Discriminante Gaussiano

Ian Gabriel Cañas Fernández. — 1092228
Instituto Tecnológico de Santo Domingo (INTEC)
Santo Domingo, D.N. República Dominicana
Machine Learning — INL367. Sección 01

***Resumen*—En este documento se presenta el proceso de identificación de personas en base a sus características para determinar si sufren o no de una enfermedad cardiovascular, basándose en las características dadas. Se presentó la estandarización de los datos brindados para la obtención de patrones y se asumió varias de las características con una distribución normal tanto individualmente como en conjunto.**

Palabras clave – Bayes, Bayes ingenuo, distribución gaussiana, salud.

***Abstract*—This document presents the process of identifying people based on their characteristics to determine whether or not they suffer from cardiovascular disease, based on the given characteristics. The standardization of the data provided to obtain patterns was presented and several of the characteristics were assumed to have a normal distribution both individually and as a whole.**

***Keywords* – Bayes, naive Bayes, Gaussian distribution, health.**

I. Introducción

El teorema de Bayes es un método de aprendizaje supervisado que se basa en la evaluación de relación entre probabilidades condicionales de eventos aleatorios, que pertenecen a un estudio. En este consideramos a los parámetros como variables aleatorias con una distribución, permitiéndonos modelar nuestra incertidumbre y a partir de ello estimarla [1].

En la presente se implementará el teorema de Bayes multivariables donde se obtendrá un set de datos que contienen doce características obtenidas en un examen médico hecho a 70 000 pacientes, dataset brindada por Svetlana Ulianova [2].

A. Objetivo general

- Clasificar las personas en un estudio sobre si sufren o no una enfermedad cardiovascular en base al conocimiento de su peso, altura y edad mediante el uso de distribución normal.

B. Objetivos específicos

- Reconocer los métodos de clasificación gaussiana.
- Destacar las ventajas y desventajas de cada método de clasificación gaussiana.
- Implementar el teorema de Bayes para machine learning.

II. Método

Para todos los casos se segmentará el conjunto de datos brindados a un grupo de 70 % que sería utilizado para el entrenamiento y la creación de los modelos; el resto de los datos sería utilizado para la comprobación de los modelos y la exactitud de estos. Para evitar posible sesgo de muestreo se inició aleatorizando los datos brindados.

A. Aplicación de Bayes univariable con características continuas

El clasificador de bayes univariable procura la distinción o clasificación de las personas en base a una única característica. Para ello se distinguió las posibilidades en base al teorema de Bayes (Ecuación 1), que describe la probabilidad de que se tenga una clase (en este caso, de estar sano o enfermo) en base a que se conoce una característica del paciente.

$$P(C_i|x) = \frac{P(C_i)P(x|C_i)}{P(x)}$$

Ecuación 1. Teorema de Bayes.

Este teorema ha sido utilizado con su equivalente que describe la probabilidad de tener dicha característica en base a que se conoce la clase mediante el uso de una función de densidad probabilística [3]. Además, se ha extrapolado a trabajar x como un vector X con todos los casos a ser descritos.

$$P(C_i|X) = \frac{P(C_i)p(X|C_i)}{p(X)}$$

Ecuación 2. Regla de Bayes.

La clase de enfermos será descrita mediante el valor 1 y la clase de sanos será descrita mediante un cero. Se empieza estableciendo los priors:

$$\hat{P}(C_i) = \frac{\sum_{t=1}^{n} x^t}{N}$$

Ecuación 3. Probabilidad a priori de cada clase.

Donde i indica la clase y t describe la cantidad de datos que se tienen, siendo x cada dato individual y N la cantidad total de datos. A partir de tal expresión obtenemos la probabilidad a priori de obtener una clase u otra.

A continuación, se seleccionan las características que se utilizarán en el modelo. En el primer segmento estaremos prediciendo con el uso de la edad, altura y peso individualmente.

En un sentido general se modela el likelihood de cada característica mediante la siguiente expresión que devuelve una constante para cada valor de x. Para poder modelar esta expresión se asume que la distribución de edades, pesos y edades cada una lleva una distribución normal, tanto en un sentido general como dada la condición de salud de cada paciente.

