Factores de Riesgo de la Obesidad Desde el Aprendizaje Maquina

G. Efraín Condés Luna

Resumen

La obesidad y las enfermedades crónicas no transmisibles, son dos grandes problemas en el sector de salud pública en nuestro país. En este trabajo, empleamos algoritmos de aprendizaje automatizado, tales como regresión logística y *random forest*, para predecir la obesidad en una base de datos con información de salud de trabajadores de la UNAM. Además, hacemos una selección de modelos con validación cruzada para cada uno de los algoritmos antes mencionados, esto con el fin de obtener los factores más influyentes en el padecimiento de la obesidad y tratar de mejorar nuestro entendimiento sobre este padecimiento.

1. Introducción

La obesidad y el sobrepeso es un padecimiento que representa un reto a nivel global en términos de salud publica, puesto a la gran cantidad de gente que la padece al rededor del mundo y a la rapidez con la que ha aumentado. En México más del 70 % de la población padece de exceso de peso, ya sea obesidad y sobrepeso, hay más hombres con sobrepeso que mujeres y hay más mujeres con obesidad que hombres. Nuestro país ocupa el primer lugar en obesidad infantil y el segundo lugar en obesidad en adultos (solo por debajo de los Estados Unidos).

El sobrepeso y la obesidad son uno de los principales precursores de enfermedades crónicas no transmisibles como la diabetes. Las enfermedades crónicas son la principal causa de incapacidad prematura en adultos en México. Por lo antes mencionado y con la esperanza de mejorar la calidad y salud de la gente en México, entender la obesidad como un padecimiento multifactorial que afecta a las masas es de suma importancia para poder proponer soluciones a estos problemas.

En este trabajo se emplearan herramientas de minería de datos y aprendizaje automatizado sobre una base de datos que cuenta con información de antropometría, pruebas de laboratorio, antecedentes personales y familiares, entre otros de más de 1975 trabajadores de la UNAM, esto con la esperanza de obtener información que ayude a mejorar nuestro entendimiento sobre la obesidad.

2. Estado del Arte

3. Metodología

En este apartado daremos una revisión a las técnicas y algoritmos que usamos en el análisis de los datos. Puesto que en la clase se cubrió el tema de regresión logística, solo se dará revisión al algoritmo *nuevo* de *random forest*.

3.1. Prueba Binomial ε

En la base de datos identificamos a la variable que deseamos predecir como la clase C, para efectos de este proyecto dicha clase es la variable de obesidad, puesto que es la característica que deseamos predecir en las personas a través de las demás variables.

Aunque tengamos información sobre muchas variables de las personas, es de esperarse que no todas las variables tengan la misma importancia o que estén correlacionadas de la misma manera respecto a la clase objetivo C. Para poder tener una idea sobre la importancia que tienen las variables respecto a la clase C se realiza una prueba que consiste en lo siguiente. Consideremos una variable X, la cual podría ser la edad, estatura, etc. Y considerese una valor x de esta variable, 3 años, 25 años, etc. Si se quisiera comparar la probabilidad de encontrar a una persona de la clase C cuando se restringe una variable a un valor especifico

con la probabilidad de encontrar a una persona de la clase C a la probabilidad de encontrar una persona de la clase C sin que se restrinja ninguna variable, un ejemplo de esto sería comparar la probabilidad de encontrar a una persona obesa si se considera solo a las personas que tienen edad de 25 años a la probabilidad de encontrar a una persona a obesa sino se restringe la edad de dicha persona.

Si $P(C)$ es la probabilidad de encontrar a la posibilidad de encontrar a una persona de la clase C sin condicionar ninguna variable y $P(C|X = x)$ es la probabilidad de encontrar a una persona de la clase C dado que la persona tiene el valor x de la variable X. Al considerar la condición de que $X = x$. Debemos restringirnos solo a la población que cumple esa característica, definamos el numero de personas que cumplen esa condición por N_x, si la condición $X = x$ no afecta nada la probabilidad de estar en la clase C la fracción de personas de las N_x que está en la clase es $N_xP(C)$, sin embargo, considerando la condición de $X = x$, la fracción de personas que están en la clase es $N_xP(C|X = x)$, para comparar estos dos casos tomamos la diferencia de la gente que está en la clase en ambos casos

