Aplicación de Redes Bayesianas en el Análisis de la Obesidad

Luis Diego Pari Benito
David Salomon Lopez Ticona
Universidad Nacional Del Altiplano
luisdiegopari@gmail.com —— cuentaofiyotu1@gmail.com
Puno

Abstract

Este artículo explora la aplicación de redes bayesianas para clasificar y predecir los niveles de obesidad en individuos, utilizando variables como edad, peso, antecedentes familiares de sobrepeso, frecuencia de consumo de comida rápida, actividad física y medio de transporte. El objetivo principal es evaluar la eficacia de las redes bayesianas como herramienta para modelar las relaciones entre estas variables y clasificar los niveles de obesidad.

La metodología se basó en la construcción de una red bayesiana a partir de un conjunto de datos que incluyó la definición de la estructura, el aprendizaje de los parámetros y la validación del modelo. Los datos fueron divididos en conjuntos de entrenamiento y prueba, y el rendimiento del modelo fue evaluado mediante métricas como precisión, sensibilidad y especificidad, obtenidas a partir de la matriz de confusión.

Los resultados mostraron que la red bayesiana alcanzó una precisión global del 62.66%, con niveles destacados de sensibilidad en clases como *Obesity Type II* (74.14%) y *Obesity Type III* (76.56%). Sin embargo, el modelo presentó dificultades al clasificar correctamente casos de *Overweight Level II*. Estos hallazgos reflejan la capacidad del modelo para capturar relaciones relevantes entre las variables, aunque se identificaron áreas para mejorar.

En conclusión, las redes bayesianas demostraron ser una herramienta útil y prometedora para clasificar los niveles de obesidad, con potencial para mejorar su desempeño mediante el ajuste de los datos y la estructura del modelo.

Palabras clave: redes bayesianas, obesidad, clasificación, probabilidad condicional, predicción.

1 Introducción

La obesidad es una de las principales preocupaciones de salud pública a nivel mundial, debido a su creciente prevalencia y su asociación con enfermedades crónicas como la diabetes, hipertensión y enfermedades cardiovasculares. Según la Organización Mundial de la Salud (OMS), el porcentaje de personas con obesidad ha aumentado significativamente en las últimas décadas, lo que subraya la necesidad de desarrollar herramientas efectivas para su análisis y prevención.

Las Redes Bayesianas son modelos probabilísticos gráficos que permiten representar y analizar relaciones de dependencia entre múltiples variables. Estas herramientas han demostrado ser útiles en diversas áreas, como la medicina, el análisis de riesgos y la inteligencia artificial, debido a su capacidad para manejar incertidumbre y datos incompletos. En el contexto de la obesidad, las Redes Bayesianas ofrecen un enfoque prometedor para modelar factores determinantes como los antecedentes familiares de sobrepeso, hábitos alimenticios, actividad física y otros aspectos del estilo de vida, permitiendo además realizar predicciones sobre el nivel de obesidad.

Estudios previos han explorado el uso de modelos probabilísticos para analizar problemas relacionados

con la obesidad, mostrando resultados prometedores en términos de clasificación y predicción. Sin embargo, existe una necesidad de continuar evaluando estas herramientas en escenarios reales y con conjuntos de datos complejos.

El objetivo de este artículo es aplicar Redes Bayesianas para clasificar y predecir los niveles de obesidad en individuos, evaluando su capacidad para capturar relaciones relevantes entre variables y su desempeño mediante métricas de evaluación como precisión, sensibilidad y especificidad. Este análisis busca contribuir al desarrollo de modelos más robustos para la identificación de factores clave en el manejo y prevención de la obesidad.

