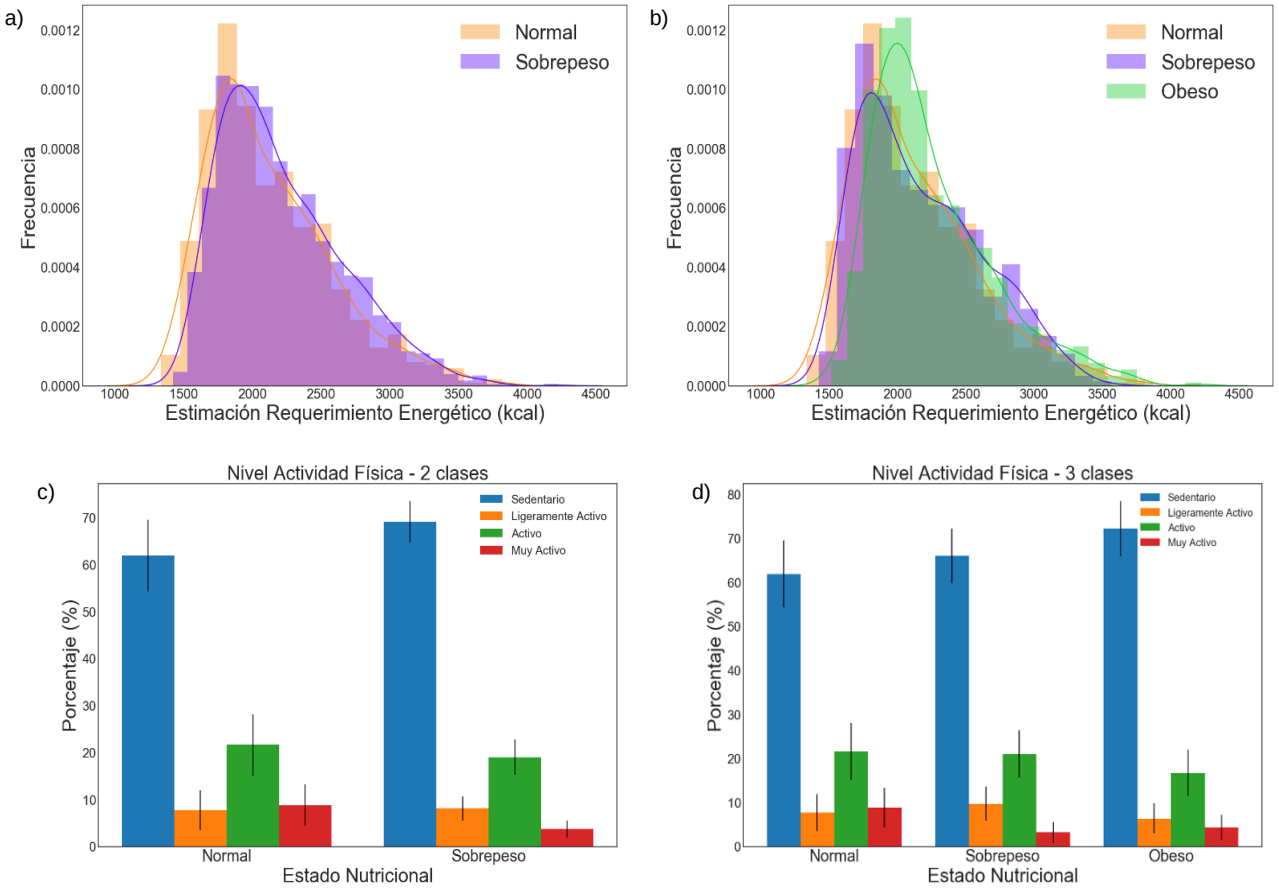


Figura 11. Descripción variables de actividad. a) Gráfico distribución de EER para población (2 clases). b) Gráfico distribución de EER para población (3 clases). c) Gráficos de porcentaje de población según nivel de actividad (2 clases). d) Gráficos de porcentaje de población según nivel de actividad (3 clases).

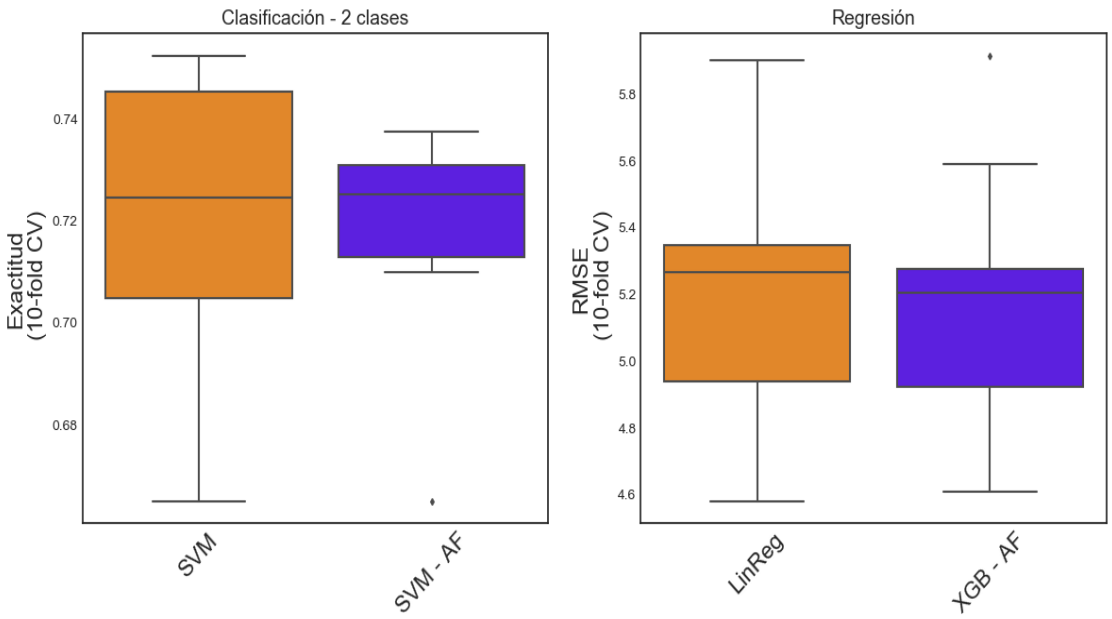


Figura 12. Comparación mejor desempeño con variables de actividad física. a) Comparación de mejores predictores con y sin variable de actividad física como predictora en clasificación de 2 clases. b) Comparación de mejores predictores con y sin variable de actividad física como predictora en regresión.

7. DISCUSIÓN

Los resultados obtenidos, considerando diferentes *sets* de datos, referentes a la dieta, variables sociodemográficas y actividad física, y la implementación de diferentes algoritmos, muestran que, si bien se puede predecir el estado nutricional en base a lo que comen las personas, el desempeño de estos predictores es insuficiente para alcanzar el valor que se propuso en la hipótesis de esta tesis (90% de exactitud), obteniendo, en el mejor de los casos, un 72,2%. Lo anterior puede tener varias explicaciones, las cuales se discuten a continuación.

Al comparar las diferentes bases de datos de consumo de alimentos que se utilizaron en esta tesis, los mejores resultados se obtuvieron a partir del consumo mensual de alimentos no agrupados combinado con variables sociodemográficas, con el algoritmo SVM (72,2% de exactitud). Sin embargo, a diferencia de la base de datos que utiliza variables procesadas para reportar los nutrientes que consumen las personas en un día, las bases de datos de consumo mensual agrupado y no agrupado se trabajaron en base al consumo en cantidades (gramos o mililitros) y la frecuencia reportada. También se debe considerar que, si bien algunos algoritmos muestran mejor desempeño que otros, en general estas diferencias no son significativas en el caso de este trabajo, comportamiento que se mantiene inclusive en las otras implementaciones metodológicas realizadas (clasificación de 3 clases y regresión), lo que se puede interpretar como cierta estabilidad en el desempeño, independiente del algoritmo que se implemente. Lo mencionado da indicios de que las limitaciones halladas tienen que ver con características intrínsecas de los datos. Otro factor a considerar es que, si bien las métricas resumen de buena forma el desempeño de los algoritmos, al explorar con detalle las matrices de confusión cuando existe desbalance de clases, como es el caso para la clasificación de 2 clases, donde el grupo “Sobrepeso” dobla al grupo “Normal”, las métricas pueden estar presentando buenos resultados debido a una sobrestimación de la clase mayoritaria[90].

A partir de la revisión en detalle de la documentación que provee la Encuesta Nacional de Consumo de Alimentos (ENCA, Informe Final[82]), los autores de esta describen las limitaciones metodológicas de la encuesta en sí y de los instrumentos utilizados para caracterizar la dieta, cuyo propósito principal es caracterizar el consumo de alimentos a nivel poblacional. En general, los datos que se obtienen desde los instrumentos para caracterizar la dieta (cuestionarios de frecuencia de consumo y recordatorio de 24 horas) están sujetos al sesgo de los sujetos[36]. Becker *et al*. reconocen las limitaciones de este tipo de instrumentos: en las encuestas recordatorias de 24 horas, las personas tienden a subestimar lo que consumen, mientras que encuestadas por su consumo de alimentos en un periodo de tiempo más largo tienden a sobreestimarlo[95]. Se tiene conocimiento con mayor certeza sobre el *misreporting* de energía más que de en el caso del reporte de comidas o nutrientes directamente. Para detectar estos casos, se utilizan biomarcadores o estimaciones de referencia del gasto energético, con umbrales de corte definidos para determinar el sub o sobre-reporte. El sub-reporte está asociado consistentemente a varias características personales, además del IMC en sí. También se debe considerar que existen ciertos alimentos más susceptibles al sub-reporte. Considerando esto, se constatan limitaciones en la calidad de los datos, relacionados con el sesgo, que alimentan los modelos predictivos implementados en esta tesis, por lo que los resultados deben ser interpretados considerando esto.

Por otro lado, al considerar la obesidad desde la perspectiva biopsicosocial[96,97], la caracterización y modelamiento de esta tiene igual o mayor dificultad que otros problemas de relevancia epidemiológica. En este sentido, dentro de las premisas de esta tesis se postulaba implícitamente que los algoritmos de análisis de AM serían capaces de reflejar esta complejidad de interacciones considerando solo los datos de dieta de las personas. Ya sea por las características de los datos o la complejidad del problema, esto se logró de forma parcial, ya que al incluir variables sociodemográficas como predictoras, los algoritmos mejoraron su capacidad predictiva del estado nutricional. Esta mejora en los resultados concuerda con

trabajos anteriores que describen cómo variables sociodemográficas, tales como el nivel socioeconómico y el género, se asocian a la obesidad[98–100].