$$N_xP(C|X = x) - N_xP(C) = N_x(P(C|X = x) - P(C)). \tag{1}$$

Notemos que de esta expresión que si al restringir $X = x$ se aumenta la probabilidad de estar en la clase C entonces la expresión 1 será positiva y si en cambio disminuye la probabilidad será negativa. Esto nos dice si el valor x está correlacionado positiva o negativamente, sin embargo no nos da una noción precisa de que tan grande es esta correlación. Para poder cuantificar la diferencia entre $N_xP(C|X = x)$ y $N_xP(C)$ debemos usar una medida propia de la distribución del problema, por lo tanto usaremos la desviación estándar de la distribución. Calcularemos por cuantas desviaciones estándares $N_xP(C|X = x)$ se desvía de $N_xP(C)$. A dicha cantidad la llamaremos ε y está dada la siguiente ecuación

$$\varepsilon = \frac{N_x(P(C|X = x) - P(C))}{\sqrt{N_x(P(C)(1 - P(C)))}}. \tag{2}$$

El valor de ε se puede calcular, dad una clase C, para cada valor x de cada variable X. A continuación mencionaremos como se calculó la ε ya en la practica, N_x es el numero total de personas cuyo valor de la variable X es x, $P(C|X = x) = N_{cx}/N_x$ donde N_cx es el numero de personas que pertenecen a la clase C y cuyo valor de la variable X es x y $P(C)$ es la fracción que representa la clase C del total de la muestra.

Como se menciona en [1], podemos usar al valor de ε como una medida qué impacto tiene un valor de un atributo sobre la clase objetivo C. Siendo que valores de $\varepsilon > 2$ indican que dicho valor de la variable es predictiva para la clase y valores de $\varepsilon < 2$ indican que es predictiva para el complemento de la clase C^c. Como veremos más adelante, los valores de ε serán utilizados para hacer una primera selección de variables para los modelos predictivos de cada clase.

3.2. Arboles de Decisión & *Random Forest*

Los arboles de decisión son modelos estadísticos que son usados en clasificación, el cual tiene la característica de representar los resultados en una estructura de árbol, en el cual cada nodo divide los datos usando los distintos valores de una variable. Para clasificar una muestra $\mathbf{X} = \mathbf{x}$ se hace pasara dicha muestra a través del árbol obtenido en el entrenamiento del algoritmo, desde el nodo raíz, que representa la variable más importante para la predicción, hasta alguna hoja que representa a una clase.

La forma en la que se crea el árbol puede expresarse de la siguiente forma [2]:

1. Se ponen todos los datos de entrenamiento en la raíz.

2. Se selecciona la *mejor* variable para dividir los datos usando medidas estadísticas.

3. Se van particionando recursivamente los datos de entrenamiento hasta que queden solo elementos de la misma clase, que solo quede un elemento o que se terminen las variables para dividir los datos.

Existen varias medidas estadísticas para seleccionar la *mejor* variable, la que usamos en los arboles de decisión en este trabajo emplean la ganancia de Gini. Para definir la ganancia Gini primero definamos la impureza Gini, para un conjunto de datos D, en el cual se tienen un total de N_C clases en las cuales clasificar, la impureza Gini se define como la probabilidad de clasificar incorrectamente a un elemento del conjunto de datos que se ha escogido de forma aleatoria, la cual es

$$G(D) = \sum_{C \in Clases} p(i)(1 - p(i)). \tag{3}$$

Con esto podemos definir la ganancia Gini de la siguiente manera, supongamos que en un nodo t del árbol se tiene un conjunto de datos D y queremos calcular la ganancia Gini que se obtiene al separar los datos respecto a su valor en la variable X que puede tomar los valores $x_1, x_2, ..., x_n$. Al separar el conjunto D, en un nodo t, por los valores de la variable X se obtienen D_i subconjuntos de los datos D con $i = 1, ..., n$, de esta forma la ganancia Gini se calcula de la siguiente manera

$$Gain(t, X) = G(D) - \sum_{i=1}^{n} w_i G(D_i), \quad \text{donde} \quad w_i = \frac{N_i}{N} = \frac{|D_i|}{|D|}. \tag{4}$$

Entre mayor sea la ganancia Gini mejor será la separación que da la variable X. De esta manera, al principio de la construcción del árbol, en la raíz, se dividirán los datos con la variable que de la mejor ganancia Gini, después para cada nodo se escogerá, de entre las variables que queden, la variable que tenga la mejor ganancia Gini y así sucesivamente hasta que queden solo elementos de la misma clase, que solo quede un elemento o que se terminen las variables para dividir los datos.