La distribución normal, también llamada distribución de Gauss, se describe mediante las siguientes expresiones:

$$p(x|C_1) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

Ecuación 4. Likelihood de una característica dada una clase.

$$\mu_i = \frac{\sum_{t=1}^{n} x^t}{N}$$

Ecuación 5. Media de un conjunto de datos.

$$\sigma_i = \sqrt{\frac{\sum_{t=1}^{n} (x_i^t - \mu_i)^2}{N}}$$

Ecuación 6. Desviación estándar de un conjunto de datos.

Donde σ_i describe la desviación estándar del conjunto de datos de entrenamiento de la clase i y μ_i describe la media del muestreo de dicha clase. Paralelamente se obtiene la misma expresión generalizada a la población, obteniéndose el modelo de la evidencia de los datos.

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Ecuación 7. Evidencia de una característica.

Por cuestión de simplicidad, para la toma de decisiones, al aplicar logaritmo natural (que expresaremos como log en vez de ln), se trabajará con la siguiente expresión discriminante para la toma de decisiones:

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log \hat{P}(C_i)$$

Ecuación 8. Función discriminante univariable.

Para la toma de decisión de prosigue introduciendo la característica en tal función discriminante para ambas clases, y se seleccionará la característica con mayor valor de salida.

Los resultados de este método se presentan en la sección III. A. en la que se presentan las matrices de confusión respecto a las decisiones tomadas considerando los modelos de densidad probabilística estimados desde los datos de prueba.

B. Bayes multivariable

Conociendo el procedimiento de clasificación por Bayes univariable, se asume que existe la posibilidad de que un único valor podría no ser tan efectivo como varios valores para la identificación de las enfermedades cardiovasculares.

1) Peso y altura como características

Se prosiguió considerando una interrelación entre el peso y la altura que nos pueda describir si se sufre o no de una enfermedad cardiovascular, para mantener la rigurosidad, se asumen ambas variables como dependientes entre sí y se extrapoló la distribución normal para que describa una densidad probabilística dado un peso y una altura determinada.

Antes de la extrapolación de la distribución de define el vector media y la matriz de covarianza mediante las siguientes expresiones.

$$E[X] = \mu = [\mu_1, \dots, \mu_d]^T$$

Ecuación 9. Vector media.

$$\Sigma \equiv cov(X) = E[(X-\mu)(X-\mu)^T] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

Ecuación 10. Definición de matriz de covarianza.

Si se considera la matriz de covarianza como no lineal, se trabajaría con la siguiente extrapolación de la campana gaussiana a más de una variable:

$$p(x|C_1) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-\mu_i)^T\Sigma_i^{-1}(X-\mu_i)}$$

Ecuación 11. Distribución normal multivariable.

Se genera el modelo en base a dicha distribución normal y se vuelve a aplicar el teorema de Bayes, solo que para esta ocasión se trabajó con su versión simplificada:

$$g_i(x) = x^T W_i x + w_i^T x + w_{i0}$$

Ecuación 12. Función discriminante Bayes multivariable.

tal que,

$$W_i = -\frac{1}{2}S_i^{-1}$$
$$w_i = S_i^{-1} m_i$$
$$w_{i0} = -\frac{1}{2} m_i^T S_i^{-1} m_i - \frac{1}{2}\log|S_i| + \log \hat{P}(C_i)$$

Donde las m_i y S_i representan las estimaciones del vector media y de la matriz de covarianza para cada clase. Los valores resultantes para las matrices de covarianzas, los vectores de media y priors se presentan en la Tabla 1.

Tabla 1. Covarianzas, medias y priors.