En este mismo contexto, la dieta es solo uno de los factores que influye en el exceso de adiposidad en las personas. La causa de obesidad definida por la OMS hace referencia al desbalance energético, donde el consumo de calorías es mayor al gasto energético[1], y, si bien la dieta está relacionada con este fenómeno, no es el único factor que lo afecta. En este sentido, con los datos disponibles, la única aproximación al gasto energético que se puede obtener de la ENCA es la EER (*Estimated Energy Requirement)*, la cual se calcula en base a las características de las personas (edad, género, talla y peso) y al nivel de actividad física que realizan[92,101]. Sin embargo, al igual que los datos de dieta, estos datos son reportados por los individuos mediante cuestionarios, es decir, no son mediciones directas del gasto energético real[82]. A raíz de esto, no es de extrañar que la inclusión de las variables de actividad física no mejore las predicciones del estado nutricional, a pesar de que la definición de obesidad haga referencia a este factor.

En general, el sesgo en los datos de fuentes debido al instrumento, limitaciones del auto reporte y la complejidad de la condición de obesidad, se puede enmarcar en que, si bien se cuenta con una gran cantidad de información de representatividad poblacional de lo que consumen las personas, estos datos representan una medición observacional en un momento determinado (transversal), la cual no es suficiente para determinar de forma certera la causalidad de la condición que se está explorando[102,103]. En esta tesis se exploraron métodos predictivos del estado nutricional exclusivamente en base a datos de consumo de alimentos reportados para una población representativa de Chile, y solo limitándose a ese nivel de análisis, los resultados son insuficientes como para determinar si la dieta causa sobrepeso u obesidad o si las personas con sobrepeso u obesidad consumen ciertos alimentos en cantidades características. En base al análisis realizado y con los datos disponibles para esta tesis, no parece existir un patrón (qué alimentos y en qué cantidades)

diferenciador entre las personas con estado nutricional normal versus las personas con sobrepeso u obesidad.

Al generar implementaciones estratificadas por edad y género, se puede observar una tendencia a mejora de la exactitud de los modelos a medida que aumenta la edad. Por otra parte, se pueden apreciar diferencias en el gráfico de distribución de gasto energético por grupos etarios, donde los estratos de menor edad mayor cantidad de sujetos que gastan más[104]. Sin embargo, los modelos de clasificación generados no alcanzan mejores resultados de predicción que los que considera toda la población, excepto para el caso de mujeres entre 49 y 58 años con variables sociodemográficas como predictores y con el algoritmo XGB. En cambio, en el caso de la regresión, en general, los estratos de género masculino muestran mejores resultados (menor RMSE) para la predicción del IMC.

Al obtener las 20 variables más influyentes de los modelos predictivos con mejor desempeño, los cuales todos incluyen variables de dieta y variables sociodemográficas, se observa que todas las variables sociodemográficas, tanto para la clasificación como para la regresión, se encuentran presentes dentro de estas 20, lo que coincide con lo ya mencionado: el desempeño mejora al incluir estas variables como predictoras. Dentro de estas 20 variables se encuentran, además, para el caso de la clasificación (lugar que ocupan en el ranking): cereales y pastas (4), carnes procesadas (5), azúcar en bebidas (10), azúcar (12), cereales procesados (15), cereales de desayuno (16), endulzantes (17) y bebidas azucaradas (18). En el caso de la regresión: carne procesada (4), cereales y pastas (5), azúcar (6), lácteos altos en grasas (7), azúcar en bebidas (12), endulzantes (14) y bebidas azucaradas (17). Al implementar modelos predictivos con variables seleccionadas que se asocian con dietas altas en azúcar, se observa que la predicción del estado nutricional y el IMC no mejoran ni empeoran, lo que se puede interpretar como que, si bien no son mejores predictores, si son capaces de clasificar y predecir de igual forma que el *set* completo de variables.

Si bien la hipótesis de esta tesis se considera rechazada al no alcanzar el 90% de exactitud con ningún algoritmo ni *set* de datos, al comparar este trabajo con el estudio de metodología similar y comparable, publicado por Selya y Anshutz[35] (*Advanced Data Analytics in Health*), revela que los resultados obtenidos en esta tesis son similares. Los autores de este artículo muestran la implementación de clasificadores (análisis de discriminantes, SVM y NN) para predecir la obesidad de una población específica (mujeres blancas, no hispánicas, mayores de 21 años), con datos obtenidos de la *National Health and Nutrition Examination Survey* en USA (NHANES) referentes a dieta, actividad física y la edad durante los años 2011 y 2013. Para este estudio, la información de la dieta es obtenida de un cuestionario en el cual se les pregunta a las personas por lo consumido (alimentos y bebidas) en las 24 horas anteriores, con lo cual se calcula el contenido nutricional de todo lo consumido en el periodo. El mejor clasificador que los autores obtuvieron alcanza una exactitud cercana al 59%, y consideran este caso como el más exitoso, ya que en la matriz de confusión del clasificador se observa que se logra diferenciar cada una de las categorías que pretenden predecir (*"Obese" / "Not obese"*) por sobre el azar (más de 50%), a diferencia de todos los otros algoritmos implementados por ellos, los cuales muestran predicciones sesgadas a la clase mayoritaria. En los resultados presentados en esta tesis, si bien la exactitud es mayor, las matrices de confusión (Anexo 5) muestran el mismo fenómeno de sesgo, debido al desbalance de las clases. Este desbalance es característico de los datos en estudio (72,7% clase sobrepeso versus 27,3% normal) y si bien se utilizaron técnicas de balance, estas no lo compensan, lo que aporta un indicio adicional a lo anteriormente mencionado, que los datos no son suficientes para diferenciar entre ambas clases. A pesar de estas limitaciones, los autores consideran que los clasificadores pueden predecir con seguridad el riesgo de ciertas condiciones de salud (como la obesidad), comparados con regresiones convencionales. Destacan, además, que el uso de clasificadores para predecir obesidad confirma que la combinación de los componentes de la dieta tiene poder explicativo sobre la obesidad mayor al que puede ser alcanzado con la suma de los componentes, lo cual valida las dificultades que se encuentran en la investigación actual de los determinantes nutricionales de la obesidad (es decir, que

la asociación entre macronutrientes individuales y la obesidad es inconsistente y de efectos débiles).

Finalmente, las proyecciones de esta tesis son variadas. Desde el punto de vista metodológico, se podrían realizar implementaciones más exhaustivas de métodos y algoritmos para descubrir patrones alimentarios o grupos de alimentos que sean mejores predictores del estado nutricional. Por otro lado, explorar la implementación de predictores posterior a realizar la descomposición nutricional de los alimentos y estimar el consumo de los nutrientes en base al reporte de frecuencia de consumo mensual podría resultar en modelos con mejor desempeño, considerando que el consumo mensual debiera ser más representativo de lo que comen habitualmente los individuos, y, por lo tanto, predecir de mejor forma el estado nutricional. Otra limitación importante, ya mencionada, es que se cuenta con datos de reportes transversales. Una propuesta interesante sería investigar métodos para simular datos de dieta y así obtener patrones de consumo de alimentos que predigan obesidad con mejor exactitud, contrastando eventualmente los datos simulados con los reales. Por otro lado, considerando la identificación de las variables (ya sean alimentos, grupos de alimentos o nutrientes) que muestran mayor influencia en la predicción del estado nutricional, sería provechoso utilizar estas variables para desarrollar estudios de seguimiento de cohortes prospectivos de personas con patrones similares y característicos, como es el caso de la cohorte Seguimiento Universidad de Navarra (SUN), enfocado en Dieta Mediterránea[105], con el fin de establecer causalidad con variables de dieta más acotadas.

8. CONCLUSIÓN

Basándose en los resultados de esta tesis, la hipótesis propuesta es rechazada, ya que ninguna combinación de datos predictivos y algoritmos implementados alcanzo el 90% de exactitud, llegando en el mejor de los casos al 72,2%. Al comparar los resultados de este trabajo con el estudio de metodología equivalente, los resultados son similares.

El autorreporte de consumo de alimentación y actividad física de una población en cierto momento tiene características y limitaciones determinadas, las cuales condicionan la implementación y desempeño de los métodos de análisis. En este caso, a partir del autorreporte transversal no es posible determinar causalidad entre las variables predictoras y la variable respuesta (estado nutricional). Si bien, de los resultados obtenidos no se puede establecer causalidad, sí se pueden obtener directrices en cuanto a cuáles variables son relevantes para predecir la obesidad con este set de datos y métodos implementados, que en conjunto proporcionen información para el diseño de futuras investigaciones con métodos que sí permitan establecer causalidad.

El estudio de la dieta como determinante de la obesidad aún se encuentra en desarrollo, y el enfoque que utiliza metodologías de AM presentado en esta tesis es una de las formas de abordarlo. Un modelo que logre clasificar el estado nutricional basado en el consumo de alimentos tiene varias aplicaciones potenciales en el contexto de la vida diaria de las personas, considerando que el uso masivo de dispositivos móviles permite recolección instantánea de información, lo que facilita obtener datos que alimenten estos modelos y que informen a los sujetos de su riesgo de tener sobrepeso u obesidad en base a lo que consumen. En este escenario, obtener modelos precisos que logren predecir estados o condiciones de salud determinados se vuelve esencial y el uso de los métodos para generarlos debe ser realizado de forma responsable, considerando sus posibles implicancias en la salud de las personas.