Una vez que se tenga completo el árbol de decisión para clasificar a una combinación de todas las variables $\mathbf{X} = \mathbf{x}$ se seguirá el camino que indique el árbol de decisión y se le clasificará con la clase que predomine en la hoja en la que acabe.

El algoritmo de *random forest* consiste en crear un bosque $\mathcal{B}$ que resulta de entrenar N_T arboles de decisión, de tal forma que cada árbol T sea entrenado con un subconjunto D_T escogido de forma aleatoria del conjunto de datos completos D. Además de que se escogen de forma aleatoria las variables que de entrada que se consideran para dividir los nodos de t [3]. Después de entrenar los N_T arboles para clasificar una muestra $\mathbf{X} = \mathbf{x}$ se clasifica en todos los árboles de decisión y se toma como la clase a la clase que predomine en todas la clasificaciones que dieron los arboles de decisión.

El algoritmo de *random forest* nos permite asignar una importancia a cada una de las variables para la clasificación empleando la ganancia de Gini [3], si definimos la importancia de una variable X en un árbol T como

$$Imp_T(X) = \sum_{t \in \mathcal{N}(T,X)} p(t) Gain(t, X), \tag{5}$$

donde $\mathcal{N}(T, X)$ es el conjunto de nodos de T en los cuales se usó la variable X para dividir los datos y $p(t) = N_t/N$, siendo N_t el numero de datos que llegan al nodo t y N el total de datos. Usando la ec.5 podemos definir la importancia de una variable en un *random forest* $\mathcal{B}$ como

$$Imp(X) = \frac{1}{N_T} \sum_{T \in \mathcal{B}} Imp_T(X). \tag{6}$$

Podemos entender a $Imp(X)$ como la suma de las ganancias de Gini pesadas $p(t)Gain(t, X)$ para todos los nodos donde X fue usada, promediada obre todos los N_T arboles del bosque.

4. Resultados

4.1. Selección de Variables A Través del Valor ε

Definir a la clase de obesidad, es decir al subconjunto de trabajadores que padece obesidad, es bastante sencillo, si una persona tiene un indice de masa corporal (IMC) mayor o igual a 30 se le etiquetará como obesa.

Para hacer el análisis que a continuación se presenta se descartaron ciertas variables de la base de datos. Se quitaron las variables que de conocerse su valor puede saberse automáticamente si la persona es obesa o no, tales como (`AIMC`, `Aedad_obes`, `dich_obes`), además, se removieron las variables de carácter antropométrico que predicen *muy bien* a la clase de obesidad, esto se hizo removiendo las variables antropométricas que tengan algún valor con un valor de $\varepsilon > 2$. Descartar estas variables para el análisis tiene dos propósitos, el primero es no usar información redundante, no tiene sentido usar información que nos permita saber si una persona es obesa *a priori*, el segundo es permitir que otras variables que tal vez no son *tan evidentes* se manifiesten como predictoras de la clase de obesidad. El conjunto de variables removidas, bajo el criterio antes mencionado, de la base de datos es la siguiente:

- `AIMC`: Categorización del indice de masa corporal.
- `IMC`: Indice de masa corporal.
- `Aedad_obes`: Edad a la que se fue diagnosticado con obesidad.
- `dich_obes`: Variable que indica si la persona ha sido diagnosticada con obesidad o no.
- `diastolica`: Medida de la presión diastólica
- `sistolica`: Medida de la presión sistólica
- `Acintura`: Medida de la cintura.
- `Apeso`: Medida del peso.
- `Atalla`: Medida de la estatura de la persona.
- `Abrazo`: Medida de la estatura de la persona.