Expresión	Valor
$\widehat{\boldsymbol{P}}(\boldsymbol{C_0})$	0.5
$\widehat{\boldsymbol{P}}(\boldsymbol{C_1})$	0.5
$\boldsymbol{\mu_{sano}}$	71.6405 164.4515
$\boldsymbol{\mu_{enfermo}}$	76.8273 164.2330
$\boldsymbol{\Sigma_{EPA,sanos}}$	178.8008 34.7626 34.7626 66.5160
$\boldsymbol{\Sigma_{EPA,enfermos}}$	225.5469 34.4144 34.4144 69.2336

Los resultados de este método se presentan en la sección III. B. en la que se presenta la estimación respecto a las decisiones tomadas considerando el modelo de densidad probabilística bivariable estimado desde los datos de prueba.

2) Peso, altura y edad como características independientes (Naive Bayes)

El clasificador de Naive Bayes consiste en la consideración de dos o más características utilizadas para el análisis de probabilidad de una clase son independientes entre sí y que el valor resultante se relaciona a la distancia de Mahalanobis hacia dichas características, que depende tan solo de la varianza de

cada una de estas.

La función de densidad probabilística utilizada para Naive Bayes se deriva de la Ecuación 11 y se presenta en la Ecuación 13, más adelante, en la Ecuación 14 se presenta la simplificación, que funciona como función discriminante utilizada para la clasificación de los datos.

$$p(x) = \frac{1}{\sqrt{2\pi}\prod_{i=1}^{d}\sigma_i} e^{-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i^2}\right)^2}$$

Ecuación 13. Densidad probabilística para entradas independientes; Naive Bayes.

$$g_i(x) = -\frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j^t - m_{ij}}{s_j}\right)^2 + \log \hat{P}(C_i)$$

Ecuación 14. Discriminante gaussiano para entradas independientes.

3) Peso, altura y edad como características correlacionadas

Finalmente se prueba con la consideración de la edad, el peso y la altura como variables correlacionadas entre sí, por lo que se asume las Ecuación 11 y Ecuación 12. Las funciones discriminantes se utilizan para predecir la clase a la que pertenecerá cada caso, esto se hace por medio de la comparación de los resultados arrojados por la función discriminante evaluada en cada conjunto de características y seleccionando el valor más grande.

III. Resultados

Antes de la distinción de los métodos utilizados, los datos obtenidos para el prior de cada clase, el modelo de los likelihoods, las medias y variaciones estándar de cada variable se listan en la Tabla 2.

Tabla 2. Estimaciones probabilísticas.

Expresión	Valor
$\hat{\boldsymbol{P}}(\boldsymbol{C_0})$	0.5
$\hat{\boldsymbol{P}}(\boldsymbol{C_1})$	0.5
$\boldsymbol{\mu_{sano}}$	18 884 164 72
$\boldsymbol{\mu_{enfermo}}$	20 064 164 77
$\boldsymbol{\Sigma_{EPA,sanos}}$	6.1268 −0.0010 0.0022 −0.0010 0.0001 0.0000 0.0022 0.0000 0.0002
$\boldsymbol{\Sigma_{EPA,enfermos}}$	5.3444 −0.0021 −0.0016 −0.0021 0.0001 0.0000 −0.0016 0.0000 0.0002
$\boldsymbol{\Sigma_{Naive,sanos}}$	6 126 800 0.0000 0.0000 0.0000 100.0000 0.0000 0.0000 0.0000 200.0000
$\boldsymbol{\Sigma_{Naive,enfermos}}$	5 344 400 0.0000 0.0000 0.0000 100.0000 0.0000 0.0000 0.0000 200.0000

A. *Bayes univariable*

Implementando Bayes mediante el uso de Matlab, en base a las expresiones descritas en la sección II. A. se puede observar en la Ilustración 1 el código implementado en un sentido general y en la Ilustración 2 la distribución normal que modela la característica de la edad de personas enfermas. Véase la Ilustración 3 para su paralelo en la clase de las personas sanas

