**Análisis y relación de importancia de hábitos alimenticios según Dataset de estudio “Estimation of obesity levels based on eating habits and physical condition”**.

Daniel Cabrera Diaz, Edwin Huaman Curo y Carlos Rebaza Valdivia

Secciones:

I Introducción

1. Objetivo del Proyecto

II Estado del Arte

III Diseño del Experimento

1. Descripción de conjunto de datos
2. Estadística Descriptiva y Visualización de Datos
3. Metodología

# **INTRODUCCIÓN**

Obesidad es hoy en día uno de los problemas de salud predominantes tanto en países desarrollados como países en vía de desarrollo,

Para poder mantener un ritmo de vida saludable es necesario que la cantidad de calorías tomadas sean equivalentes a la cantidad de calorías gastadas, si la cantidad de calorías tomadas de los alimentos son mayor a la cantidad de calorías gastadas esta diferencia se convierte en grasa y la acumulación de esta grasa es llamada obesidad.

**Hechos sobre obesidad y sobrepeso según la OMS:**

* En 2016 más de 1.9 mil millones de adultos mayores de 18 años tenían sobrepeso. De estos más de 650 millones tenían obesidad.
* En 2016, 39% de adultos mayores a 18 (39% de hombres y 40% de mujeres) tenían sobrepeso.
* Entre el 13% de la población total de adultos (11% hombres y 15% mujeres) tenían obesidad en el 2016.
* La prevalencia mundial de obesidad se triplico entre los años 1975 y 2016
* Cuando en 1975 menos del 1% de los niños y adolescentes entre 5-19 años eran obesos en 1975, más de 124 millones padecen de obesidad (6% mujeres y 8% hombres)
* Sobrepeso y obesidad están más relacionadas mundialmente a muertes que desnutrición o problemas de bajo peso. En la población total existen más personas con sobrepeso que personas con bajo peso, esto ocurre en todas las regiones excepto África y Asia.

1. Objetivo del Proyecto

Con este proyecto se busca extraer las características más importantes del dataset ***“Estimation of obesity levels based on eating habits and physical condition Data Set”,*** dichas importancias nos permitirán analizar qué hábitos tienen una fuerte relación con la obesidad de las personas, establecer valores mínimos y máximos, para mantener un peso óptimo, además de predecir si con los valores de algunos hábitos se habrá una tendencia o relación con obesidad. También se buscará la relación entre otras enfermedades tales como el cáncer u otras.

# ESTADO DEL ARTE

Los autores en Ref. [3] resaltan el esfuerzo de los investigadores para identificar los factores que provocan la obesidad de forma temprana y crean un método inteligente basado en técnicas supervisadas y no supervisadas de minería de datos, para la predicción de esta enfermedad con el fin de ayudar a las personas y profesionales a tener un estilo de vida más saludable. Para su investigación, utilizaron un dataset de estudiantes entre 18 y 25 años recolectados de instituciones de Colombia, México y Perú. Crean modelos basados en Árboles de decisión (DT) y Máquina de Soporte de Vectores (SVM) para predecir los niveles de obesidad y un modelo de agrupamiento K-Means para validar los modelos predictivos. Para los modelos utilizaron las siguientes métricas de evaluación; el mejor resultado obtenido es el DT con los siguientes resultados Accuracy (98.5%), Recall (98.5), True Positives (98.5), False Positives (0.2%) y área ROC (99.5%). Estos resultados superan a otros estudios previos que obtuvieron valores entre 75% y 85% y el modelo creado puede ser de utilidad para analizar la relevancia de otros métodos para el estudio de diferentes enfermedades o patologías, la detección temprana y adecuada, y minimizar el impacto de dichas enfermedades en la sociedad.

El autor en Ref. [2] pone en relevancia como los problemas de peso afectan negativamente a casi todas las funciones fisiológicas del cuerpo e incluyen una gran amenaza para la salud pública. utiliza un dataset para la predicción de los niveles de obesidad en personas de 14 a 61 años con diversos hábitos alimentarios y condiciones físicas en los países de Colombia, México y Perú. Crea un modelo de red neuronal convolucional donde se encuentran la importancias de todas las variables en el modelo expuesto, el modelo encontró las siguientes métricas de evaluación Accuracy (82%), F.Kappa (78%). También realiza análisis de la matriz de confusión, el autor recomienda estudiar la relación entre la obesidad y el Covid19 en términos de factores de riesgo con otro modelo de aprendizaje automático.