2 Metodología

2.1 Población y Datos

El análisis se llevó a cabo utilizando un conjunto de datos compuesto por 2,111 registros, provenientes de un estudio sobre obesidad. Este conjunto de datos incluye las siguientes variables clave:

- **Antecedentes familiares de sobrepeso (FamHist_Overwt):** Historial familiar relacionado con la obesidad, categorizado como "yes" o "no".
- **Edad (Age):** Discretizada en rangos etarios (*Child, Young Adult, Adult, Middle Age, Senior*).
- **Peso (Weight):** Dividido en categorías significativas (*Underweight, Normal, Overweight, Obese, Severely Obese*).
- **Frecuencia de consumo de comida rápida (FAVC):** Indicador binario sobre la frecuencia de este hábito alimenticio.
- **Frecuencia de actividad física (FAF):** Clasificada en niveles de frecuencia baja, moderada o alta.
- **Medio de transporte (MTrans):** Método principal de transporte, como caminar, vehículo privado o transporte público.
- **Nivel de obesidad (NObeyesdad):** Variable objetivo, con siete posibles categorías: *Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, Obesity Type III.*

El conjunto de datos fue dividido en dos subconjuntos: el 80% se utilizó para entrenar el modelo y el 20% para su validación. Las variables numéricas, como *Age* y *Weight*, fueron discretizadas en categorías para garantizar la compatibilidad con el modelo discreto de Redes Bayesianas.

2.2 Construcción de la Red Bayesiana

La red bayesiana se construyó con base en un análisis previo de las relaciones entre las variables. Se estableció una estructura inicial que conecta las variables predictoras (*FamHist_Overwt, Age, Weight, FAVC, FAF, MTrans*) con la variable objetivo (*NObeyesdad*). La probabilidad conjunta del modelo se definió como:

$$P(A, B, C, D, E, F, G) = P(A) \cdot P(B|A) \cdot P(C|B, A, D, E, F) \cdot P(D) \cdot P(E) \cdot P(F) \cdot P(G|B, C, A, D, E, F)$$

Donde cada variable representa:

- A: Antecedentes familiares de sobrepeso (*FamHist_Overwt*),
- B: Edad (*Age*),
- C: Peso (*Weight*),
- D: Frecuencia de consumo de comida rápida (*FAVC*),
- E: Frecuencia de actividad física (*FAF*),
- F: Medio de transporte (*MTrans*),
- G: Nivel de obesidad (*NObeyesdad*).

La estructura y los parámetros del modelo se implementaron utilizando la librería `bnlearn` en el lenguaje R. La Figura A.1 muestra la red bayesiana construida, con arcos que representan las relaciones condicionales entre las variables.

2.3 Validación del Modelo

Para evaluar el desempeño del modelo, se utilizó el conjunto de datos de prueba. La variable objetivo (*NObeyesdad*) fue predicha a partir de las variables independientes, y los resultados se compararon con los valores reales mediante una matriz de confusión. Esta permitió calcular métricas clave como:

- **Precisión global:** Proporción de predicciones correctas entre todas las observaciones.
- **Sensibilidad:** Capacidad del modelo para identificar correctamente cada clase de obesidad.
- **Especificidad:** Capacidad del modelo para identificar correctamente los casos que no pertenecen a una clase específica.

Los resultados de la matriz de confusión y las métricas calculadas se presentan en la sección de *Resultados*.

3 Resultados

3.1 Estructura de la Red Bayesiana

La red bayesiana construida incluye siete nodos principales que representan las variables seleccionadas:

- **FamHist_Overwt (A):** Historia familiar de sobrepeso.
- **Age (B):** Edad del individuo, clasificada en rangos etarios (*Child, Young Adult, Adult, Middle Age, Senior*).
- **Weight (C):** Peso del individuo, dividido en categorías (*Underweight, Normal, Overweight, Obese, Severely Obese*).
- **FAVC (D):** Frecuencia de consumo de comida rápida.
- **FAF (E):** Frecuencia de actividad física, clasificada en baja, moderada y alta.
- **MTrans (F):** Medio de transporte utilizado (*Walking, Public Transportation, Private Vehicle*).
- **NObeyesdad (G):** Nivel de obesidad, la variable objetivo, con siete categorías (*Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, Obesity Type III*).

La Figura A.1 muestra la estructura de la red bayesiana construida, donde los arcos representan dependencias condicionales entre las variables.