Igualmente, la gráfica de histograma frente a modelo estimado tanto para la altura (Ilustración 4 e Ilustración 5) como el peso (Ilustración 6 e Ilustración 7) se presentan a continuación.

```
%% Priors, características y likelihoods

% Para conocer el prior de los enfermos en el dataset, se extrae su
% proporción en base a los datos de prueba.

prior = [sum(traindata.cardio==0) sum(traindata.cardio==1)]./length(traindata.cardio);
PC1 = prior(2);

sanos = traindata(traindata.cardio==0, :);
enfermos = traindata(traindata.cardio==1, :);

likeSano = [mean(sanos.age) std(sanos.age); mean(sanos.height) std(sanos.height); mean(sanos.weight) std(sanos.weight)];
likeEnfermo = [mean(enfermos.age) std(enfermos.age); mean(enfermos.height) std(enfermos.height); mean(enfermos.weight) std(enfermos.weight)];

%% Gráfico de la edad de las personas enfermas versus su estimación
figure
edadEnfermoEstimado = normrnd(likeEnfermo(1, 1), likeEnfermo(1,2), [length(enfermos.id) 1]);
histogram(enfermos.age)
hold on
histogram(edadEnfermoEstimado)

title("Edad de las personas enfermas")
legend ("Datos reales", "Modelo estimado")
xlabel("Edad (días)")
ylabel("Cantidad de enfermos")

clear edadEnfermoEstimado
```

Ilustración 1. Código implementado para generación de modelo e histograma.

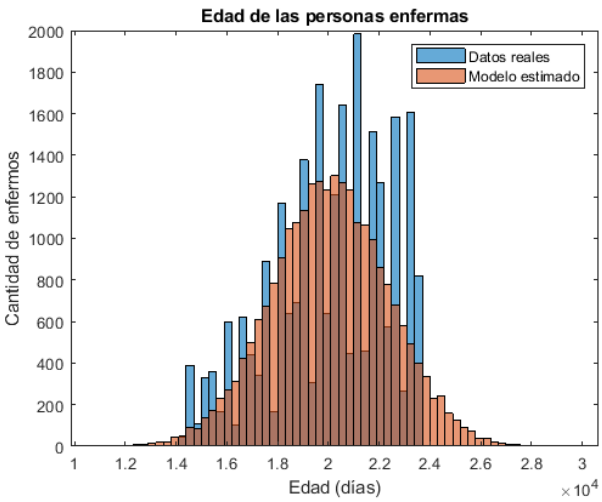

Ilustración 2. Histograma de edades de personas enfermas frente a modelo generado.

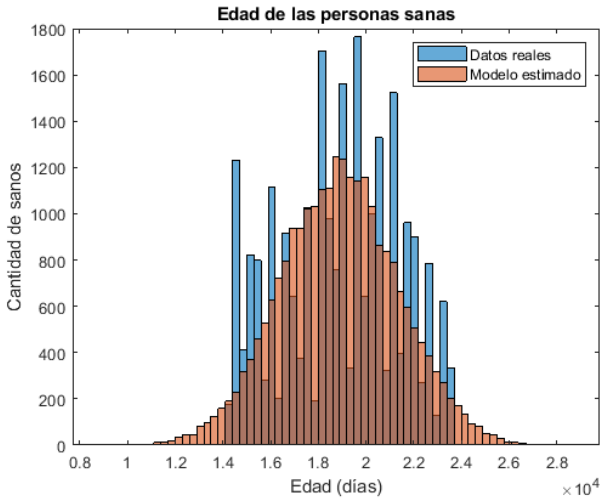

Ilustración 3. Histograma de edades de personas sanas frente a modelo generado.

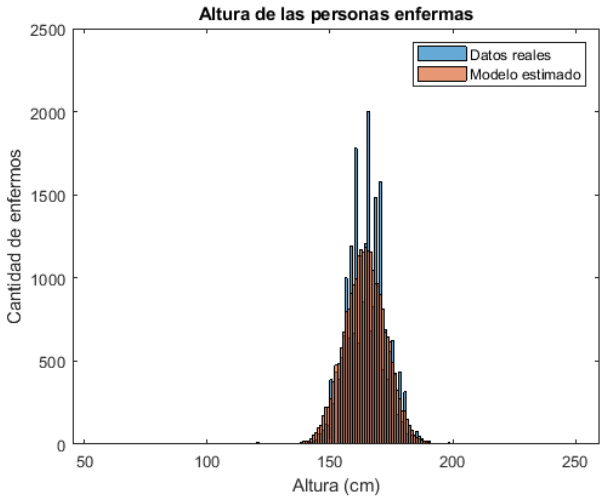

Ilustración 4.. Histograma de alturas de personas enfermas frente a modelo generado.