Los autores en ref [6] aplicaron la estrategia SMOTE para el balanceo de datos sintéticos, a medida de obtener más muestras para la característica target, con esta estrategia se logra conseguir que la data sintética tenga un patrón similar a la data real para poder ser utilizada con mayor precisión en los modelos.

Los autores en ref [5] los autores abordan el problema con el uso de algoritmos de Random Forest y Redes Neuronales Artificiales, utilizando una estrategia de optimización brindada por la herramienta WEKA con CV parameter selection, y comparan parámetros usados con parametros optimos para seleccionar el mejor resultado.

En [7] se presenta una comparación entre cuatro algoritmos de predicción los cuales son: árboles de decisión, clasificador bayesiano, clasificadores basados en reglas de asociación y las redes neuronales. En dicha investigación, se utiliza el cálculo del IMC como indicador del sobrepeso, como veremos en muchas de las investigaciones presentadas. Los resultados de esta investigación son prometedores, presentando altos porcentajes de exactitud de 83.6%, especificidad de 97.8% y sensibilidad de 12.3% para el de árboles de decisión, por ejemplo.

Del mismo modo, en [8], nos presentan algunos algoritmos utilizados para la predicción de la obesidad, tales como: k-NN, algoritmo de SVM, árboles de decisión, regresión logística, random forest, gradient boosting, entre otros. En dicha investigación los autores demuestran que el uso de la regresión logística tiene la mayor exactitud en cuanto a la predicción con un 97.09%, teniendo también alta precisión, mientras que por otro lado el algoritmo de gradient boosting presenta una exactitud de 64.08% lo cual no lo hace confiable. Esto nos demuestra, siendo una investigación actual, que uno de los algoritmos más confiables es el de regresión logística.

El autor en ref [9] presenta un artículo sobre la obesidad en Perú, la clasificación en grupos etarios, los porcentajes relacionados con género, prevención y relación con hábitos y comidas Patrones dietarios Actividad física mostrando una analitica entre la clasificación de muestras y aportando recomendaciones de para prevenir la obesidad.

Los autores en ref [10] muestran un análisis de las diferentes tendencias del sobrepeso en la sociedad peruana entre los años 2007 y 2014, dichas tendencias muestran una pendiente de crecimiento.

# DISEÑO DEL EXPERIMENTO

## Descripción del conjunto de datos

El dataset incluye datos para predecir la obesidad de personas dependiendo de algunos hábitos y preferencias de dieta.

El dataset contiene 2111 instancias de las cuales 77% fueron generadas de manera artificial con la herramienta Weka tool and SMOTE, y 23% de la data fue colectada a través de una encuesta online en un portal web donde participaron personas de México, Colombia y Perú.

Tabla 1

Descripción de atributos

| Atributo | Descripción | Valores |
| --- | --- | --- |
| Gender | Género de la persona | Male Female |
| Age | Edad de la persona | Valor Numerico |
| Height | Estatura de la persona | Valor Numerico |
| Weight | Peso de la Persona | Valor Numerico |
| Family history with Obesity | Cuenta con familiar con obesidad | Yes No |
| FAVC | Consumo de comida alta en calorias | Yes No |
| FCVC | Consumo de Vegetales | Always Sometimes Rarely |
| NCP | Número de comidas al dia | 1 2 3  more than 3 |
| CAEC | Consumo de alimentos entre comidas | Always Usually Sometimes Rarely |
| CH2O | Consumo de Agua | Less than a liter More than 2 liters Between 1 and 2 liters |
| CALC | Consumo de Alcohol | Less than a liter More than 2 liters Between 1 and 2 liters |
| SCC | Monitorea el consumo de calorías | Yes No |
| FAF | Frecuencia de actividad fisica | to 2 days 1 to 4 days 3 to 6 days 5 No physical activity |
| TUE | Tiempo de usao de dispositivos de tecnologia | 0 to 2 3 to 5 |
| MTRANS | Medio de transporte usado | Transportation used Public transportation Motorbike Bike Walking Automobile |
| Smoking | Si persona fuma o no | Yes No |
| NObesity | Clasificador de Obesidad | Insufficient Weight, Normal Weight Obesity Type I, Obesity Type II Obesity Type II Overweight Level I, Overweight Level II |