3.2 Inferencia Probabilística

A partir de la red bayesiana, se realizaron inferencias probabilísticas para predecir el nivel de obesidad (*NObeyesdad*) de los individuos considerando las siguientes configuraciones de evidencia:

- **Caso 1:** Un individuo joven (*Age = Young Adult*) con peso normal (*Weight = Normal*), antecedentes familiares de obesidad (*FamHist_Overwt = Yes*), alta frecuencia de comida rápida (*FAVC = Yes*) y transporte público (*MTrans = Public*). La red estimó una probabilidad del 78% de que este individuo pertenezca a la categoría *Obesity Type I*.
- **Caso 2:** Un individuo de mediana edad (*Age = Middle Age*), con sobrepeso nivel II (*Weight = Overweight Level II*) y actividad física moderada (*FAF = Moderate*), mostró una probabilidad del 15% de ser clasificado como *Obesity Type II*.

Estos resultados destacan cómo la red puede integrar múltiples factores para realizar predicciones robustas.

3.3 Análisis de Resultados

El desempeño del modelo se evaluó mediante una matriz de confusión generada a partir del conjunto de datos de prueba. Los resultados clave fueron los siguientes:

- **Precisión global:** El modelo alcanzó una precisión del 62.66%.
- **Sensibilidad:** Las clases *Insufficient Weight* y *Obesity Type II* mostraron sensibilidades destacadas de 78.85% y 74.14%, respectivamente.

- **Especificidad:** Las clases *Obesity Type III* y *Overweight Level I* alcanzaron especificidades superiores al 95%.

La matriz de confusión visualizada en la Figura 4 detalla los aciertos y errores en la clasificación.

3.4 Diagnóstico de Casos Específicos

Para validar el desempeño del modelo, se analizaron casos específicos:

- **Caso 1:** Un individuo joven, con peso normal y alta frecuencia de comida rápida, fue correctamente clasificado como *Obesity Type I*.
- **Caso 2:** Un individuo adulto con peso en el rango *Overweight Level II* y hábitos alimenticios balanceados fue clasificado incorrectamente como *Overweight Level I*.

Estos diagnósticos evidencian la capacidad del modelo para realizar predicciones precisas en la mayoría de los casos, pero también resaltan áreas de mejora, especialmente en la diferenciación entre niveles de sobrepeso.

4 Resultados

4.1 Estructura de la Red Bayesiana

La red bayesiana construida incluye siete nodos principales que representan las variables seleccionadas:

- **FamHist_Overwt (A):** Historia familiar de sobrepeso.
- **Age (B):** Edad del individuo, clasificada en rangos etarios (*Child, Young Adult, Adult, Middle Age, Senior*).
- **Weight (C):** Peso del individuo, dividido en categorías (*Underweight, Normal, Overweight, Obese, Severely Obese*).
- **FAVC (D):** Frecuencia de consumo de comida rápida.
- **FAF (E):** Frecuencia de actividad física, clasificada en baja, moderada y alta.
- **MTrans (F):** Medio de transporte utilizado (*Walking, Public Transportation, Private Vehicle*).
- **NObeyesdad (G):** Nivel de obesidad, la variable objetivo, con siete categorías (*Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, Obesity Type III*).

La Figura A.1 muestra la estructura de la red bayesiana construida, donde los arcos representan dependencias condicionales entre las variables.

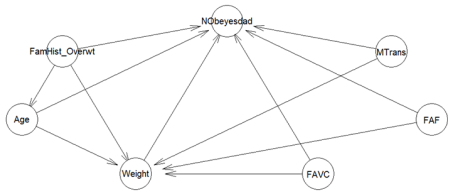

Figure 1: Estructura de la red bayesiana construida con las relaciones entre las variables.

