Estimación del nivel de obesidad basado en los hábitos de alimentación y condiciones físicas

John Sebastian Estrada Durango
Departamento de Ingeniería de Sistemas
Universidad de Antioquia
Medellín, Colombia
johns.estrada@udea.edu.co

Steven Henao Mejía
Departamento de Ingeniería de Sistemas
Universidad de Antioquia
Medellín, Colombia
steven.henao@udea.edu.co

Resumen – Este documento trata del entrenamiento de un modelo de Machine Learning con el cual se busca estimar el nivel de obesidad basado en los hábitos de alimentación y condiciones físicas de cada individuo. Para el desarrollo del proyecto se implementó la técnica de aprendizaje de Máquinas de Soporte Vectorial, una estrategia de validación cruzada para hallar los mejores parámetros del modelo y una métrica de desempeño basada en balance accuracy.

Palabras clave – Machine Learning, validación, accuracy, MVS

I. INTRODUCCIÓN

Para este problema se busca clasificar el nivel de obesidad de cada individuo según los datos suministrados por el mismo. La obesidad tradicionalmente se le ha definido como un aumento en la proporción del tejido adiposo corporal, o bien como un aumento patológico del tejido adiposo en relación al tejido magro. La Organización Mundial de la Salud (OMS) define la obesidad como una acumulación anormal o excesiva de grasa que puede ser perjudicial para la salud. Esta clasificación se puede usar para el campo de la salud ya que ayudaría a realizar una primera clasificación de los pacientes y de esta manera, el personal de la salud podría determinar cuáles pacientes se deberían tratar con mayor importancia, dado que de acuerdo con datos de la OMS, hace diez años había en el mundo un aproximado de 330 millones de adultos obesos; en 2005 alcanzó los 400 millones de personas, y se calcula que para el año 2015 habrá por lo menos 2,300 millones de individuos con sobrepeso y más de 700 millones con obesidad.

Este documento cuenta con tres partes: En la parte II se describe el problema de ML que se desea solucionar, se enumeran y se explica la codificación de cada variable y se muestra el estado del arte. La parte III corresponde al entrenamiento y evaluación del modelo, donde se muestran los resultados obtenidos y la implementación de las técnicas.

II. COMPRENSIÓN DEL PROBLEMA DE ML

A. Descripción del problema

Para este problema se busca clasificar el nivel de obesidad de cada individuo según los datos suministrados relacionados a los diferentes hábitos de su rutina diaria. De esta manera, los resultados obtenidos se pueden usar para el campo de la salud ya que ayudaría a realizar una primera clasificación de los pacientes y así, el personal de la salud podría determinar la prioridad con la cuál deberían atender y dar tratamientos específicos a los pacientes según su estado de obesidad, e identificar los riesgos que esto implica en determinados tratamientos. Por lo tanto, se puede identificar que el problema a tratar es de clasificación.

B. Descripción de las variables del sistema

Variables de entrada[1]:

1. Gender: Es una variable categórica con dos posibles valores (female, male).
2. Age: Variable númerica.
3. Height: Variable numérica en metros.
4. Weight: Variable numérica en kilogramos.
5. history_with_overweight: Variable categórica con dos posibles valores (yes, no), define si tiene antecedentes familiares con problemas de obesidad.
6. FAVC: Variable categórica con dos posibles valores (yes, no), define si come comidas de altas calorías frecuentemente.
7. FCVC: Variable numérica con tres posibles valores (1 = never, 2 = sometimes y 3 = always), define la frecuencia en que se come vegetales.
8. NCP: Variable numérica con la cantidad de comidas principales que consume diariamente.
9. CAEC: Variable categórica con cuatro posibles valores (no, sometimes, frequently, always).
10. SMOKE: Variable categórica con dos posibles valores (yes, no).
11. CH2O: Variable numérica con la cantidad de agua que bebe diariamente en litros.
12. SCC: Variable categórica con dos posibles valores (yes, no), define si controla las calorías que consume a diario.
13. FAF: Variable numérica con la frecuencia en que realiza actividad física.

14. TUE: Variable numérica con cantidad de horas que le dedica a dispositivos tecnológicos.
15. CALC: Variable categórica con cuatro posibles valores (I do not drink, sometimes, frequently, always). Para saber la frecuencia en que bebe alcohol.
16. MTRANS: Variable categórica con cinco posibles valores (Automobile, motorbike, bike, public transport, walk).

Variable a predecir:

la variable a predecir es NObesity que cuenta con los siguientes valores: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II y Obesity Type III. Los cuales son el nivel de obesidad en que se encuentra la persona.

Por otra parte, el dataset no cuenta con valores faltantes, pero sí con muchos valores categóricos, lo que implicó realizar un preprocesado de los datos el cual se realizó implementando el método get_dummies() de pandas que consiste en convertir variables categóricas en variables ficticias o indicadoras [2]. Para el caso de la variable a predecir se reemplazaron los diferentes valores por números de 1 a 7 con el fin de identificar las clases, siendo 1 equivalente a Insufficient Weight y 7 a Obesity Type III.