## Estadistica Descriptiva y visualización de datos

En la figura [Fig 2] se puede apreciar la distribución de la clasificación de obesidad según el género en los datos reales 23% del dataset.

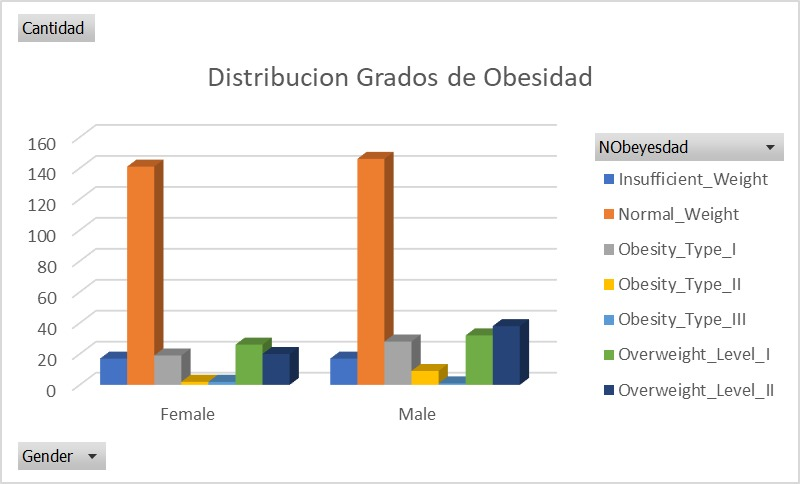


Fig 2. Distribución según grados de obesidad

En la figura 3 [Fig 3]. basado en los primeros 498 instancias que representan el 23% de la data se calculó el IMC y se puede ver su relación con género, donde se ve que la mayoría de las personas en la data muestran que no tienen obesidad ni sobrepeso.

Las características Height(Estatura en metros) y Weight(Peso en kilogramos), están relacionadas a la clasificación de obesidad debido a que mantienen una relación con el IMC (Índice de masa corporal)

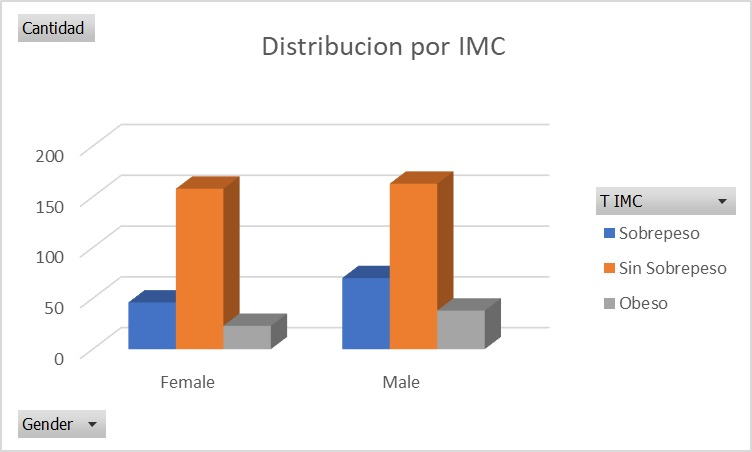


Fig 3. Distribución por IMC

En la figura 4 [Fig 4]. se puede apreciar la distribución de fumadores en base a su IMC, se puede notar una relación casi directamente proporcional.