9. BIBLIOGRAFÍA

1. World Health Organization. WHO | Obesity and overweight. WHO. http://www.who.int/mediacentre/factsheets/fs311/en/. Published 2017. Accessed July 16, 2017.
2. Ministerio de Salud Chile. *Norma Para La Evaluacion Nutricional de Niños, Niñas y Adolescentes de 5 Años a 19 Años de Edad*. Santiago, Chile; 2016.
3. Ministerio de Salud Chile, Organización Panamericana de la Salud, Organización Mundial de la Salud. *Referencia OMS Para La Evaluacion Antropometrica En Niños y Niñas Menores de 6 Años*.; 2006.
4. Schwartz MW, Seeley RJ, Zeltser LM, et al. Obesity Pathogenesis: An Endocrine Society Scientific Statement. *Endocr Rev*. 2017;38(June 2017):267-296. doi:10.1210/er.2017-00111.
5. Ludwig D, Ebbeling C. The carbohydrate-insulin model of obesity: Beyond “calories in, calories out.” *JAMA Intern Med*. July 2018. http://dx.doi.org/10.1001/jamainternmed.2018.2933.
6. World Health Organization. WHO | Noncommunicable diseases. WHO. http://www.who.int/mediacentre/factsheets/fs355/en/. Published 2017. Accessed July 15, 2017.
7. Swinburn BA, Sacks G, Hall KD, et al. The global obesity pandemic: shaped by global drivers and local environments. *Lancet*. 2011;378(9793):804-814. doi:10.1016/S0140-6736(11)60813-1.
8. Popkin BM. *The World Is Fat : The Fads, Trends, Policies, and Products That Are Fattening the Human Race*. Avery; 2009.
9. Popkin BM. Contemporary nutritional transition: determinants of diet and its impact on body composition. *Proc Nutr Soc*. 2012;70(1):82-91. doi:10.1017/S0029665110003903.Contemporary.
10. Bedregal P, Margozzini P, González C. *Informe Final: Estudio de Carga de Enfermedad y Carga Atribuible*.; 2008. http://www.cienciasdelasalud-udla.cl/portales/tp76246caadc23/uploadImg/File/Informe-final-carga-Enf-2007.pdf.
11. Ministerio de Salud Chile, Departamento de Epidemiología. *Reporte de Vigilancia de Enfermedades No Transmisibles (ENT)*.; 2011. https://www.paho.org/hq/dmdocuments/2012/ENT-I-Reporte-Vigilancia-2011.pdf. Accessed July 11, 2018.

12. Escobar MC, Báez L, Cozzaglio M, et al. *ENFERMEDADES NO TRANSMISIBLES*.; 2013. http://www.redcronicas.cl/wrdprss_minsal/wp-content/uploads/2014/04/Enfermedades-no-Transmisibles-en-Chile-2013.pdf. Accessed July 11, 2018.
13. Ministerio de Salud Chile, Departamento de Epidemiología. *ENCUESTA NACIONAL DE SALUD 2016-2017 Segunda Entrega de Resultados*.; 2018. http://www.minsal.cl/wp-content/uploads/2017/11/ENS-2016-17_PRIMEROS-RESULTADOS.pdf. Accessed July 10, 2018.
14. INTA. Encuesta Nacional de Salud 2016-2017: Obesidad y falta de conciencia de la sociedad chilena | INTA. http://inta.cl/encuesta-nacional-de-salud-2016-2017-obesidad-y-falta-de-conciencia-de-la-sociedad-chilena/. Accessed July 10, 2018.
15. Ministerio de Salud Chile, P. Universidad Católica de Chile, Universidad Alberto Hurtado. *Encuesta Nacional de Salud ENS Chile 2009-2010*.; 2014. http://www.minsal.gob.cl/portal/url/item/bcb03d7bc28b64dfe040010165012d23.pdf.
16. Butland B, Jebb S, Kopelman P, et al. *Foresight Tackling Obesities : Future Choices – Project Report*.; 2007. Goverment Office for Science, UK.
17. Garawi F, Devries K, Thorogood N, Uauy R. Global differences between women and men in the prevalence of obesity: is there an association with gender inequality? *Eur J Clin Nutr*. 2014;68(10):1101-1106. doi:10.1038/ejcn.2014.86.
18. Kanter R, Caballero B. Global Gender Disparities in Obesity: A Review. *Adv Nutr*. 2012;3(4):491-498. doi:10.3945/an.112.002063.
19. Chooi YC, Ding C, Magkos F. The epidemiology of obesity. *Metabolism*. 2019;92:6-10. doi:10.1016/J.METABOL.2018.09.005.
20. Baum CL, Ruhm CJ. Age, socioeconomic status and obesity growth. *J Health Econ*. 2009;28(3):635-648. doi:10.1016/J.JHEALECO.2009.01.004.
21. Rao M, Afshin A, Singh G, Mozaffarian D. Do healthier foods and diet patterns cost more than less healthy options? A systematic review and meta-analysis. *BMJ Open*. 2013;3(12):e004277. doi:10.1136/bmjopen-2013-004277.
22. Northstone K, Emmett PM. Dietary patterns of men in ALSPAC: associations with socio-demographic and lifestyle characteristics, nutrient intake and comparison with women's dietary patterns. *Eur J Clin Nutr*. 2010;64(9):978-986. doi:10.1038/ejcn.2010.102.
23. Estaquio C, Druesne-Pecollo N, Latino-Martel P, Dauchet L, Hercberg S, Bertrais S. Socioeconomic Differences in Fruit and Vegetable Consumption among Middle-Aged

French Adults: Adherence to the 5 A Day Recommendation. *J Am Diet Assoc*. 2008;108(12):2021-2030. doi:10.1016/J.JADA.2008.09.011.

24. Malon A, Deschamps V, Salanave B, et al. Compliance with French Nutrition and Health Program Recommendations Is Strongly Associated with Socioeconomic Characteristics in the General Adult Population. *J Am Diet Assoc*. 2010;110(6):848-856. doi:10.1016/j.jada.2010.03.027.
25. Lallukka T, Laaksonen M, Rahkonen O, Roos E, Lahelma E. Multiple socio-economic circumstances and healthy food habits. *Eur J Clin Nutr*. 2007;61(6):701-710. doi:10.1038/sj.ejcn.1602583.
26. Harrington J, Fitzgerald AP, Layte R, Lutomski J, Molcho M, Perry IJ. Sociodemographic, health and lifestyle predictors of poor diets. *Public Health Nutr*. 2011;14(12):2166-2175. doi:10.1017/S136898001100098X.
27. McNaughton SA, Ball K, Crawford D, Mishra GD. An Index of Diet and Eating Patterns Is a Valid Measure of Diet Quality in an Australian Population. *J Nutr*. 2008;138(1):86-93. doi:10.1093/jn/138.1.86.
28. Darmon N, Ferguson EL, Briend A. A Cost Constraint Alone Has Adverse Effects on Food Selection and Nutrient Density: An Analysis of Human Diets by Linear Programming. *J Nutr*. 2002;132(12):3764-3771. doi:10.1093/jn/132.12.3764.
29. Darmon N, Ferguson E, Briend A. Do economic constraints encourage the selection of energy dense diets? *Appetite*. 2003;41(3):315-322. http://www.ncbi.nlm.nih.gov/pubmed/14637330. Accessed December 31, 2018.
30. Drewnowski A, Rehm CD, Arterburn D. The geographic distribution of obesity by census tract among 59 767 insured adults in King County, WA. *Int J Obes (Lond)*. 2014;38(6):833-839. doi:10.1038/ijo.2013.179.
31. Ross R, Janssen I. Physical activity, total and regional obesity: dose-response considerations. *Med Sci Sports Exerc*. 2001;33(6 Suppl):S521-7; discussion S528-9. http://www.ncbi.nlm.nih.gov/pubmed/11427779. Accessed December 31, 2018.
32. Levine JA, Lanningham-Foster LM, McCrady SK, et al. Interindividual Variation in Posture Allocation: Possible Role in Human Obesity. *Science (80-)*. 2005;307(5709):584-586. doi:10.1126/science.1106561.
33. Church TS, Thomas DM, Tudor-Locke C, et al. Trends over 5 Decades in U.S. Occupation-Related Physical Activity and Their Associations with Obesity. Lucia A, ed. *PLoS One*. 2011;6(5):e19657. doi:10.1371/journal.pone.0019657.
34. Walker TB, Parker MJ. Lessons from the war on dietary fat. *J Am Coll Nutr*.

2014;33(4):347-351. doi:10.1080/07315724.2013.870055.

35. Selya A, Anshutz D. Machine Learning for the Classification of Obesity from Dietary and Physical Activity Patterns. In: Giabbanelli P, Papageorgiou E, Mago V, eds. *Advanced Data Analytics in Health*. Vol 93. 1st ed. Switzerland: Springer; 2018:77-97. doi:10.1007/978-3-319-77911-9.
36. Beechy L, Galpern J, Petrone A, Das SK. Assessment tools in obesity — Psychological measures, diet, activity, and body composition. *Physiol Behav*. 2012;107(1):154-171. doi:10.1016/J.PHYSBEH.2012.04.013.
37. Moussavi N, Gavino V, Receveur O. Could the Quality of Dietary Fat, and Not Just Its Quantity, Be Related to Risk of Obesity? *Obesity*. 2008;16(1):7-15. doi:10.1038/oby.2007.14.
38. van Dam RM, Seidell JC. Carbohydrate intake and obesity. *Eur J Clin Nutr*. 2007;61(S1):S75-S99. doi:10.1038/sj.ejcn.1602939.
39. Jacobs DR, Steffen LM. Nutrients, foods, and dietary patterns as exposures in research: a framework for food synergy. *Am J Clin Nutr*. 2003;78(3):508S-513S. doi:10.1093/ajcn/78.3.508S.
40. Ioannidis J. The Challenge of Reforming Nutritional Epidemiologic Research. *J Am Med Assoc*. 2018;320. doi:10.1038/srep26983.
41. Arvaniti F, Panagiotakos DB. Healthy Indexes in Public Health Practice and Research: A Review. *Crit Rev Food Sci Nutr*. 2008;48(4):317-327. doi:10.1080/10408390701326268.
42. Fransen HP, Ocké MC. Indices of diet quality. *Curr Opin Clin Nutr Metab Care*. 2008;11(5):559-565. doi:10.1097/MCO.0b013e32830a49db.
43. Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol*. 2002;13(1):3-9. http://www.ncbi.nlm.nih.gov/pubmed/11790957. Accessed March 27, 2019.
44. Gao SK, Beresford SA, Frank LL, Schreiner PJ, Burke GL, Fitzpatrick AL. Modifications to the Healthy Eating Index and its ability to predict obesity: the Multi-Ethnic Study of Atherosclerosis. *Am J Clin Nutr*. 2008;88(1):64-69. doi:10.1093/ajcn/88.1.64.
45. Guo X, Warden BA, Paeratakul S, Bray GA. Healthy Eating Index and obesity. *Eur J Clin Nutr*. 2004;58(12):1580-1586. doi:10.1038/sj.ejcn.1601989.
46. Lassale C, Fezeu L, Andreeva VA, et al. Association between dietary scores and 13-year weight change and obesity risk in a French prospective cohort. *Int J Obes*.