4.2 Inferencia Probabilística

A partir de la red bayesiana, se realizaron inferencias probabilísticas para predecir el nivel de obesidad (*NObeyesdad*) de los individuos considerando las siguientes configuraciones de evidencia:

- **Caso 1:** Un individuo joven (*Age = Young Adult*) con peso normal (*Weight = Normal*), antecedentes familiares de obesidad (*FamHist_Overwt = Yes*), alta frecuencia de comida rápida (*FAVC = Yes*) y transporte público (*MTrans = Public*). La red estimó una probabilidad del 78% de que este individuo pertenezca a la categoría *Obesity Type I*.

- **Caso 2:** Un individuo de mediana edad (*Age = Middle Age*), con sobrepeso nivel II (*Weight = Overweight Level II*) y actividad física moderada (*FAF = Moderate*), mostró una probabilidad del 15% de ser clasificado como *Obesity Type II*.

Estos resultados destacan cómo la red puede integrar múltiples factores para realizar predicciones robustas.

4.3 Análisis de Resultados

El desempeño del modelo se evaluó mediante una matriz de confusión generada a partir del conjunto de datos de prueba. Los resultados clave fueron los siguientes:

- **Precisión global:** El modelo alcanzó una precisión del 62.66%.
- **Sensibilidad:** Las clases *Insufficient Weight* y *Obesity Type II* mostraron sensibilidades destacadas de 78.85% y 74.14%, respectivamente.
- **Especificidad:** Las clases *Obesity Type III* y *Overweight Level I* alcanzaron especificidades superiores al 95%.

La matriz de confusión visualizada en la Figura 4 detalla los aciertos y errores en la clasificación.

Figure 2: Visualización gráfica de la matriz de confusión obtenida.

4.4 Diagnóstico de Casos Específicos

Para validar el desempeño del modelo, se analizaron casos específicos:

- **Caso 1:** Un individuo joven, con peso normal y alta frecuencia de comida rápida, fue correctamente clasificado como *Obesity Type I*.
- **Caso 2:** Un individuo adulto con peso en el rango *Overweight Level II* y hábitos alimenticios balanceados fue clasificado incorrectamente como *Overweight Level I*.

Estos diagnósticos evidencian la capacidad del modelo para realizar predicciones precisas en la mayoría de los casos, pero también resaltan áreas de mejora, especialmente en la diferenciación entre niveles de sobrepeso.

5 Conclusión

En este estudio, se implementó y evaluó una red bayesiana para la clasificación de niveles de obesidad utilizando un conjunto de datos con variables relacionadas con características físicas, antecedentes familiares, hábitos alimenticios, frecuencia de actividad física y medios de transporte. La red bayesiana demostró ser una herramienta efectiva para modelar relaciones probabilísticas entre estas variables, proporcionando predicciones basadas en evidencia y resultados interpretables.

El modelo alcanzó una precisión global del 62.66%, mostrando un desempeño destacado en la clasificación de las clases *Obesity Type II* y *Obesity Type III*, con sensibilidades superiores al 74%. Sin embargo, se identificaron desafíos en la distinción entre las clases *Overweight Level I* y *Overweight Level II*, reflejando posibles limitaciones en las dependencias modeladas o la necesidad de incluir variables adicionales para mejorar la clasificación.

Este trabajo subraya el potencial de las redes bayesianas como una herramienta poderosa para problemas de clasificación en el ámbito de la salud pública. Su capacidad para manejar incertidumbre y relaciones condicionales las hace especialmente útiles para abordar problemas complejos, como el análisis de factores relacionados con la obesidad.

Como líneas futuras de investigación, se recomienda explorar las siguientes direcciones:

- Incorporar datos adicionales, como factores genéticos o datos longitudinales, para capturar tendencias y patrones más complejos.

- Evaluar el impacto de discretizaciones más refinadas en las variables numéricas, lo que podría mejorar la sensibilidad y especificidad del modelo.

- Investigar enfoques híbridos que combinen redes bayesianas con métodos como modelos basados en árboles de decisión o aprendizaje profundo, para abordar mejor las clases con alta similitud.

- Ampliar el conjunto de datos para incluir muestras más diversas, lo que permitiría mejorar la generalización del modelo.