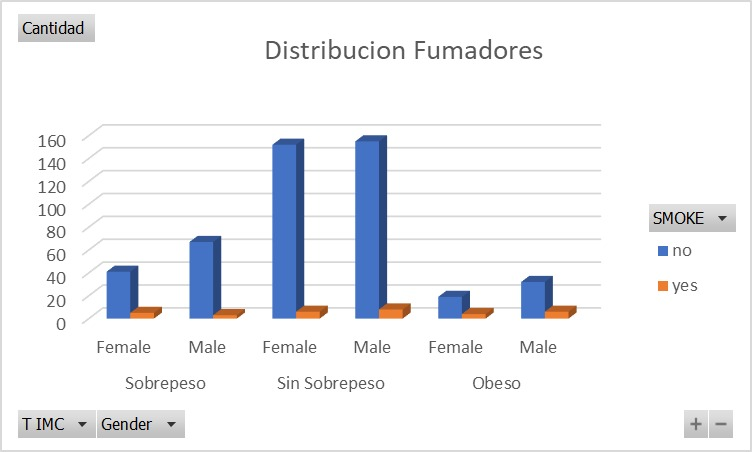


Fig 4. Distribución Fumadores - IMC

En la figura 5 [Fig 5] se observa una distribución similar a la de fumadores.

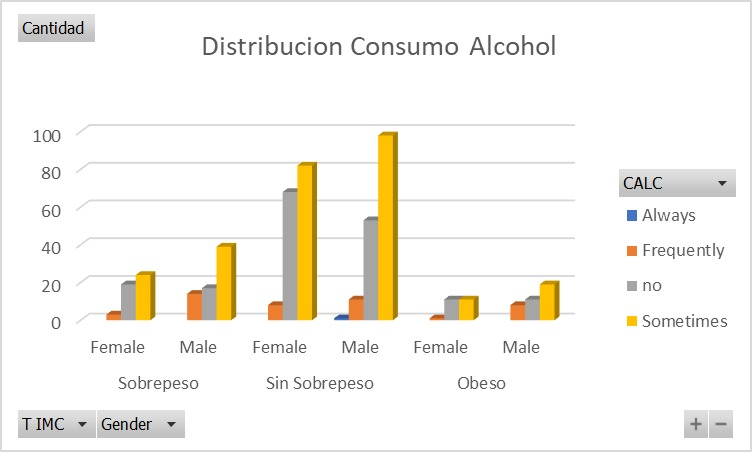


Fig 5. Distribución Consumo de Alcohol - IMC

En la figura 6 [Fig 6] Se observa un patrón similar al de la figura 5 en cuanto a la relación de consumo entre comidas.

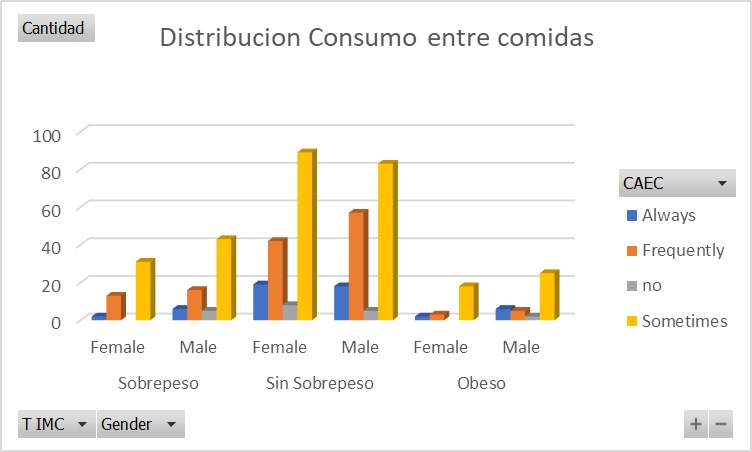


Fig 6. Distribución Consumo alimentos   
entre comidas - IMC

En la figura 7 se observa un patrón similar a la relación con la obesidad si algún familiar también padece de ella