2012;36(11):1455-1462. doi:10.1038/ijo.2011.264.

47. Quatromoni PA, Pencina M, Cobain MR, Jacques PF, D'Agostino RB. Dietary Quality Predicts Adult Weight Gain: Findings from the Framingham Offspring Study*. *Obesity*. 2006;14(8):1383-1391. doi:10.1038/oby.2006.157.
48. Romaguera D, Norat T, Vergnaud A-C, et al. Mediterranean dietary patterns and prospective weight change in participants of the EPIC-PANACEA project. *Am J Clin Nutr*. 2010;92(4):912-921. doi:10.3945/ajcn.2010.29482.
49. Sares-Jäske L, Knekt P, Lundqvist A, Heliövaara M, Männistö S. Dieting attempts modify the association between quality of diet and obesity. *Nutr Res*. 2017;45:63-72. doi:10.1016/j.nutres.2017.08.001.
50. Buckland G, Bach A, Serra-Majem L. Obesity and the Mediterranean diet: a systematic review of observational and intervention studies. *Obes Rev*. 2008;9(6):582-593. doi:10.1111/j.1467-789X.2008.00503.x.
51. Togo P, Osler M, Sørensen T, Heitmann B. Food intake patterns and body mass index in observational studies. *Int J Obes*. 2001;25(12):1741-1751. doi:10.1038/sj.ijo.0801819.
52. Kipnis V, Midthune D, Freedman L, et al. Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutr*. 2002;5(6a):915-923. doi:10.1079/PHN2002383.
53. Freedman LS, Commins JM, Moler JE, et al. Pooled Results From 5 Validation Studies of Dietary Self-Report Instruments Using Recovery Biomarkers for Energy and Protein Intake. *Am J Epidemiol*. 2014;180(2):172-188. doi:10.1093/aje/kwu116.
54. Gibson R. *Principles of Nutritional Assessment*. Second Edi. New York: Oxford University Press; 2005.
55. Kutner MH, Nachtsheim CJ, Neter J, Li W. *Applied Linear Statistical Models*. 5th ed. McGraw-Hill; 2005. http://books.google.fr/books?id=0xqCAAAACAAJ&dq=intitle:Applied+linear+statistical+models+djvu&hl=&cd=1&source=gbs_api.
56. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA*. 2018;319(13):1317. doi:10.1001/jama.2017.18391.
57. Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Machine Learning Techniques for Prediction of Early Childhood Obesity. *Appl Clin Inform*. 2015;06(03):506-520. doi:10.4338/ACI-2015-03-RA-0036.
58. Michie D, Spiegelhalter D, Taylor C. *Machine Learning, Neural and Statistical*

Classification.; 1994. doi:10.2307/1269742.

59. Bishop CM. *Pattern Recognition and Machine Learning*. (Jordan M, Kleinberg J, Scholkopf B, eds.). Singapore: Springer; 2006. doi:10.1117/1.2819119.
60. James G, Witten D, Tibshirani R, Hastie T. An Introduction to Statistical Learning with Applications in R. *Book*. 2013:431. doi:10.1007/978-1-4614-7138-7.
61. Maimon O, Rokach L. *Data Mining and Knowledge Discovery Handbook*. 2nd ed. Springer; 2010. doi:10.1007/978-0-387-09823-4.
62. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. *ACM Comput Surv*. 1999;31(3):264-323. doi:10.1145/331499.331504.
63. Tobar F. Maquinas que aprenden. *Rev BITS Ciencias*. 2017;15:36-43.
64. Pohlmeyer EA, Mahmoudi B, Geng S, Prins NW, Sanchez JC. Using reinforcement learning to provide stable brain-machine interface control despite neural input reorganization. *PLoS One*. 2014;9(1). doi:10.1371/journal.pone.0087253.
65. Nau C, Ellis H, Huang H, et al. Exploring the forest instead of the trees: An innovative method for defining obesogenic and obesoprotective environments. *Heal Place*. 2015;35:136-146. doi:10.1016/j.healthplace.2015.08.002.
66. Schapire RE. *The Boosting Approach to Machine Learning An Overview*.; 2002. www.research.att.com/. Accessed March 7, 2019.
67. Seyednasrollah F, Mäkelä J, Pitkänen N, et al. Prediction of Adulthood Obesity Using Genetic and Childhood Clinical Risk Factors in the Cardiovascular Risk in Young Finns Study. *Circ Cardiovasc Genet*. 2017;10(3):e001554. doi:10.1161/CIRCGENETICS.116.001554.
68. Sze MA, Schloss PD. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. *MBio*. 2016;7(4):e01018-16. doi:10.1128/mBio.01018-16.
69. Thaiss CA, Itav S, Rothschild D, et al. Persistent microbiome alterations modulate the rate of post-dieting weight regain. *Nature*. 2016;540(7634):544-551. doi:10.1038/nature20796.
70. Lee BJ, Kim KH, Ku B, Jang J-S, Kim JY. Prediction of body mass index status from voice signals based on machine learning for automated medical applications. *Artif Intell Med*. 2013;58(1):51-61. doi:10.1016/J.ARTMED.2013.02.001.
71. Giabbanelli PJ, 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 Nutr*. 2016;19(9):1543-1551. doi:10.1017/S1368980016000185.

72. Lazarou C, Karaolis M, Matalas A-L, Panagiotakos DB. Dietary patterns analysis using data mining method. An application to data from the CYKIDS study. *Comput Methods Programs Biomed*. 2012;108(2):706-714. doi:10.1016/J.CMPB.2011.12.011.
73. Kastorini C-M, Papadakis G, Milionis HJ, et al. Comparative analysis of a-priori and a-posteriori dietary patterns using state-of-the-art classification algorithms: A case/case-control study. *Artif Intell Med*. 2013;59(3):175-183. doi:10.1016/J.ARTMED.2013.08.005.
74. Thangamani D, Sudha P. *Identification Of Malnutrition With Use Of Supervised Datamining Techniques-Decision Trees And Artificial Neural Networks*.; 2014. www.ijecs.in. Accessed March 27, 2019.
75. Einsele F, Sadeghi L, Jenzer H, Ingold R. A Study about Discovery of Critical Food Consumption Patterns Linked with Lifestyle Diseases using Data Mining Methods. *Proc Int Conf Heal Informatics*. 2015:239-245. doi:10.5220/0005170402390245.
76. Weber I, Achananuparp P. Insights from Machine-Learned Diet Success Prediction. In: *Biocomputing 2016*. WORLD SCIENTIFIC; 2016:540-551. doi:10.1142/9789814749411_0049.
77. Hossain M, Mullally C, Asadullah MN. Alternatives to calorie-based indicators of food security: An application of machine learning methods. *Food Policy*. March 2019. doi:10.1016/J.FOODPOL.2019.03.001.
78. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16(6):321-332. doi:10.1038/nrg3920.
79. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V., Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8-17. doi:10.1016/J.CSBJ.2014.11.005.
80. Ball N, Btunner R. Data Mining and Machine Learning in Astronomy. *Int J Mod Phys D*. 2010;19(07):1049-1106. doi:10.1142/S0218271810017160.
81. Wei-Yang Lin, Ya-Han Hu, Chih-Fong Tsai. Machine Learning in Financial Crisis Prediction: A Survey. *IEEE Trans Syst Man, Cybern Part C (Applications Rev*. 2012;42(4):421-436. doi:10.1109/TSMCC.2011.2170420.
82. Departamento de Nutrición, Escuela de Nutrición, Escuela de Salud Pública, Cetro de Microdatos. *Encuesta Nacional de Consumo Alimentos*.; 2014. http://web.minsal.cl/sites/default/files/ENCA-INFORME_FINAL.pdf.
83. Ministerio de Salud Chile. ENCUESTA DE CONSUMO ALIMENTARIO EN CHILE (ENCA) - Ministerio de Salud - Gobierno de Chile. http://www.minsal.cl/enca/.

Accessed July 10, 2018.

84. Departamento de Nutrición, Escuela de Nutrición, Escuela de Salud Pública, Cetro de Microdatos. *Manual Usuario Encuesta Nacional Consumo Alimentario 2010-2011*.; 2011.
85. Hastings C, Mosteller F, Tukey JW, Winsor CP. Low Moments for Small Samples: A Comparative Study of Order Statistics. *Ann Math Stat*. 1947;18(3):413-426. doi:10.1214/aoms/1177730388.
86. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297. doi:10.1007/BF00994018.
87. Introduction to Boosted Trees — xgboost 0.72 documentation. https://xgboost.readthedocs.io/en/latest/model.html#tree-boosting. Accessed July 10, 2018.
88. Tuning the hyper-parameters of an estimator — scikit-learn 0.19.1 documentation. http://scikit-learn.org/stable/modules/grid_search.html. Accessed July 10, 2018.
89. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J Mach Learn Res*. 2016;18:1-5. doi:http://www.jmlr.org/papers/volume18/16-365/16-365.pdf.
90. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*. 2009;21(9):1263-1284. doi:10.1109/TKDE.2008.239.
91. Tsoumakas G, Katakis I, Vlahavas I. Mining Multi-label Data. In: Maimon O, Rokach L, eds. *Data Mining and Knowledge Discovery Handbook*. 2nd ed. Springer; 2010:667-685. doi:10.1017/CBO9781107415324.004.
92. Street F. *Dietary Reference Intakes*.; 2000. doi:10.17226/9956.
93. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2012;12:2825-2830. doi:10.1007/s13398-014-0173-7.2.
94. Jones E, Oliphant T, Peterson P. SciPy: Open source scientific tools for Python. doi:10.1109/MCSE.2007.58.
95. Becker W, Welten D. Under-reporting in dietary surveys--implications for development of food-based dietary guidelines. *Public Health Nutr*. 2001;4(2B):683-687. doi:10.1079/PHN2001154.
96. Engel GL. The need for a new medical model: a challenge for biomedicine. *Science*. 1977;196(4286):129-136. doi:10.1126/SCIENCE.847460.
97. Borrell-Carrió F, Suchman AL, Epstein RM. The biopsychosocial model 25 years later: principles, practice, and scientific inquiry. *Ann Fam Med*. 2004;2(6):576-582.

doi:10.1370/afm.245.