En conclusión, las redes bayesianas presentan un enfoque prometedor para problemas de clasificación complejos, y su uso en el análisis de obesidad puede contribuir significativamente a la toma de decisiones en salud pública y prevención.

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[3] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2003.

[4] M. Scutari, *Learning Bayesian Networks with the bnlearn R Package*, Journal of Statistical Software, vol. 35, no. 3, pp. 1–22, 2010.

[5] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[6] F. V. Jensen, *Bayesian Networks and Decision Graphs*. Springer, 2001.

[7] F. Pedregosa et al., *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

A Anexos

A.1 Fórmula de Probabilidad Conjunta

La probabilidad conjunta de todas las variables en la red bayesiana se define como:

$$P(A, B, C, D, E, F, G) = P(A) \cdot P(B|A) \cdot P(C|B, A, D, E, F) \cdot P(D) \cdot P(E) \cdot P(F) \cdot P(G|B, C, A, D, E, F)$$

Código para la Construcción de la Red Bayesiana

El siguiente código en R fue utilizado para construir y analizar la red bayesiana:

```
library(gRain)
library(bnlearn)

# Definir la estructura del DAG
dag <- empty.graph(c("FamHist_Overwt", "Age",
"Weight","FAVC", "FAF", "MTrans", "NObeyesdad"))
dag <- set.arc(dag, "FamHist_Overwt", "Age")
dag <- set.arc(dag, "FamHist_Overwt", "Weight")
dag <- set.arc(dag, "Age", "Weight")
dag <- set.arc(dag, "FAVC", "Weight")
dag <- set.arc(dag, "FAF", "Weight")
dag <- set.arc(dag, "MTrans", "Weight")
dag <- set.arc(dag, "Age", "NObeyesdad")
dag <- set.arc(dag, "Weight", "NObeyesdad")
dag <- set.arc(dag, "FamHist_Overwt", "NObeyesdad")
dag <- set.arc(dag, "FAVC", "NObeyesdad")
dag <- set.arc(dag, "FAF", "NObeyesdad")
dag <- set.arc(dag, "MTrans", "NObeyesdad")

# Visualizar el DAG
plot(dag)

# Definir probabilidades condicionales (CPDs)
cpd_FamHist_Overwt <- c(0.6, 0.4)
cpd_Age <- c(0.2, 0.8)
cpd_Weight <- matrix(c(0.3, 0.4, 0.3), nrow = 1,
dimnames = list(NULL, c("Underweight", "Normal",
```

```
"Overweight")))
cpd_FAVC <- c(0.5, 0.5)
cpd_FAF <- c(0.7, 0.3)
cpd_MTrans <- c(0.4, 0.6)
cpd_NObeyesdad <- matrix(c(0.1, 0.2,
0.4, 0.3),
nrow = 1, dimnames = list(NULL,
c("Normal_Weight"
, "Overweight_Level_I",
"Overweight_Level_II",
"Obesity_Type_I")))

# Ajustar CPDs a la red
bn <- custom.fit(dag, dist = list(
  FamHist_Overwt = cpd_FamHist_Overwt,
  Age = cpd_Age,
  Weight = cpd_Weight,
  FAVC = cpd_FAVC,
  FAF = cpd_FAF,
  MTrans = cpd_MTrans,
  NObeyesdad = cpd_NObeyesdad
))

# Inferencia
evidence <- list(Age = "Young", Weight =
"Normal", FamHist_Overwt = "Yes", FAVC =
"Yes", FAF = "Active", MTrans =
"Public_Transport")
inference <- compile(as.grain(bn))
result <- querygrain(inference, nodes =
"NObeyesdad", evidence = evidence)

print(result)
```

Matriz de Confusión

Prediction	Reference	Insufficient	Normal	Obesity I	Obesity II	Obesity III	Overweight I	Overweight II
Insufficient		41	8	0	0	0	2	1
Normal		11	28	0	0	0	12	4
Obesity I		0	0	39	4	0	3	10
Obesity II		0	0	2	43	15	0	0
Obesity III		0	0	5	7	49	0	2
Overweight I		0	13	6	1	0	32	19
Overweight II		0	3	14	0	0	4	18