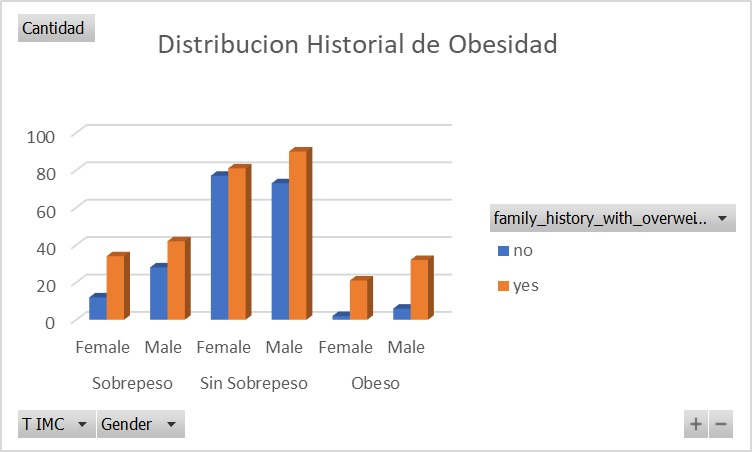


Fig 7. Distribución de IMC y familiar con obesidad.

## Metodología

Se ha agregado la variable índice de masa corporal, de acuerdo con la fórmula.



La relación entre el IMC y la categorización de obesidad es casi directa tal como muestra la Fig. 1. Se quita las variables asociadas al IMC, Estatura y Peso de las variables de estudio, por su aporte casi nulo al objetivo del problema.

Para el problema de agrupamiento o clusterización, se considerarán todas las variables, excepto Estatura, Peso y las categorías de obesidad.

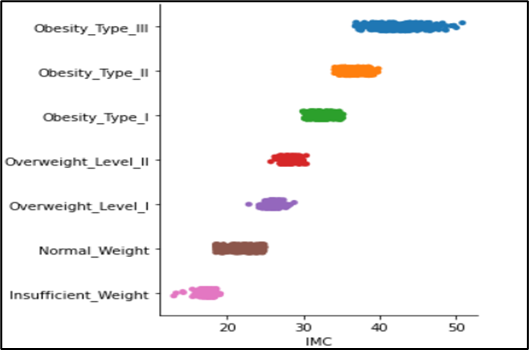


Fig. 8. Relación de IMC con la obesidad

Para validar los modelos predictivos se utilizarán las siguientes métricas

1. **Exactitud** **(Accuracy)**, se necesita que los modelos de clasificación clasifiquen de forma correcta de cada clase o tipo de obesidad.

2. **Precisión**, para entender intuitivamente el poder o habilidad de nuestro modelo predictivo para no equivocarse con falsos positivos.

3. **Sensibilidad** **(Recall)**, se necesita comprender la habilidad del clasificador para encontrar todos los verdaderos positivos.

4. **Matriz de confusión**, se necesita comprender correctamente si nuestro modelo tiene la capacidad discriminatoria de inferir la clase correcta de tipo de obesidad.

**Algoritmos empleados**

Los algoritmos que vamos a poner a prueba en el presente trabajo son: árboles de decisión (DT), Random Forest (RF) y una Red Neuronal Artificial (RNA), donde evaluaremos el mejor modelo utilizando las métricas de evaluación descritas anteriormente.

**Estrategia de validación para el ajuste de hiper parámetros**

Para los árboles de decisión, utilizaremos el cálculo de la medida Gini que nos servirá para determinar la impureza que tiene nuestro modelo.

En el caso de Random Forrest, una métrica importante de calcular es el Out Of Bag (OOB) la cual usaremos para este caso.

Para el Caso RNA, vamos a cambiar los siguientes hiperparámetros con el objetivo de encontrar un mejor modelo.

Número y tamaño de capas, Learning rate, Dropout, Función de Activación.