98. Manios Y, Androutsos O, Katsarou C, et al. Prevalence and sociodemographic correlates of overweight and obesity in a large Pan-European cohort of preschool children and their families: The ToyBox-study. *Nutrition*. 2018;55-56:192-198. doi:10.1016/J.NUT.2018.05.007.
99. Marqueta de Salas M, Martín-Ramiro JJ, Juárez Soto JJ. Sociodemographic characteristics as risk factors for obesity and overweight in Spanish adult population. *Med Clínica (English Ed*. 2016;146(11):471-477. doi:10.1016/J.MEDCLE.2016.07.001.
100. Saïle R, Msaad R, Mohtadi K, et al. Prevalence of Obesity and Associated Sociodemographic Factors in Casablanca, Morocco. *Atheroscler Suppl*. 2018;32:76-77. doi:10.1016/J.ATHEROSCLEROSISSUP.2018.04.233.
101. Mifflin MD, St Jeor ST, Hill L a, Scott BJ, Daugherty S a, Koh YO. A new predictive equation for resting energy expenditure in healthy individuals. *Am J Clin Nutr*. 1990;51:241-247.
102. Jelinek GA. Determining Causation from Observational Studies: A Challenge for Modern Neuroepidemiology. *Front Neurol*. 2017;8:265. doi:10.3389/fneur.2017.00265.
103. Glymour MM, Spiegelman D. Evaluating Public Health Interventions: 5. Causal Inference in Public Health Research-Do Sex, Race, and Biological Factors Cause Health Outcomes? *Am J Public Health*. 2017;107(1):81-85. doi:10.2105/AJPH.2016.303539.
104. Jackson A, Stanforth P, Gagnon J, et al. The effect of sex, age and race on estimating percentage body fat from body mass index: The Heritage Family Study. *Int J Obes*. 2002;26:789-796. doi:10.1038=sj.ijo.0802006.
105. Carlos S, De La Fuente-Arrillaga C, Bes-Rastrollo M, et al. Mediterranean diet and health outcomes in the SUN cohort. *Nutrients*. 2018;10(4):1-24. doi:10.3390/nu10040439.
106. Amigo H, Pizarro M, Bustos P, et al. *Anexos ENCA*.; 2010.

ANEXOS

Anexo 1. Flujos de entrevistas de Encuesta Nacional de Consumo de Alimentos

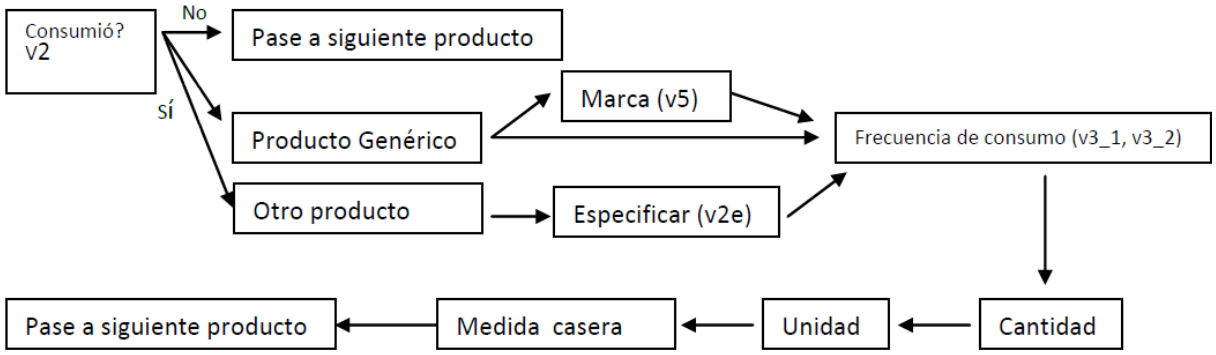


Figura 13. Flujo de entrevista Encuesta de Tendencia de Consumo Cuantificado, Fuente: Manual de Usuario ENCA 2010-2011[84],

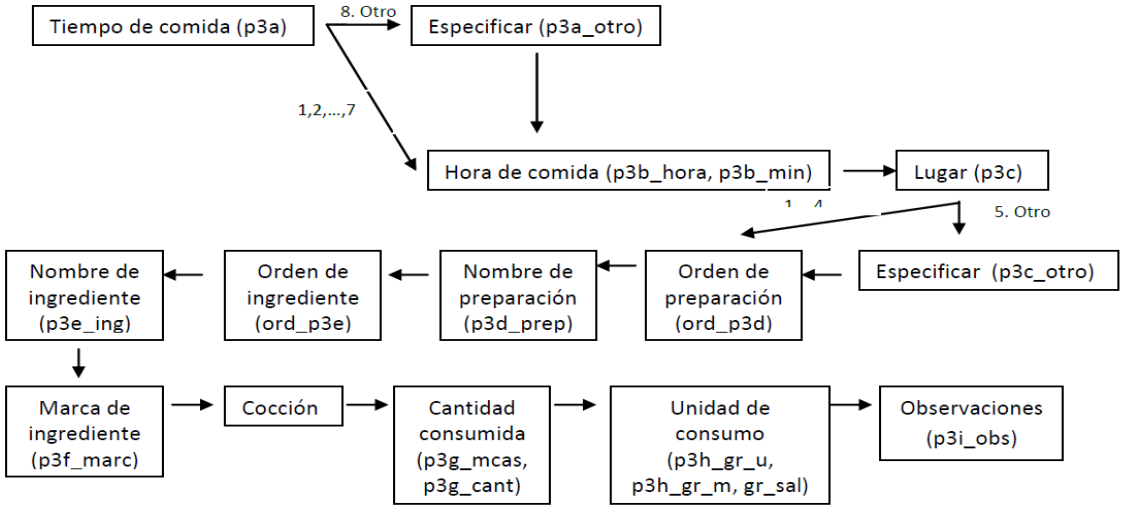


Figura 14. Flujo de entrevista Encuesta de Recordatorio de 24 horas, Fuente: Manual de Usuario ENCA 2010-2011[84],

Anexo 2. Lista de alimentos y agrupación consultados en ENCA

Lista de alimentos consultados en Encuesta de Tendencia de Consumo Cuantificado (ETCC) y su agrupación utilizados en esta tesis. Obtenido de documentación ENCA Anexos[106]. Disponible en: http://web.minsal.cl/enca/ [83].

Tabla 9. Variables consumo mensual agrupado y no agrupado (g/ml).

Grupo de Alimentos	Alimentos			
Cereales y Pastas	chuchoca	mote maíz	arroz	arroz integral
	chuño	mote trigo	fideos	fideos integrales
	harina tostada	sémola	fideos ravioles	fideos torteloni
	maicena	quínoa	masa de lasaña	fideos agnolotti
Pan	pan amasado	pan hallulla sin sal	pan molde integral	galleta de agua light
	pan baguette	pan hotdog	pan paris	galleta de agua sin sal
	pan croissant	pan marraqueta	pan pita blanco	galleta de soda
	pan dobladita	pan marraqueta integral	pan pita integral	galleta de soda light
	pan frica	pan marraqueta sin sal	pan tortilla de rescoldo	galleta integral
	pan hallulla	pan molde blanco	tortilla de maíz	galleta integral light
	pan hallulla integral	pan molde blanco light	galleta de agua	
Cereales de Desayuno	cereal avena con azúcar	cereal granola con azúcar	cereal integral	cereal nestum avena
	cereal avena con chocolate	cereal hojuelas maíz azucarado	cereal nestum 3 cereales	cereal nestum frutilla
	cereal cerelac 5 cereales	cereal inflado	cereal nestum arroz	cereal nestum miel
Cereales procesados	galleta dulce con sabor			
	galleta dulce rellena con crema			
Leguminosas frescas,	arvejas naturales	habas	papa	piñón

Tabla 9. Variables consumo mensual agrupado y no agrupado (g/ml).

Grupo de Alimentos	Alimentos			
papas y otros	castaña de cajú	ñoquis	castaña	yuca
	choclo natural			
Verduras en general	acelga cocida	repollito de Bruselas	Apio	pepino
	alcachofa	zapallo camote	Berro	perejil
	berenjena	zapallo italiano	ciboulette	pimentón
	betarraga cocida	cebolla	cilantro	repollo
	brócoli	cebollín	cochayuyo	rúcula
	champiñón	tomate	diente de dragón	ulte
	coliflor	zanahoria	lechuga	brote de alfalfa
	esparrago	achicoria	Luche	endivia
	espinaca cocida	Ají	palmito en conserva	nabo
	porotos verdes	Ajo	penca	
Frutas en general	arándano	guindas en conserva	Papaya	guayaba
	chirimoya	Higo	papaya en conserva	limón
	ciruela	Kiwi	pepino dulce	lúcuma
	clementina	macedonia en conserva	Pera	maqui
	damasco	Mandarina	Piña	maracuyá
	damasco en conserva	Mango	piña en conserva	murtilla
	durazno	mango en conserva	Plátano	tuna
	durazno en conserva	Manzana	Pomelo	jugo natural limón
	frambuesa	Melón	Sandia	ciruela deshidratada

Tabla 9. Variables consumo mensual agrupado y no agrupado (g/ml).