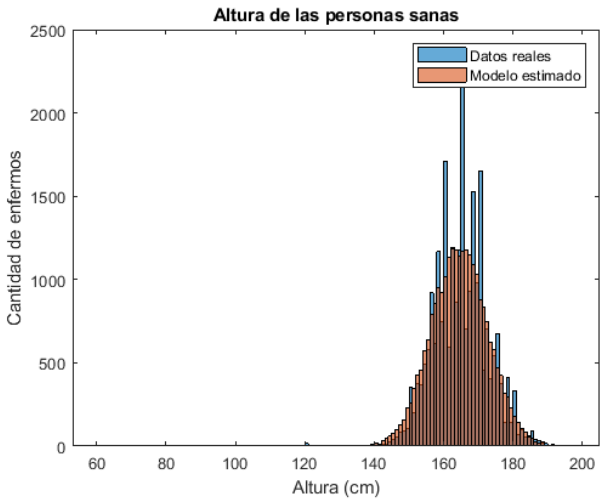

Ilustración 5. Histograma de alturas de personas sanas frente a modelo generado.

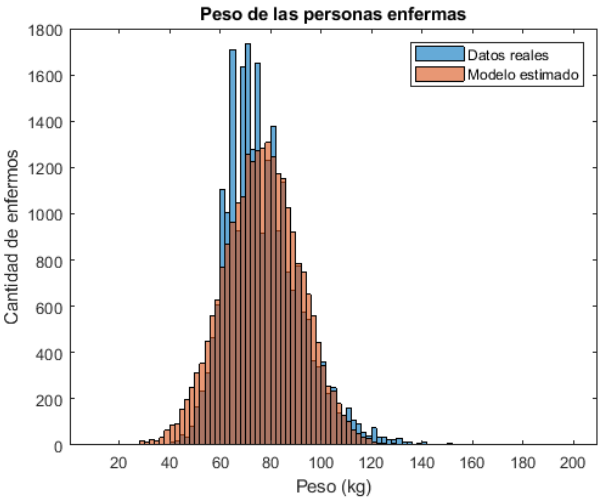

Ilustración 6. Histograma de pesos de personas enfermas frente a modelo generado.

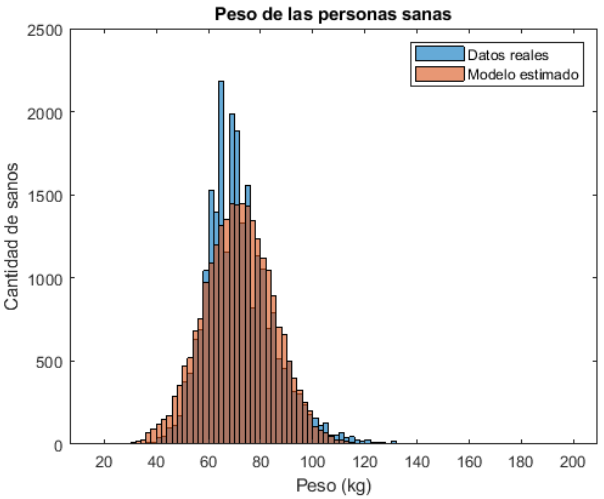

Ilustración 7. Histograma de pesos de personas sanas frente a modelo generado.