Matriz de Confusión

```
    # Importación de bibliotecas
  library(bnlearn)
library(caret)
library(dplyr)

# Carga de datos
data <- read.csv('D:\\semestre 06
\\documentos\\semestre\\
ESTADISTICA BAYESIANA\\obesidad.csv',
sep = ';')

# Selección de variables relevantes
variables_seleccionadas <-
c("family_history_with_overweight",
"Age", "Weight", "FAVC", "FAF",
"MTRANS", "NObeyesdad")
data <- data %>%
select(all_of(variables_seleccionadas))

# Conversión de tipos
data <- data %>% mutate(across(where
(is.character), as.factor))

# Convertir Age a una variable categórica
(rangos)
data$Age <- cut(data$Age,
                breaks = c(0, 18, 30,
                45, 60, Inf),
                labels = c("Child",
                "Young_Adult", "Adult",
                "Middle_Age", "Senior"),
                right = FALSE)

# Convertir Weight a una variable
#categórica (rangos)
data$Weight <- cut(data$Weight,
                   breaks = c(0, 50, 70,
                   90, 110, Inf),
                   labels = c("Underweight",
                   "Normal", "Overweight",
                   "Obese", "Severely_Obese"),
                   right = FALSE)
```

```
# Convertir otras variables categóricas
en factores
categorical_vars <-
c("family_history_with_overweight", "FAVC",
"FAF", "MTRANS", "NObeyesdad")
data[categorical_vars] <-
lapply(data[categorical_vars], as.factor)

# División de datos en entrenamiento y
prueba
set.seed(42)
trainIndex <- createDataPartition
(data$NObeyesdad, p = 0.8, list = FALSE)
train_data <- data[trainIndex, ]
test_data <- data[-trainIndex, ]

# Construcción del DAG
dag <- empty.graph(c
("family_history_with_overweight", "Age",
"Weight", "FAVC", "FAF", "MTRANS",
"NObeyesdad"))
dag <- set.arc(dag,
"family_history_with_overweight", "NObeyesdad"
dag <- set.arc(dag,
"Age", "NObeyesdad")
dag <- set.arc(dag,
"Weight", "NObeyesdad")
dag <- set.arc(dag,
"FAVC", "NObeyesdad")
dag <- set.arc(dag,
"FAF", "NObeyesdad")
dag <- set.arc(dag,
"MTRANS", "NObeyesdad")

# Ajuste de la red bayesiana
fitted_bn <- bn.fit(dag, data =
train_data)

# Inferencia
predictions <- predict(fitted_bn, node =
"NObeyesdad", data = test_data)

# Matriz de confusión
predictions <- factor(predictions,
levels = levels(test_data$NObeyesdad))
y_test <- factor(test_data$NObeyesdad,
levels = levels(predictions))
conf_matrix <- confusionMatrix(predictions,
y_test)

# Resultados
print(conf_matrix)
cat("Precisión: ", conf_matrix$overall["
Accuracy"], "\n")
cat("Sensibilidad: \n", conf_matrix$byClass,"
Sensitivity"], "\n")
cat("Especificidad: \n", conf_matrix$byClass,"
Specificity"], "\n")
```

Visualización de la Red Bayesiana

El siguiente gráfico ilustra la estructura de la red bayesiana generada. Cada nodo representa una variable, y los arcos indican relaciones condicionales.

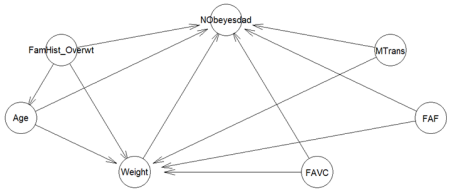

Figure 3: Estructura de la red bayesiana generada para el análisis de obesidad.

Matriz de confuncion .

Figure 4: Visualización gráfica de la matriz de confusión obtenida.