Habiendo calculando el valor de ε para cada valor de cada variable en la base de datos respecto a la clase de obesidad, y descartando las variables mencionadas anteriormente, seleccionaremos las 30 variables que tengan los valores de ε más altos. Solo nos quedamos con las *mejores* 30 para reducir la dimensionalidad del conjunto de datos a analizar, de tal manera que las implementaciones de técnicas de selección de modelos y de validación que usaremos más adelantee se puedan llevar a cabo en un tiempo de computo razonable.

Las 30 variables *más predictivas* fueron las siguientes:

1. `peso_act`: Autovaloración de la persona sobre su peso actual (muy malo, malo, bueno, regular, etc.).
2. `Ainsulina`: Medida de insulina.
3. `peso1`: Autovaloración de la persona sobre su peso hace 1 año (muy malo, malo, bueno, regular, etc.).
4. `peso_ehoy`: Autovaloración de la persona sobre su peso en la actualidad (1-10).
5. `peso5`: Autovaloración de la persona sobre su peso hace 5 años.
6. `aes_peso`: Estimación en kilogramos que la persona tiene sobre su peso actual.
7. `sindro_meta`: Indica si la persona padece de sindrome metabolico.
8. `dich_hiper`: Indica si la persona ha sido diagnosticada con hipertensión.
9. `condi_act`: Autovaloración de la persona sobre su condición física actual.
10. `Aedad_hiper`: Edad a la que la persona fue diagnosticada con hipertensión.
11. `peso10`: Autovaloración de la persona sobre su peso hace 10 años.
12. `Ahba`: Nivel de hemoglobina.
13. `peso_acc`: Acciones que la persona desearía tomar respecto a su peso (bajar, subir, etc.).

14. `salud_act`: Autoevaluación de la persona sobre su salud actual (bueno, malo, muy malo, etc.).
15. `estatura`: Estimación que la persona tiene sobre su estatura.
16. `condi1`: Autoevaluación de la persona sobre condición física hace un año (bueno, malo, muy malo, etc.).
17. `mad_diab`: Indica si su madre es diabetica.
18. `id_gestud`: Grado de estudios.
19. `tgb_com`: Medida de los triglicéridos.
20. `com_relrec`: Estimación de la persona sobre cuanto come en relación a lo recomendado.
21. `glu_com`: Medida de la glucosa.
22. `Apuesto`: Indica el puesto que tiene la persona en la UNAM.
23. `Aher6_edad`: Edad del sexto hermano.
24. `estres10`: Autoevaluación del nivel de estrés que sufría hace 10 años.
25. `Aedad_diab`: Edad a la que fue diagnosticado con diabetes.
26. `condi5`: Autoevaluación de la persona sobre su condición física hace 5 años.
27. `Ahij_sobrepeso`: Numero de hijos que tienen sobrepeso.
28. `Atio_diabeticos`: Numero de tíos diabéticos.
29. `Ahij2_edad`: Edad del segundo hijo.
30. `locout20`: Numero de veces que comía la persona en locales fuera de la UNAM hace 20 años.

El lector deberá tener presente que los modelos y discusiones subsecuentes partirán de la selección de variables que hemos hecho aquí.

4.2. Regresión Logística

Se realizó una selección de modelos usando una regresión logística usando validación cruzada de 50 particiones, el código de todo el trabajo puede encontrarse en [4]. El método para seleccionar las variables consiste en elegir primero al mejor modelo con una variable después se seleccionó al mejor modelo de dos variables que incluya a la primera seleccionada y así sucesivamente hasta las 30 variables y se seleccionó al mejor modelo de entre esos 30 modelos seleccionados.

La fig. 1 muestra la gráfica de los errores en función del numero de variables en el modelo.

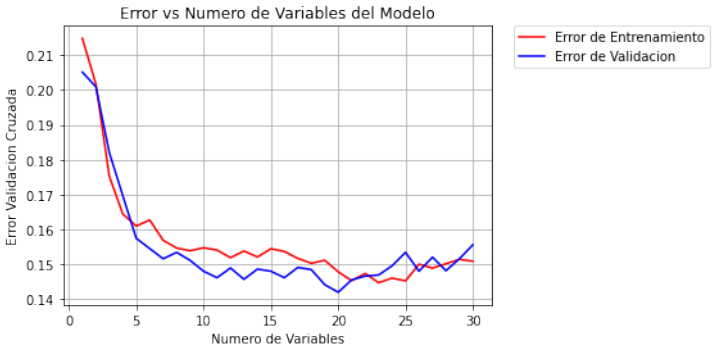

Figura 1: Error vs Numero de Variables del Modelo usando RL.