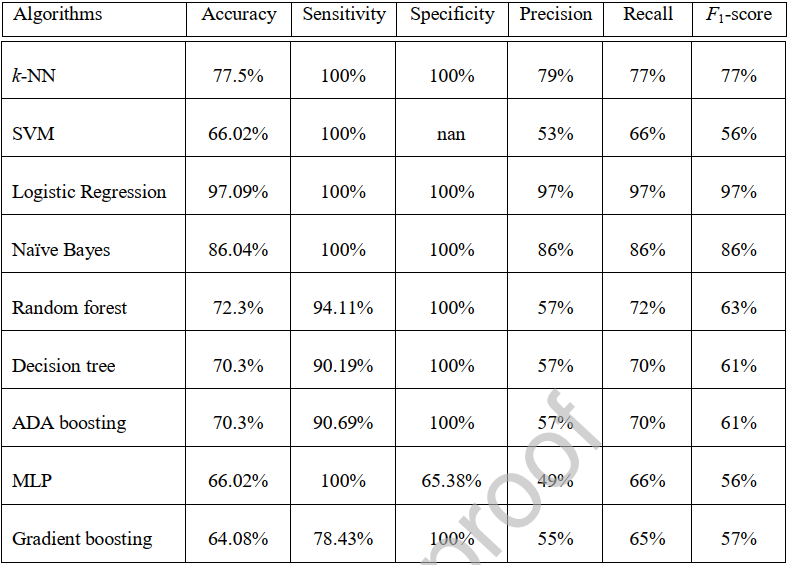
# EXPERIMENTACIÓN Y RESULTADOS

Modelos de predicción

En [8], se puede observar que los autores usan diversos modelos de predicción, entre ellos se encuentran el k-NN, SVM, Random forest, Decision Tree, Logistic Regression, entre otros. Lo cual se puede ver en la tabla 2.

Tabla 2

Comparativa de métricas de evaluación del estudio en [8]

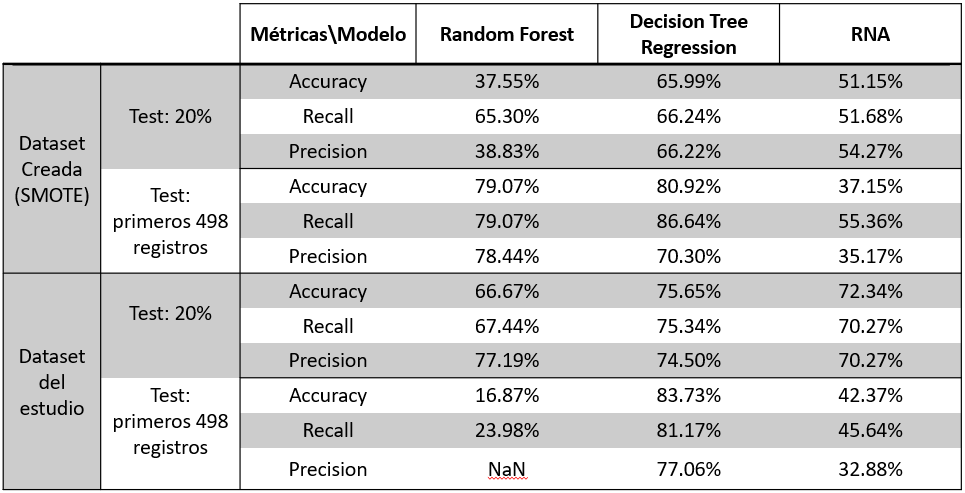


En base a las métricas de evaluación, determinan que Logistic Regression es el método adecuado para la predicción de sus datos obteniendo un 97.09%.

Siguiendo este mismo ejemplo, realizamos la evaluación de métricas de nuestros modelos a investigar (Random Forest, Decision Tree, RNA). Nosotros hemos considerado centrarnos en tres métricas de evaluación fundamentales, las cuales son accuracy, recall y precision. Se puede ver en la tabla 3 los resultados de dichas métricas mencionadas:

Tabla 3

Comparativa de métricas de evaluación del presente estudio



Podemos observar que obtenemos los mejores resultados en la dataset del estudio, donde consideramos utilizar como test los primeros 498 registros. Con estas características, se obtuvo un accuracy del 83.73%, recall de 81.17% y precision del 77.06% para el modelo de Árbol de decisión. Estos resultados se pueden evidenciar en la matriz de confusión, donde se ve que los datos se concentran en los cuadrantes de aciertos en la predicción, tal como se observa en la fig. 9