Grupo de Alimentos	Alimentos			
	frutilla	Membrillo	Uva	huesillo
	frutilla en conserva	Mora	Calafate	pasas
	Grosella	Naranja	Caqui	damasco deshidratado
	Guinda	Níspero	granada	manzana deshidratada
Lácteos altos en grasas	leche cultivada	leche líquida entera sin sabor con lactosa	leche purita fortificada	leche de cabra
	leche en polvo entera sin sabor con lactosa	leche líquida entera sin sabor sin lactosa	leche purita mama	leche de oveja
	leche líquida entera con sabor normal con lactosa	leche purita cereal	yogurt entero sin sabor	
Quesos	queso mantecoso	queso crema	queso de oveja	queso gouda
	queso chanco	queso de cabra	queso Edam	queso parmesano
				queso azul
Lácteos medios en grasas	bebida láctea años dorados	leche líquida semidescremada sin sabor con lactosa	quesillo sin sal	leche en polvo semidescremada con omega 3 con lactosa
	leche en polvo semidescremada sin sabor con lactosa	leche líquida semidescremada sin sabor sin lactosa	queso fresco	leche líquida semidescremada con calcio con lactosa
	leche líquida semidescremada con sabor normal con lactosa	quesillo	leche en polvo de soya	leche líquida semidescremada con sabor normal sin lactosa
Lácteos bajos en grasa	bebida láctea probiótico light	leche en polvo descremada sin sabor con lactosa	leche líquida descremada sin sabor con lactosa	yogurt descremado con sabor
	leche cultivada light	leche en polvo descremada sin sabor sin lactosa	leche líquida descremada sin sabor sin lactosa	leche líquida descremada con calcio sin lactosa

Tabla 9. Variables consumo mensual agrupado y no agrupado (g/ml).

Grupo de Alimentos	Alimentos			
	leche en polvo descremada con calcio con lactosa	leche líquida descremada con calcio con lactosa	quesillo light	leche líquida descremada con omega 3 con lactosa
	leche en polvo descremada con fibra con lactosa	leche líquida descremada con fibra con lactosa	yogurt descremado con fruta	
	leche en polvo descremada con omega 3 con lactosa	leche líquida descremada con sabor diet con lactosa	yogurt descremado con frutos secos	
Lácteos medios en grasa, ricos en HdeC	bebida láctea probiótico normal	postre envasado arroz con leche	yogurt entero con fruta	postre envasado mousse
	leche condensada	postre envasado flan sabor vainilla	yogurt entero con frutos secos	leche condensada light
	postre envasado 1+1 light	postre envasado leche asada	yogurt entero con sabor	leche de burra
	postre envasado 1+1			
Carnes Rojas	carne de alpaca	carne de chivo	carne de pollo pana	carne de vacuno molida corriente
	carne de cabrito costillar	carne de chivo costillar	carne de vacuno asado carnicero	carne de vacuno molida tártaro
	carne de cabrito pulpa	carne de conejo	carne de vacuno asado de tira	carne de vacuno osobuco
	carne de cerdo cazuela	carne de cordero chuleta	carne de vacuno asiento	carne de vacuno pana
	carne de cerdo chuleta	carne de cordero costillar	carne de vacuno cazuela	carne de vacuno plateada
	carne de cerdo costillar	carne de cordero guatitas	carne de vacuno charqui	carne de vacuno pollo ganso
	carne de cerdo criadilla	carne de cordero osobuco	carne de vacuno corazón	carne de vacuno posta negra
	carne de cerdo lomo	carne de cordero pernil	carne de vacuno criadilla	carne de vacuno posta rosada
	carne de cerdo osobuco	carne de cordero pulpa	carne de vacuno filete	carne de vacuno prieta

Tabla 9. Variables consumo mensual agrupado y no agrupado (g/ml).

Grupo de Alimentos	Alimentos			
	carne de cerdo pernil	carne de equino posta	carne de vacuno guatita	carne de vacuno riñón
	carne de cerdo pulpa	carne de llama	carne de vacuno guachalomo	carne de vacuno seso
	carne de cerdo riñón	carne de oveja	carne de vacuno lengua	carne de vacuno sobrecostillas
	carne de cerdo seso	carne de pollo contre	carne de vacuno lomo liso	carne de vacuno ubres
	carne de cerdo ubres	carne de pollo corazón	carne de vacuno lomo vetado	chunchules
Aves	carne de pavo ala	carne de pavo pechuga	carne de pollo cogote	carne de pollo pechuga
	carne de pavo molida	carne de pavo trutro	carne de pollo espinazo	carne de pollo rabadilla
	carne de pavo osobuco	carne de pollo ala	carne de pollo pata	carne de pollo trutro
Pescados y mariscos	marisco cholga en aceite en conserva	marisco ostra fresca	pescado congrio	pescado reineta
	jaiba	marisco picoroco	pescado congrio dorado	pescado robalo
	marisco almeja en conserva	navajuela	pescado corolado	pescado salmón
	marisco almeja fresca	pescado albacora	pescado corvina	pescado sardina en conserva
	marisco calamar	pescado albacorilla	pescado jibia	pescado sardina natural
	marisco camarón congelado	pescado atún en aceite en conserva	pescado jurel en conserva	pescado sargo
	marisco camarón fresco	pescado atún en agua en conserva	pescado jurel natural	pescado sierra
	marisco cholga natural	pescado baunco	pescado lenguado	pescado tollo
	marisco chorito en aceite en conserva	pescado blanquillo	pescado lisa	pescado tomollo
	marisco chorito natural	pescado caballa	pescado merluza	pescado trucha arcoiris
	marisco erizo	pescado cabinza	pescado mero	pescado vidriola

Tabla 9. Variables consumo mensual agrupado y no agrupado (g/ml).

Grupo de Alimentos	Alimentos			
	marisco macha en conserva	pescado carpa	pescado pejegallo	piure
	marisco macha natural	pescado centolla	pescado pejerrey	
	marisco ostión	pescado cojinova	pescado pejesapo	
Carnes procesadas	carne vegetal	cecina jamonada	gorda de vacuno	hamburguesa de vacuno light
	cecina arrollado huaso	cecina mortadela	gorda de vacuno	queso cabeza
	cecina jamón ahumado	cecina salame	hamburguesa de cerdo	vienesa de cerdo
	cecina jamón cocido	cecina salchichón cerveza	hamburguesa de pavo	vienesa de pavo
	cecina jamón de pavo	gorda de cerdo	hamburguesa de pollo	vienesa de pollo
	cecina jamón de pollo	gorda de pollo	hamburguesa de vacuno	vienesa de vacuno
Huevos	huevo clara	huevo entero	huevo yema	huevo de codorniz
Leguminosa Seca	garbanzos	lentejas	porotos granados	porotos negros
	porotos tórtolas			
Aceites y grasas poliinsaturados	aceite de soya	aceite maíz	aceite maravilla	aceite mezcla
	aceite pepita de uva			
Aceites y grasas monoinsaturados	aceite canola	aceite de palta	aceite oliva	
Aceites y grasas saturadas	cecina tocino	crema batida	longaniza	margarina
	chicharrones	crema chantilly	manteca animal	mayonesa
	choricillo	crema espesa	mantequilla	pate de ternera
	chorizo	crema espesa light	mantequilla sin sal	crema ácida

Tabla 9. Variables consumo mensual agrupado y no agrupado (g/ml).

Grupo de Alimentos	Alimentos			
Alimentos ricos en lípidos, monoinsaturados	aceituna	avellana	nuez	pistacho
	almendra	maní	palta	snack salado maní con sal
Azúcares	azúcar	azúcar light		
Otros azúcares	caramelo duro	jugo en polvo normal	manjar light	miel de abeja
	jalea	leche condensada	mermelada	miel de palma
	jarabe para postre sabor frambuesa	manjar	mermelada diet	leche condensada light
Bebidas y refrescos azucarados	bebida de fantasía coca cola normal	bebida isotónica	jugo liquido normal	helado de yogurt
	bebida energética normal	jugo liquido light	helado de agua	helado de crema
Golosinas y otros alimentos dulces	cacao en polvo	chocolate solido dulce	pastel pie	snack dulce queque de vainilla
	chocolate con frutos secos	pastel alfajor bañado en chocolate	pastel tartaleta de fruta	torta de bizcocho con manjar y crema
	chocolate en polvo dulce	pastel Berlín relleno con crema pastelera	snack dulce cuchuflí con baño de chocolate	torta de mil hojas manjar
	chocolate relleno	pastel kuchen de fruta	snack dulce cuchuflí relleno con manjar	torta de panqueque
Bebidas y refrescos libres de calorias	bebida de fantasía coca cola light	bebida de fantasía coca cola zero	jugo en polvo light	
Endulzantes no nutritivos	endulzante liquido sacarina	endulzante liquido sucralosa		
Alcohol	licor cerveza	licor vino blanco	licor aguardiente	licor vodka
	licor chicha	licor vino rose	licor pisco	licor whisky
	licor cola de mono	licor vino tinto	licor ron	
	licor champaña	licor sidra	licor tequila	

Tabla 9. Variables consumo mensual agrupado y no agrupado (g/ml).

Grupo de Alimentos	Alimentos			
Crema Años Dorados	crema años dorados			
Bebida Láctea Años Dorados	bebida láctea años dorados			
Leche Purita	leche purita fortificada	leche purita cereal	leche purita mama	alimento mi sopita

Variables de nutrientes consumidos utilizados en esta tesis, obtenidos de la base de datos ENCA_Er24h_Nutrientes, construida a partir de la Encuesta Recordatoria de 24hr, cómo se describe en el Manual de Usuario ENCA[84]. Disponible en: http://web.minsal.cl/enca/ [83]

Tabla 10. Variables Nutrientes

Calorías (Kcals)
Proteínas (g/día)
Hidratos de Carbono (g/día)
Lípidos (g/día)
Grasas Saturadas (mg/día)
Grasas Monoinsaturadas (mg/día)
Grasas poliinsaturadas (mg/día)
Colesterol (mg/día)

Lista de variables sociodemográficas utilizados en esta tesis, obtenidos de la base de datos ENCA_NSE, descritos en Manual de Usuario ENCA[84]. Disponibles en: http://web.minsal.cl/enca/ [83].

Tabla 11. Variables Sociodemográficas

Edad (años)
Género (hombre/mujer)
Área (urbana/rural)
Macrozona (Norte/Centro-Norte/Centro/Centro-Sur/Sur)
Nivel Socioeconómico (Bajo/Medio-Bajo/ Medio/Medio-Alto/Alto)

Anexo 3. Variables alto consumo de calorías seleccionadas

Tabla 12. Variables de consumo mensual agrupado seleccionadas para exploración de grupos de alimentos con alto consumo de calorías.