El modelo que tuvo el mejor rendimiento fue el que ocupó las variables `condi1`, `peso_ehoy`, `Ainsulina`, `aes_peso`, `peso1`, `tgb_com`, `peso5`, `peso_act`, `glu_com`, `mad_diab`, `Ahba`, `com_relrec`, `peso_acc`, `Aedad_diab`, `salud_act`, `condi_act`, `estatura`, `sindro_meta`, `id_gestud` y `locout20`. Dicho modelo tuvo el siguiente rendimiento:

- El modelo predijo en promedio correctamente el 84.56 % de los datos de prueba.

- El modelo tuvo en promedio un error cuadrático medio del 0.15 en los datos de prueba.

Estos valores y los que se presentarán más adelante para los demás algoritmos se obtuvieron con una validación cruzada de 100 particiones.

4.3. *Random Forest*

Para saber cuantos arboles son *suficientes* para entrenar los clasificadores con *random forest* hicimos una gráfica para ver la convergencia del porcentaje de aciertos contra el numero de arboles empleados en el entrenamiento del *random forest*, con el fin de ver en que valores empieza a converger el porcentaje de aciertos y saber que numero de arboles usar en las pruebas posteriores. En la fig. 2 podemos ver la gráfica de dichos cálculos, puede verse que el porcentaje de aciertos empieza a converger para valores mayores que 2000 por lo tanto el numero de arboles que usaremos en los modelos que mencionaremos a continuación serán entrenados invariablemente con este numero de arboles.

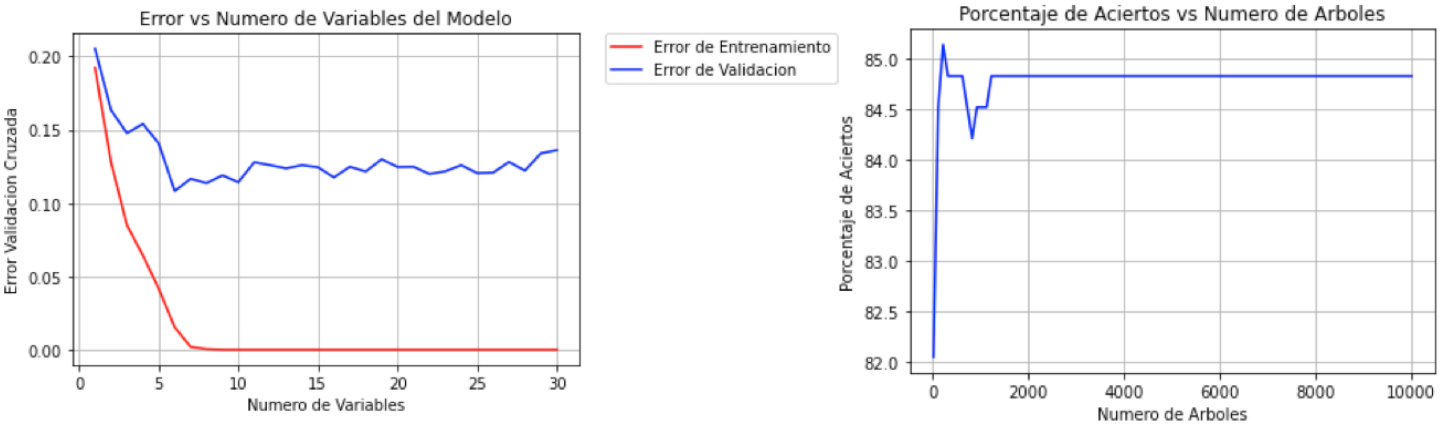

Figura 2: Gráfica de error vs. numero de variables (izquiera) y del porcentaje de errores vs. el numero de arboles empleados para el entrenamiento del *random forest*.