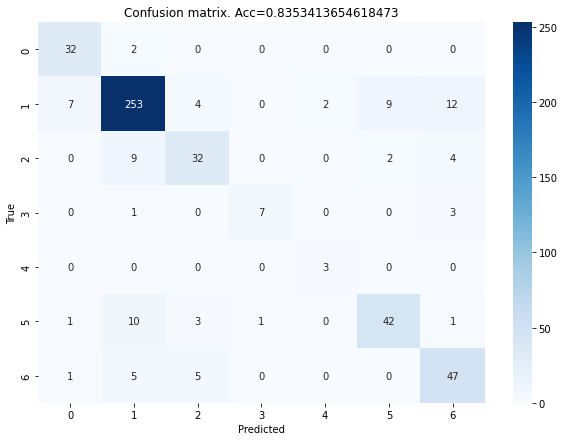


Fig 9. Matriz de confusión del mejor resultado obtenido (Árbol de decisión).

Modelo de Clustering

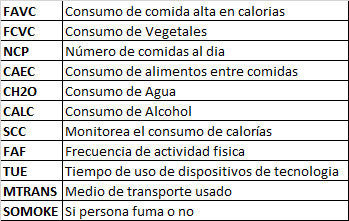
En [3] los autores crearon un modelo de segmentación con 4 clusters, considerando todas las variables del dataset, estos cuatro clusters son definidos con un perfil de persona.

En este estudio se está obviando las variables que no se consideran hábitos y solo se mantienen algunas condiciones físicas.

Finalmente se consideraron las variables para el modelo de segmentación descritas en la tabla 4.

Tabla 4

Atributos para el modelo de segmentación



El modelo de segmentación fue procesado con un número de clusters entre 2 y 5. Encontrándose que la segmentación para 2 grupos es la que mejor diferencia con el objetivo de la obesidad según se observa en el fig. 10

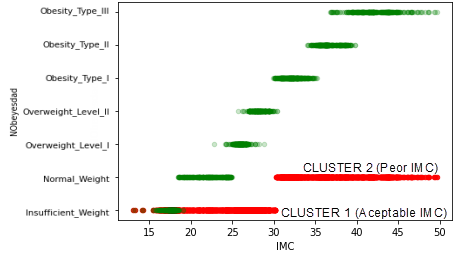
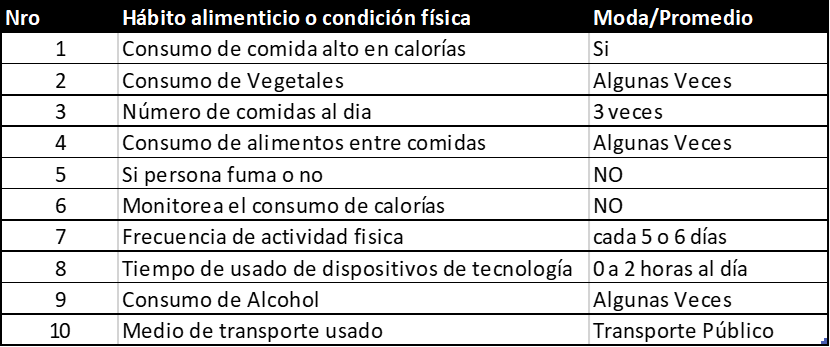


Fig. 10. Relación de IMC con niveles de obesidad y nuevo agrupamiento.

Finalmente se ha obtenido el perfil de una persona que pertenece al Cluster 1 utilizando el promedio y la moda según el tipo de dato evaluado.

Tabla 5

Perfil de una persona con nivel de IMC aceptable



# DISCUSIÓN

Modelo Random Forest

El Modelo random forest mejora al trabajar con data generada a través de SMOTE NC, el uso de datos discretos en los hábitos también mejoran la performance del modelo random forest.

Modelo Árbol de decisión

En el modelo de árbol de decisión se puede ver, por el contrario, que trabajando con la data del estudio se obtienen mejores resultados. Además, nos percatamos de que al escalar la data se distorsionan mucho los resultados, haciendo que las métricas no sean buenas, además entorpece la predicción, lo cual se evidencia en la matriz de confusión.