Retomando el hecho que a partir de dicho modelo se hace la estimación mediante el uso de las Ecuación 2 y Ecuación 8 como expresiones de predicción tal y como se presentan en la Ilustración 8 y Ilustración 9.

```
%% 1.4 Evidencia de cada característica

pEdad = [mean(traindata.age) std(traindata.age)]; %./length(traindata.age);
pAltura = [mean(traindata.height) std(traindata.height)]; %./length(traindata.height);
pPeso = [mean(traindata.weight) std(traindata.weight)]; %./length(traindata.weight);

pdfEdad = pdf('Normal', testdata.age, pEdad(1), pEdad(2)); % p(x) / x = Edad
pdfAltura = pdf('Normal', testdata.height, pAltura(1), pAltura(2));  % p(x) / x = altura
pdfPeso = pdf('Normal', testdata.weight, pPeso(1), pPeso(2));  % p(x) / x = Peso

meanEdadSano = mean(sanos.age);
meanAlturaSano = mean(sanos.height);
meanPesoSano = mean(sanos.weight);

stdEdadSano = std(sanos.age);
stdAlturaSano = std(sanos.height);
stdPesoSano = std(sanos.weight);


pdfEdadSano = pdf('Normal', testdata.age, meanEdadSano, stdEdadSano); % p(x) / x = Edad
pdfAlturaSano = pdf('Normal', testdata.height, meanAlturaSano, stdAlturaSano);  % p(x) / x = altura
pdfPesoSano = pdf('Normal', testdata.weight, meanPesoSano, stdPesoSano);  % p(x) / x = Peso
```

Ilustración 8. Uso directo de teorema de Bayes.

```
%% Clasificación de datos utilizando la edad como variable independiente
    for j = 1:length(testdata.id); % Repeticiones hasta la cantidad de personas en el test
        % Funciones discriminantes
        gSanos = -0.5*log(2*pi)-log(likeSano(1, 2))-(testdata.age(j)-likeSano(1,1))^2/(2*likeSano(1, 2))+log(prior(1));
        gEnfermos = -0.5*log(2*pi)-log(likeEnfermo(1, 2))-(testdata.age(j)-likeEnfermo(1,1))^2/(2*likeEnfermo(1, 2))+log(prior(1));

        if gSanos > gEnfermos
            estado = 0; % Persona sana
        else
            estado = 1; % Persona no sana
        end

        testdata.predictionEdad(j) = estado;

        if estado == testdata.cardio(j)
            cont = cont + 1;
            if estado == 0
                tp(1) = tp(1) + 1;
            else
                tn(1) = tn(1) + 1;
            end
        else
            if estado == 0
                fp(1) = fp(1) + 1;
            else
                fn(1) = fn(1) + 1;
            end

        end
    end
```

Ilustración 9. Uso de función discriminante.

A continuación, se presentan los resultados de exactitud obtenidos presentado mediante proporción de aciertos (Ilustración 10), que fue medida llevando un conteo de la cantidad de verdaderos positivos y verdaderos negativos sobre la cantidad total de predicciones; así como se presentan las matrices de confusión (Ilustración 11, Ilustración 12 y Ilustración 13).

```
aciertosEdad =

    0.5930


aciertosAltura =

    0.5080


aciertosPeso =

    0.5691
```

Ilustración 10. Proporción de aciertos.

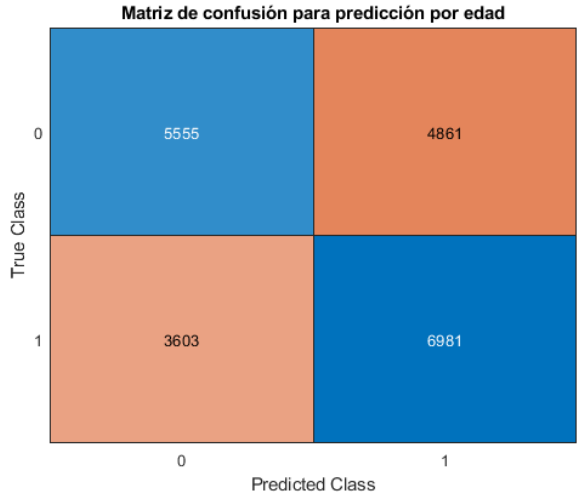

Ilustración 11. Matriz de confusión para predicción por edad.