Para la selección de modelos en el caso del *random forest* se hizo lo siguiente, se obtuvo la lista de las variables ordenada por su importancia, se evaluó con validación cruzada el modelo que tenía a ls primer varabile más importante, después a la que tenía a las dos más importantes y así sucesivamente, al final se eligió entre todos estos modelos al que tuviera el mejor desempeño. En la fig. 2 se pueden ver las curvas de los errores en los conjuntos de entrenamiento y de validación en función del numero de variables del modelo.

El modelo que tuvo el mejor rendimiento fue el que ocupó las variables `aes_peso`, `peso_ehoy`, `Ainsulina`, `peso_act`, `peso1` y `estatura`. Dicho modelo tuvo el siguiente rendimiento:

- El modelo predijo en promedio correctamente el 88.56 % de los datos de prueba.
- El modelo tuvo en promedio un error cuadrático medio del 0.11 en los datos de prueba.

5. Conclusiones

- De los modelos encontrados el que tuvo el mejor rendimiento fue el que usó *random forest* y las variables `aes_peso`, `peso_ehoy`, `Ainsulina`, `peso_act`, `peso1` y `estatura`, logrando un 88.56 % de aciertos en los datos de prueba. El mejor modelo usando regresión logística fue el que ocupó las variables `condi1`, `peso_ehoy`, `Ainsulina`, `aes_peso`, `peso1`, `tgb_com`, `peso5`, `peso_act`, `glu_com`, `mad_diab`, `Ahba`, `com_relrec`, `peso_acc`, `Aedad_diab`, `salud_act`, `condi_act`, `estatura`, `sindro_meta`, `id_gestud` y `locout20`, obteniendo un 84.56 % de aciertos en los datos de prueba.

- Las variables que tienen en común la selección de variables hecha por los dos algoritmos son `Ainsulina`, `aes_peso`, `estatura`, `peso1`, `peso_act`, `peso_ehoy`. Todas estas variables, a excepción de la `Ainsulina` que es la medida de la insulina, tienen que ver con la autovaloración que tienen las personas sobre su cuerpo, en especifico sobre su peso y estatura. Esta es una situación notable, puesto que dichas variables son más de índole cualitativo, a diferencia de las variables que representan pruebas de laboratorio que son totalmente cuantitativas. Esto resulta favorable si se pensara implementar algún programa de salud en contra de la obesidad, puesto que la gente parece estar consciente de su estado de salud, hablando dentro el marco de la obesidad, es decir, en caso de planearse un programa en contra de la obesidad, convencer a la gente de que padece de obesidad y que ésta afecta su salud no debería ser unos de los principales problemas a resolver.

- El hecho de que la variable `Ainsulina` aparezca en la intersección de las dos selecciones de variables puede ser un indicador de como es que el padecimiento de la diabetes está tan correlacionado con la obesidad.

- Además de variables que tienen que ver con la autovaloración de las personas sobre su salud y de variables que tienen que ver con enfermedades crónicas o pruebas de laboratorio, entre las selecciones de variables aparecen variables como `id_gestud`. La explicación a que esta variable aparezca podría deberse a que dependiendo del grado de estudios que tiene la persona será el puesto que tenga como trabajador de la UNAM y dependiendo del puesto que tenga en la UNAM tendrá una rutina diaria mas o menos sedentaria y con unos u otros hábitos alimenticios que son variables que pueden ser predictoras *intuitivas* de la obesidad. Sin embargo, las hipótesis aquí planteadas son mera especulación, para futuros estudios se podrían realizar pruebas de correlación entre las variables mencionadas para la clase de obesidad y así comprobar o desmentir lo antes mencionado.

Referencias

[1] Ana E. Ruiz L, Christopher R. Stephens S & Hugo Flores, *Una generalización del clasificador Naive Bayes para usarse en bases de datos con dependencia de variables.* CRC Press, 2017.

[2] Mohssen M. et al. *Machine Learning Algorithms and Applications.*

[3] Louppe G., Wehenkel L., Sutera A. & Geurts P. *Understanding variable importances in Forests of randomized trees. Advances in Neural Information Processing Systems.* 26, 2013.

[4] https://github.com/efracondes/proyecto_machine_learning