Cereales y Pastas
Pan
Cereales de desayuno
Cereales procesados
Lácteos altos en grasas
Lácteos medios en grasa
Lácteos medios en grasa ricos en Hidratos de Carbono,
Aceites y grasas poliinsaturados
Azúcar
Otros Azúcares
Bebidas y refrescos azucarados
Golosinas y otros alimentos dulces
Endulzantes no nutritivos
Alcohol

Anexo 4. Desempeño de clasificación de estado nutricional (dos clases)

Tabla 13. Resultados de calificación de estado nutricional (dos clases).

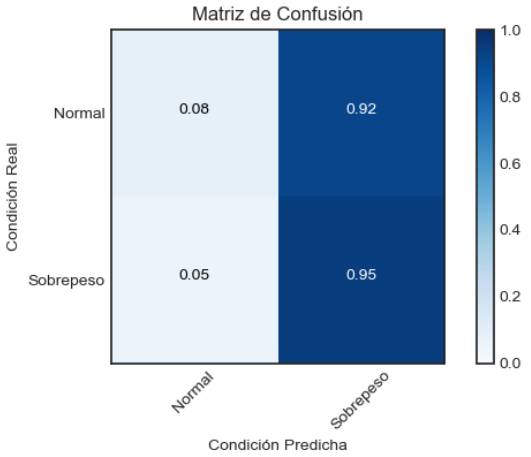
Variables Predictoras	Métricas[1]	RegLog Media ± DE	MSV Media ± DE	BA Media ± DE	RN Media ± DE	XGB Media ± DE	p-val[2]
Alim, Consumo Mensual sSOC[3]	Exactitud	61,3 ± 5,2	69,7 ± 1,5	68,6 ± 3,5	67,1 ± 3,3	67,5 ± 3,5	2:3:5:4-1
	Precisión	81,8 ± 4,3	74,1 ± 0,9	76,1 ± 2,2	76,7 ± 2,4	76,7 ± 2,4	
	Sensibilidad	61,5 ± 8,6	90,8 ± 2,2	83,9 ± 4,6	79,7 ± 4,1	80,5 ± 5,9	
	Especificidad	60,5 ± 13,6	10,4 ± 4,7	25,6 ± 10,0	31,7 ± 9,8	30,9 ± 11,1	
	F1	69,8 ± 5,4	81,6 ± 1,0	79,7 ± 2,5	78,1 ± 2,4	78,4 ± 2,9	
Alim, Consumo Mensual cSOC[3]	Exactitud	64,8 ± 6,0	72,2 ± 2,6	70,3 ± 2,4	68,6 ± 4,2	71,3 ± 2,7	2:5:3:4-1
	Precisión	83,1 ± 4,6	74,5 ± 1,2	77,4 ± 1,8	77,3 ± 2,3	78,2 ± 1,9	
	Sensibilidad	61,3 ± 9,2	94,5 ± 3,3	84,5 ± 4,5	81,5 ± 5,4	85,0 ± 5,7	
	Especificidad	61,3 ± 14,4	8,3 ± 5,2	30,3 ± 9,5	32,4 ± 8,3	32,8 ± 10,3	
	F1	73,1 ± 6,2	83,4 ± 1,7	80,7 ± 1,9	79,2 ± 3,2	81,3 ± 2,4	
Alim, Consumo Mensual Agrupado sSOC	Exactitud	58,5 ± 4,2	67,7 ± 2,9	65,0 ± 2,5	60,2 ± 2,7	62,4 ± 2,4	2:3-5:4:1
	Precisión	78,6 ± 3,7	72,6 ± 1,3	75,9 ± 1,5	73,8 ± 1,9	75,8 ± 1,4	
	Sensibilidad	59,2 ± 5,6	89,3 ± 3,5	76,0 ± 3,6	70,3 ± 2,7	71,0 ± 4,4	
	Especificidad	56,7 ± 9,9	10,4 ± 4,0	35,8 ± 5,4	33,5 ± 5,4	39,8 ± 5,9	
	F1	67,4 ± 4,1	80,1 ± 2,1	75,9 ± 2,1	72,0 ± 2,1	73,2 ± 2,4	
Alim, Consumo Mensual Agrupado cSOC	Exactitud	62,4 ± 5,9	70,6 ± 1,0	67,6 ± 3,6	62,2 ± 3,1	68,0 ± 3,3	2:5:3-1:4
	Precisión	80,6 ± 2,9	72,9 ± 0,5	77,1 ± 2,2	75,3 ± 1,8	76,4 ± 1,5	
	Sensibilidad	63,4 ± 9,3	94,8 ± 2,1	79,5 ± 6,7	71,5 ± 4,4	81,3 ± 7,8	
	Especificidad	59,6 ± 8,0	6,3 ± 3,3	37,0 ± 10,3	37,5 ± 6,2	32,7 ± 10,6	
	F1	70,6 ± 6,8	82,4 ± 0,8	78,1 ± 3,3	73,3 ± 2,7	78,5 ± 3,3	
Nutrientes sSOC	Exactitud	52,1 ± 3,4	58,0 ± 2,6	57,7 ± 3,2	50,8 ± 5,3	51,4 ± 2,8	2:3-1:5:4
	Precisión	72,3 ± 2,4	71,9 ± 1,9	72,0 ± 1,7	71,0 ± 1,8	71,8 ± 2,0	
	Sensibilidad	55,4 ± 5,6	69,3 ± 2,8	68,4 ± 4,6	54,4 ± 10,2	54,5 ± 4,2	
	Especificidad	43,6 ± 7,3	27,9 ± 6,2	29,5 ± 5,4	41,2 ± 9,6	43,1 ± 5,8	
	F1	62,6 ± 3,9	70,6 ± 2,0	70,1 ± 2,8	61,1 ± 7,3	61,9 ± 3,0	
Nutrientes cSOC	Exactitud	60,1 ± 5,0	68,8 ± 2,3	66,8 ± 3,6	64,1 ± 3,7	67,8 ± 3,2	2:5:3-4:1
	Precisión	79,3 ± 2,2	73,0 ± 1,3	76,3 ± 2,4	75,5 ± 2,1	76,4 ± 1,8	
	Sensibilidad	61,0 ± 8,8	90,7 ± 4,7	79,0 ± 7,2	75,1 ± 7,2	80,8 ± 7,0	
	Especificidad	57,5 ± 7,8	10,4 ± 7,9	34,2 ± 11,1	34,9 ± 10,7	33,2 ± 10,0	
	F1	68,6 ± 6,1	80,8 ± 1,8	77,4 ± 3,2	75,1 ± 3,7	78,3 ± 3,0	
Sociodemográfico	Exactitud	60,7 ± 5,6	62,5 ± 4,4	65,0 ± 4,1	63,5 ± 5,6	66,7 ± 3,3	5:3:4:2:1
	Precisión	79,8 ± 2,2	79,6 ± 2,2	77,0 ± 1,9	79,1 ± 1,8	77,6 ± 2,1	
	Sensibilidad	61,5 ± 10,0	65,5 ± 9,8	74,3 ± 10,3	68,1 ± 12,4	76,5 ± 8,5	
	Especificidad	58,6 ± 8,6	45,5 ± 11,8	40,1 ± 13,7	51,4 ± 13,2	40,6 ± 12,5	
	F1	69,0 ± 6,8	71,4 ± 5,4	75,2 ± 4,5	72,5 ± 6,5	76,7 ± 3,7	

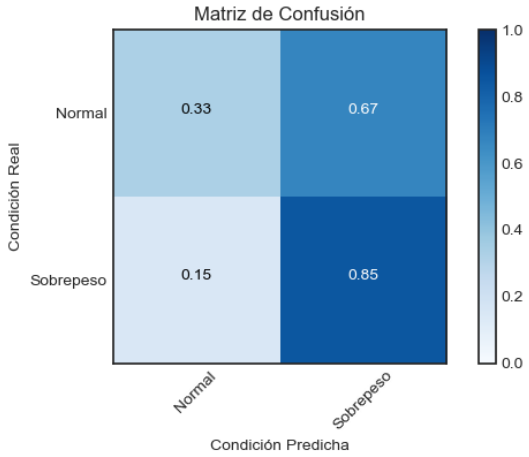
[1]Unidad de medida de métricas: porcentaje (%); [2] Se muestran resultados de test ANOVA, corrección Bonferroni (significancia *p-value* < 0,05), de mejor a peor desempeño de exactitud, 1: RegLog (Regresión Logística); 2: MSV (Máquinas Soporte Vectorial); 3: BA (Bosques Aleatorios); 4: RN (Redes Neuronales); 5: XGB (*Extreme Gradient Boosting*), ": " significa que no hay diferencias significativas en esta comparación, " - " significa que hay diferencias significativas; [3]sSOC: sin variables sociodemográficas; cSOC: con variables sociodemográficas; [4] Mejor resultado para cada métrica y *set* de datos

Anexo 5. Matrices de Confusión

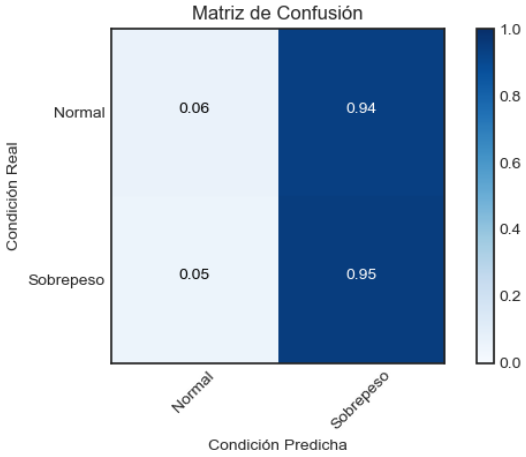
Matrices de confusión resultantes de la clasificación para 2 clases implementadas con el algoritmo SVM (izquierda) y XGB (derecha), para variables predictoras de dieta (consumo mensual no agrupado y agrupado) con variables sociodemográficas. Matrices de confusión normalizadas considerando la condición real como referencia.