Modelo Red Neuronal

En el modelo de red neuronal artificial se ha encontrado métricas de evaluación mayor a 70%, pero al evaluarlos con la totalidad de datos encuestados (primeros 498), estas métricas caen alrededor del 45%, esto es una clara evidencia de Over fitting ya que el modelo clasifica bien los datos de entrenamiento pero no los datos reales. Esto debido a la cantidad de datos limitados y creación de datos sintéticos.

# CONCLUSIONES Y TRABAJOS FUTUROS

Conclusiones.

* Se ha encontrado un modelo aceptable, para predecir si una persona tiene un nivel de obesidad, conociendo los hábitos alimenticios y algunas condiciones físicas sin la necesidad de saber su estatura y peso.
* Una Red neuronal artificial no es la mejor herramienta de predicción para un dataset limitado y con datos sintéticos. Es mejor utilizar otros modelos.
* En el árbol de decisión de tipo regresión no se pudo escalar los datos, ya que esto generaba que nuestras métricas disminuyeran considerablemente, teniendo valores inferiores a 20%.
* No es recomendable utilizar la data sintética para validación.

Trabajos futuros.

* Encontrar relación de la obesidad con enfermedades cardiacas
* Estudiar la relación de hábitos alimenticios y condiciones físicas para predecir si una persona sufrirá grave sintomatología de COVID19.
* Trabajar los modelos utilizados en este trabajo con datasets más robustos y que contengan menos datos sintéticos, para la obtención de resultados más apegados a la realidad.

# REFERENCIAS

[1] S. Manna, A.M. Jewkes, Understanding early childhood obesity risks: an empirical study using fuzzy signatures, Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on, IEEE (2014, July), pp. 1333-1339

[2] Mehmet Kivrak, Deep Learning-Based Prediction of Obesity Levels According to Eating Habits and Physical Condition, Inonu University, Faculty of Medicine Dept. of Biostatistics and Medical Informatics, Malatya, Turkey 2021

[3] Rodolfo Cañas Cervantes y Ubaldo Martinez Palacio (2022), Estimation of obesity levels based on computational intelligence, Department of computer Science and Electronics, Universidad de la costa, CUC. Faculty Sytems Engineering Program, Colombia.

[4] Hera Siddiqui;Ajita Rattani;Nikki K. Woods;Laila Cure;Rhonda K. Lewis;Janet Twomey;Betty Smith-Campbell;Twyla J. Hill, A Survey on Machine and Deep Learning Models for Childhood and Adolescent Obesity, School of Computing, Wichita State University, Wichita, KS 67260, USA 2021.

[5] Asma Alqahtani, Fatima Albuainin, Rana Alrayes, Noura Al muhanna, Eyman Alyahyan and Ezaz Aldahasi, Obesity Level Prediction Based on Data Mining Techniques, International Journal of Computer Science and Network Security, Kingdom of Saudi Arabia 2021.

[6]Fabio Mendoza Palechor\*, Alexis de la Hoz Manotas

Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico

2019

[7] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, I. Buchan y J. Keane, «Comparing data mining methods with logistic regression in childhood obesity prediction,» Information Systems Frontiers, nº 11, p. 449–460, 2009.

| [8] F. Ferdowsy, K. S. Alam Rahi, I. Jabiullah y T. Habib, «A machine learning approach for obesity risk prediction» Current Research in Behavioral Sciences, vol. 00, 2021. |
| --- |

[9]La obesidad en el Perú Obesity in Peru Jaime Pajuelo-Ramírez

http://www.scielo.org.pe/scielo.php?script=sci\_arttext&pid=s1025-55832017000200012

[10]Análisis de la tendencia del sobrepeso y obesidad en la población peruana Analysis of the overweight and obesity trend in the Peruvian population

Carolina Tarqui-Maman, Doris Alvarez-Dongo[a](https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S2174-51452017000200006#aff1), Paula Espinoza-Oriundo[a](https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S2174-51452017000200006#aff1), Jose-Sanchez-Abanto https://scielo.isciii.es/scielo.php?script=sci\_arttext&pid=S2174-51452017000200006