C. *Estado del arte*

Se han desarrollado diferentes proyectos relacionados al problema de obesidad donde se implementan diversos métodos de Machine Learning (ML) para llegar a resultados que ayuden a identificar de manera más eficiente y precisa esta problemática.

Uno de los resultados obtenidos lo podemos identificar en el artículo [3], donde los autores especifican que la obesidad se ha convertido en una epidemia mundial y en la búsqueda de identificar los factores para predecir emergencias que puedan ocurrir a causa de estos, se implementó la metodología SEMMA, tres métodos de ML (Decision trees (J48), Bayesian network y Logistic Regression) y validación cruzada, donde los mejores resultados los obtuvo el modelo de Decision Trees con un 97.4% para la precisión, un TP rate del 97.8% y un FP rate del 0.2%. Por otro lado, en el artículo [4] lo que hicieron fue predecir la obesidad infantil después de los dos años con ayuda de un sistema de apoyo a las decisiones clínicas llamado CHICA, para este caso se implementaron modelos de ML como J48, mencionado en el caso anterior, Random Tree, RandomForest, ID3, Naïve Bayes, y Bayes trained. Dando como mejor resultado una precisión del 85% y sensibilidad del 89% para el modelo ID3, al igual que en el artículo anterior se usó la validación cruzada para obtener los mejores parámetros de los modelos. De igual manera, en el artículo [5] se plantea un framework, que utiliza Naïve Bayes para la predicción y Genetic Algorithm para la optimización de los parámetros, como solución aplicada al problema de predicción de la obesidad infantil obteniendo una precisión del 92%. Finalmente, en el artículo [6], los autores plantean la obesidad como un grave problema de salud pública a nivel mundial que aumenta el riesgo de diversas enfermedades. Por consiguiente, deciden recopilar diferentes tipos de datos e implementar técnicas de aprendizaje automatizado con el fin de extraer patrones ocultos para la predicción de la mejora del estado de la obesidad. Para tal fin, hacen uso de Recurrent Neural Networks (RNN) y obtienen una precisión del 80.7%.

III. Entrenamiento y evaluación

A. *Experimentos*

Para obtener los mejores resultados de los modelos y determinar los mejores parámetros para los mismos, se implementó la metodología de validación cruzada con 5 folds y así obtener los conjuntos correspondientes tanto para entrenamiento como para prueba. Para este caso se dejaron los valores típicos de la división que consisten en un 80% para entrenamiento y 20% para prueba.

Las métricas de desempeño que se usaron fueron Balanced Accuracy (BACC); que es una medida de clasificación normalizada con respecto al número de muestras por clase [7]. Se escoge esta medida de desempeño puesto que este problema es de clasificación ordinal ya que las muestras se valoran en una escala que va de 1 a 7 en este caso.

Este conjunto de datos incluye datos para la estimación de niveles de obesidad en individuos de los países de México, Perú y Colombia, con base en sus hábitos alimenticios y condición física. Los datos contienen 17 atributos y 2111 registros, los registros están etiquetados con la variable de clase NObesity [8]. En la fig. 1. podemos observar la distribución de muestras por clase: 272 muestras para la clase 1, 287 para la clase 2, 290 para las clases 3 y 4, 351 para la clase 5, 297 para la clase 6 y 324 para la clase 7.

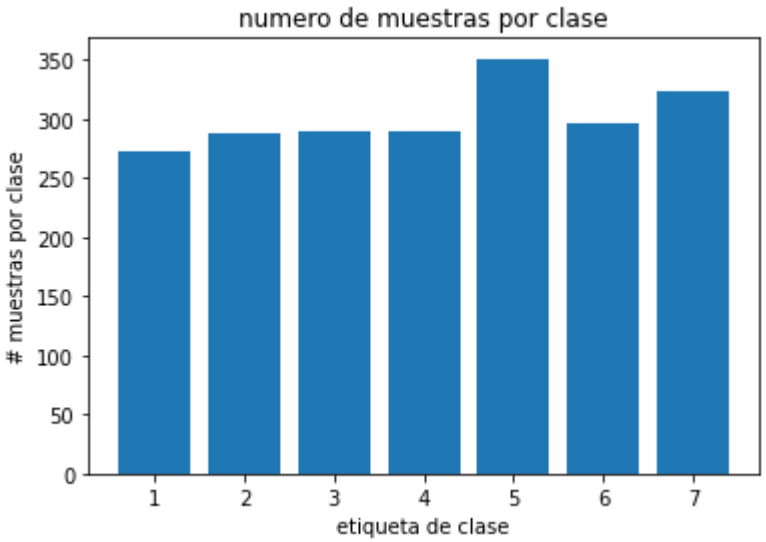

Fig. 1. Distribución de muestras por clase.