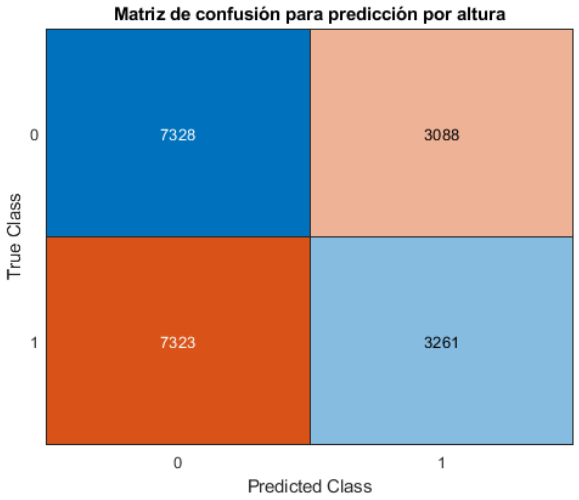

Ilustración 12. Matriz de confusión para predicción por altura.

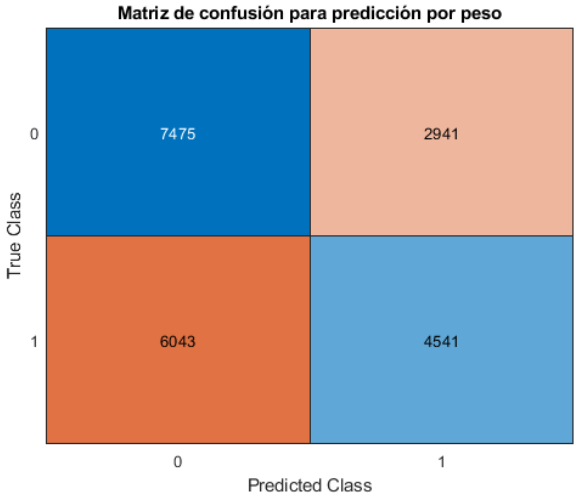

Ilustración 13. Matriz de confusión para predicción por peso.

B. Bayes multivariable: peso y altura

Retomando los valores obtenidos en la Tabla 2 y reconociendo el uso de las Ecuación 11 y Ecuación 12 se llega a determinar qué personas tienen o no una enfermedad cardiovascular en base a la relación entre tales características. En la Ilustración 14 se presenta el algoritmo implementado en Matlab y en la Ilustración 15 se presenta la tasa de aciertos.

En este caso se presenta una distribución gaussiana de dos características (Ilustración 16) generada con los datos del entrenamiento, apreciando la posibilidad de que ambas características se encuentren correlacionadas entre sí. En la gráfica de la ya mencionada distribución bivariable se puede apreciar en el color azulado la densidad probabilística para las personas sanas y en la superficie rojiza la distribución de las personas que sufren enfermedades cardiovasculares.

```
%%
predictionMultVar = nan(length(testdata.id), 1);
predictionNaiveBayes = nan(length(testdata.id), 1);
predictionDep = nan(length(testdata.id), 1);

testdata = [testdata table(predictionNaiveBayes) table(predictionMultVar) table(predictionDep)];

%%

wSano = -0.5*inv(covSano);
wEnfermo = -0.5*inv(covEnfermo);

wSano = inv(covSano)*meanSano;
wEnfermo = inv(covEnfermo)*meanEnfermo;

w0Sano = -0.5*meanSano'*inv(covSano)*meanSano - 0.5*log(det(covSano)) + log(prior(1));
w0Enfermo = -0.5*meanEnfermo'*inv(covEnfermo)*meanEnfermo-0.5*log(det(covEnfermo))+log(prior(2));

%%
cont = 0;
for i = 1:length(testdata.id)
    sano = PesAltTest(i,:)*wSano*PesAltTest(i,:)' + wSano'*PesAltTest(i,:)'+w0Sano;
    enfermo = PesAltTest(i,:)*wEnfermo*PesAltTest(i,:)' + wEnfermo'*PesAltTest(i,:)'+w0Enfermo;

    if sano > enfermo
        estado = 0; % estado de la persona es no enferma
    else
        estado = 1; % estado de la persona es enferma
    end

    testdata.predictionMultVar(i) = estado;

    % Verificación si la persona está sana o no

    if estado == testdata.cardio(i)
        cont = cont + 1;
    end

end

aciertosMultVar = cont/length(testdata.id)
```

Ilustración 14. Algoritmo implementado para Bayes multivariable (peso-altura).

```
aciertosMultVar =

    0.5763
```

Ilustración 15. Tasa de aciertos para Bayes multivariable (peso-altura).