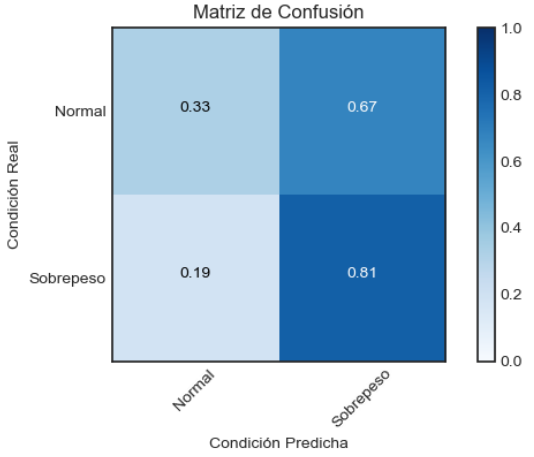
Consumo Mensual No Agrupado



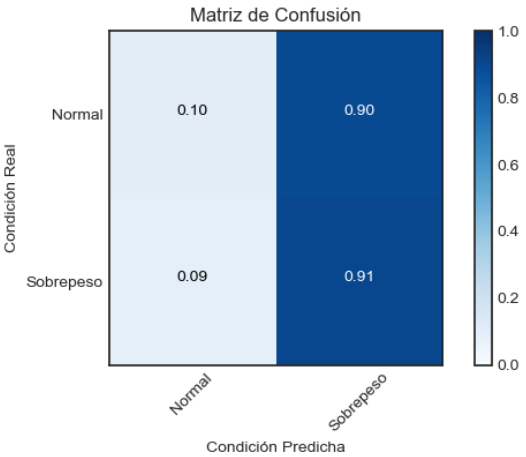


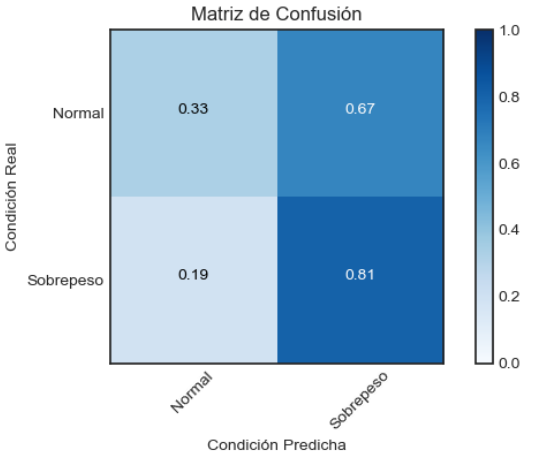
Consumo Mensual Agrupado





Nutrientes





Anexo 6. Desempeño de clasificación de estado nutricional (tres clases)

Tabla 14. Resultados de calificación de estado nutricional (tres clases).

Variables Predictoras	Métricas[1]	RegLog Media ± DE	MSV Media ± DE	BA Media ± DE	RN Media ± DE	XGB Media ± DE	p-val[2]
Alim, Consumo Mensual sSOC[3]	Exactitud	43,6 ± 2,1	43,5 ± 2,8	43,8 ± 2,8	41,4 ± 3,4	42,8 ± 4,0	3:1:2:5:4
	Precisión	43,5 ± 3,0	44,5 ± 5,2	43,5 ± 4,5	40,5 ± 3,8	41,8 ± 4,8	
	Sensibilidad	41,7 ± 2,1	40,3 ± 2,8	40,3 ± 2,6	40,2 ± 3,6	40,6 ± 4,1	
	F1	41,3 ± 2,5	38,6 ± 3,2	37,8 ± 2,9	40,0 ± 3,8	39,8 ± 4,5	
Alim, Consumo Mensual cSOC[3]	Exactitud	45,4 ± 2,5	45,2 ± 2,8	46,7 ± 3,0	40,6 ± 4,0	44,7 ± 2,9	3:1:2:5-4
	Precisión	46,6 ± 2,1	46,1 ± 3,5	47,3 ± 4,5	40,1 ± 4,6	44,4 ± 3,5	
	Sensibilidad	44,6 ± 2,6	43,9 ± 3,2	44,3 ± 3,0	39,5 ± 4,2	43,6 ± 2,6	
	F1	43,6 ± 4,6	43,5 ± 3,5	43,6 ± 3,7	39,2 ± 4,6	42,8 ± 4,5	
Alim, Consumo Mensual Agrupado sSOC	Exactitud	41,6 ± 2,6	41,2 ± 2,2	40,4 ± 3,4	38,7 ± 3,3	41,2 ± 4,1	1:2:5:3:4
	Precisión	42,6 ± 4,2	42,1 ± 5,3	40,7 ± 4,2	38,4 ± 3,7	41,4 ± 4,3	
	Sensibilidad	38,9 ± 2,7	37,6 ± 2,3	38,1 ± 3,3	37,0 ± 3,2	39,5 ± 4,1	
	F1	37,8 ± 3,2	34,5 ± 3,2	37,5 ± 3,5	36,5 ± 3,2	39,3 ± 4,3	
Alim, Consumo Mensual Agrupado cSOC	Exactitud	45,3 ± 4,5	43,9 ± 3,4	44,2 ± 4,7	40,7 ± 2,1	44,0 ± 3,7	1:3:5:2:4
	Precisión	47,9 ± 6,5	45,0 ± 4,0	45,2 ± 4,8	39,9 ± 2,4	44,9 ± 3,4	
	Sensibilidad	44,4 ± 4,4	42,4 ± 3,3	43,0 ± 4,7	38,6 ± 1,9	43,2 ± 3,7	
	F1	43,0 ± 7,2	41,6 ± 4,1	43,2 ± 4,8	37,9 ± 2,1	42,4 ± 5,0	
Nutrientes sSOC	Exactitud	39,3 ± 1,8	39,8 ± 2,7	37,9 ± 1,8	39,3 ± 2,7	37,2 ± 2,2	2:1:4:3:5
	Precisión	32,6 ± 14,0	36,2 ± 6,8	33,9 ± 3,7	34,0 ± 10,0	34,2 ± 3,4	
	Sensibilidad	34,9 ± 1,7	35,6 ± 2,6	34,7 ± 1,9	35,2 ± 2,4	34,1 ± 2,2	
	F1	28,2 ± 2,1	29,8 ± 2,7	31,6 ± 2,1	29,7 ± 2,0	31,7 ± 2,5	
Nutrientes cSOC	Exactitud	43,2 ± 3,9	44,7 ± 2,5	41,7 ± 2,0	39,6 ± 2,5	40,2 ± 3,3	2:1:3-5:4
	Precisión	43,3 ± 4,3	46,1 ± 3,3	43,0 ± 2,4	35,0 ± 10,9	40,6 ± 3,6	
	Sensibilidad	42,3 ± 3,7	43,7 ± 2,6	40,2 ± 1,8	35,6 ± 2,6	39,1 ± 2,9	
	F1	40,6 ± 7,2	42,9 ± 3,5	39,5 ± 2,4	30,1 ± 2,5	38,1 ± 4,1	
Sociodemográfico	Exactitud	43,5 ± 3,0	44,0 ± 2,6	42,5 ± 3,5	43,1 ± 3,5	41,5 ± 3,2	2:1:4:3:5
	Precisión	44,4 ± 3,1	45,0 ± 2,8	43,6 ± 3,8	43,2 ± 4,8	44,3 ± 6,6	
	Sensibilidad	42,8 ± 2,8	44,4 ± 2,6	41,5 ± 3,7	42,2 ± 3,7	40,4 ± 3,2	
	F1	41,2 ± 6,2	43,2 ± 2,8	40,6 ± 4,1	40,2 ± 6,5	38,2 ± 5,9	

[1]Unidad de medida de métricas: porcentaje (%); [2] Se muestran resultados de test ANOVA, corrección Bonferroni (significancia *p-value* < 0,05), de mejor a peor desempeño de exactitud, 1: RegLog (Regresión Logística); 2: MSV (Máquinas Soporte Vectorial); 3: BA (Bosques Aleatorios); 4: RN (Redes Neuronales); 5: XGB (*Extreme Gradient Boosting*), ": " significa que no hay diferencias significativas en esta comparación, " - " significa que hay diferencias significativas; [3]sSOC: sin variables sociodemográficas; cSOC: con variables sociodemográficas; [4] Mejor resultado para cada métrica y *set* de datos.

Anexo 7. Desempeño de predicción IMC

Tabla 15. Resultados de regresión.

Variables Predictoras	RegLin Media ± DE	MSV Media ± DE	BA Media ± DE	RN Media ± DE	XGB Media ± DE	p-val[1]
Alim, Consumo Mensual sSOC[3] (RMSE)	5,41 ± 0,42	5,59 ± 0,45	5,38 ± 0,46	6,24 ± 0,37	5,42 ± 0,45	3:1:5:2-4
Alim, Consumo Mensual cSOC[3] (RMSE)	5,28 ± 0,38	5,47 ± 0,43	5,29 ± 0,40	6,14 ± 0,30	5,31 ± 0,43	1:3:5:2-4
Alim, Consumo Mensual Agrupado sSOC (RMSE)	5,34 ± 0,40	5,59 ± 0,48	5,32 ± 0,43	5,53 ± 0,42	5,37 ± 0,41	3:1:5:4:2
Alim, Consumo Mensual Agrupado cSOC (RMSE)	5,21 ± 0,36	5,49 ± 0,42	5,21 ± 0,36	5,52 ± 0,27	5,19 ± 0,37	5:3:1:2:4
Nutrientes sSOC (RMSE)	5,42 ± 0,42	5,83 ± 0,44	5,51 ± 0,38	5,46 ± 0,43	5,54 ± 0,39	1:4:3:5:2
Nutrientes cSOC (RMSE)	5,25 ± 0,37	5,53 ± 0,39	5,31 ± 0,34	5,36 ± 0,35	5,34 ± 0,37	1:3:5:4:2
Sociodemográficos (RMSE)	5,24 ± 0,37	5,58 ± 0,50	5,31 ± 0,35	5,30 ± 0,35	5,28 ± 0,39	1:5:4:3:2

RMSE (*Root Mean Square Error)*: raíz cuadrada del error cuadrático medio

[1] Se muestran resultados de test ANOVA, corrección Bonferroni (significancia *p-value* < 0,05) de mejor a peor desempeño, 1: RegLin (Regresión Lineal); 2: MSV (Máquinas Soporte Vectorial); 3: BA (Bosques Aleatorios); 4: RN (Redes Neuronales); 5: XGB (*Extreme Gradient Boosting*), ":" significa que no hay diferencias significativas en esta comparación, "-" significa que hay diferencias significativas.

[2] Mejor resultado para cada métrica y *set* de datos.

[3]sSOC: sin variables sociodemográficas; [4]cSOC: con variables sociodemográficas.