B. *Modelo*

Para este caso se utilizó la técnica de aprendizaje de Máquinas de Soporte Vectorial, la cual es un tipo de modelo de aprendizaje que permite encontrar la mejor solución usando como criterio de ajuste la maximización del margen, siendo el margen la distancia mas corta entre la frontera de decisión y cualquiera de las muestras. Las SVM se basan en riesgo estructural, Las muestra que se encuentran muy cercanas a la frontera de decisión se les llama vectores de soporte y se concideran dificiles ya que tienen un mayor grado de incertidumbre en comparación con las muestras que estan mas alejadas del margen, por esta razón son usadas para crear la frontera. El kernel,C y gamma son los parametors de las SVM, cuando en un problema se requiere hacer una división de las muestras pero no es posible llevarla a cabo entonces se transforma la dimensión mediante el kernel, cuando C es demasiado grande los erroes cuentan con mayor peso lo que aumenta el sobre ajustes provocando el aumento del margen, y el gamma es el coeficiente del kernel.

Los hiperparametros para útilizados para la fase de validación fueron:

Parametros	Valores
C	0.001, 0.01, 0.1, 1, 10
Gamma	0.01, 0.1, 1
Kernel	‘Linear’, ‘rbf’

C. *Resultados*

Los mejores parametros fueron:

Kernel	Gamma	C	E_training	E_test
Rbf	0.10	10	0.99	0.87

La Fig. 2. Muestra la matriz de confución obtenida luego de entrenar el modelo con los mejores parámetros.

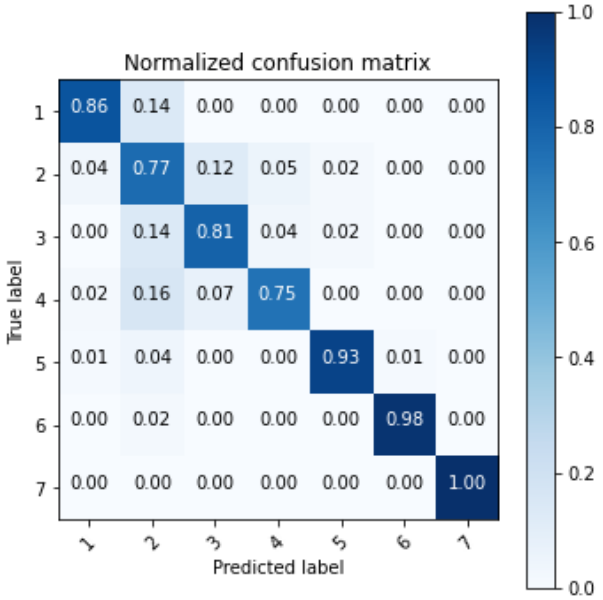

Fig. 2: Matriz de confución

REFERENCES

[1] Palechor, F. and Manotas, A., 2019. *Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico*. [online] Available at: <https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub> [Accessed 8 October 2021].

[2] Pandas.pydata.org. 2021. *pandas.get_dummies — pandas 1.3.3 documentation*. [online] Available at: <https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html> [Accessed 8 October 2021].

[3] De-La-Hoz-Correa, E., Mendoza Palechor, F., De-La-Hoz-Manotas, A., Morales Ortega, R. and Sánchez Hernández, A., 2019. *Obesity level estimation software based on decision trees*. [online] Repositorio.cuc.edu.co. Available at: <https://repositorio.cuc.edu.co/handle/11323/4176> [Accessed 8 October 2021].

[4] Dugan, T., 2015. *Machine Learning Techniques for Prediction of Early Childhood Obesity*. [online] Thieme. Available at: <https://www.thieme-connect.com/products/ejournals/html/10.4338/ACI-2015-03-RA-0036> [Accessed 8 October 2021].

[5] Muhamad, B., 2012. *A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction*. [online] Ieeexplore.ieee.org. Available at: <https://ieeexplore.ieee.org/abstract/document/6297254> [Accessed 8 October 2021].

[6] Qinghan, X., 2018. *Recurrent Neural Networks Based Obesity Status Prediction Using Activity Data*. [online] Ieeexplore.ieee.org. Available at: <https://ieeexplore.ieee.org/abstract/document/8614164> [Accessed 8 October 2021].

[7] Árias, J., 2020. *Multimodal and Multi-Output Deep Learning Architectures for the Automatic Assessment of Voice Quality Using the GRB Scale*. [online] Ieeexplore.ieee.org. Available at: <https://ieeexplore.ieee.org/document/8917580> [Accessed 8 October 2021].

[8] Mendoza, F. and Manotas, A., 2019. *UCI Machine Learning Repository: Estimation of obesity levels based on eating habits and physical condition Data Set*. [online] Archive.ics.uci.edu. Available at: <https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+> [Accessed 8 October 2021].