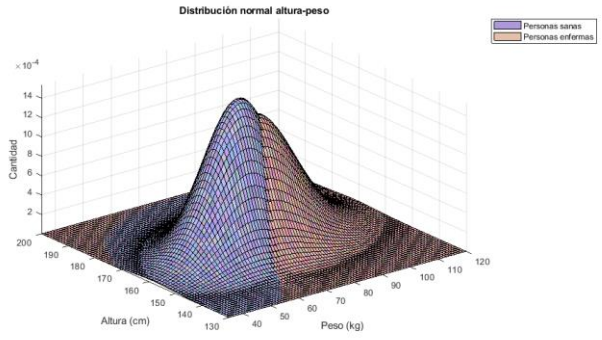

Ilustración 16. Distribución gaussiana bivariable para peso-altura.

C. *Naive Bayes*

Recordando los valores obtenidos en la Tabla 2; en esta sección ha sido replicado los procedimientos anteriores a tres variables, considerándolas como independientes entre sí mediante el uso de las expresiones Ecuación 13 y Ecuación 14. En la Ilustración 17 se muestra el algoritmo implementado en Matlab y en la Ilustración 18 se puede apreciar la tasa de acierto del modelo.

```
%% Naive Bayes asumiendo que las variables son independientes
carSano = [sanos.age(:) sanos.height(:) sanos.weight(:)]; % Matriz con peso y altura, no enfermos
carEnfermo = [enfermos.age(:) enfermos.height(:) enfermos.weight(:)]; % Matriz con peso y altura, enfermos

meanSano = mean(carSano);
meanEnfermo = mean(carEnfermo);

DesEstSano = std(carSano);
DesEstEnfermo = std(carEnfermo);

cont = 0;
for i = 1:length(testdata.id)
    X = [testdata.age(i) testdata.height(i) testdata.weight(i)];
    sano = -0.5*sum((((X-meanSano)./DesEstSano)').^2)+log(prior(1));
    enfermo = -0.5*sum((((X-meanEnfermo)./DesEstEnfermo)').^2)+log(prior(2));

    if sano > enfermo
        estado = 0;% Persona no está enferma
    else
        estado = 1;% Persona está enferma
    end

    testdata.predictionNaiveBayes(i) = estado;

    if estado == testdata.cardio(i)
        cont = cont + 1;
    end
end

aciertosNaiveIndep = cont/length(testdata.id)
```

Ilustración 17. Algoritmo implementado para Naive Bayes (edad, altura y peso).

```
aciertosNaiveIndep =

    0.6214
```

Ilustración 18. Tasa de aciertos para Naive Bayes (edad, altura y peso).

D. *Bayes multivariable: edad, altura y peso*

Retomando el empleo de Bayes a dos variables, donde se obtuvo una tasa de acierto cerca de 58 %, se ha procurado considerar, además, la edad como una característica importante.

Empleando el mismo algoritmo, en este caso se ha implementado tanto directamente mediante la Ecuación 1 (véase la Ilustración 19), así como mediante su simplificación vista en la Ecuación 12 con fines comparativos (véase la Ilustración 20).

Además, en la Ilustración 21 se puede apreciar la tasa de aciertos obtenida por ambos métodos.

```
%% Naive Bayes asumiendo que las variables son dependientes

meanSano = mean(carSano)';
meanEnfermo = mean(carEnfermo)';

covSano = cov(carSano);
covEnfermo = cov(carEnfermo);

RSano = corrcov(covSano);
REnfermo = corrcov(covEnfermo);

mean3 = mean([traindata.age traindata.height traindata.weight])';
cov3 = cov([traindata.age traindata.height traindata.weight]);


NaiveTest = [testdata.age(:) testdata.height(:) testdata.weight(:)];

pdf3Enfermo = mvnpdf(NaiveTest, meanEnfermo', covEnfermo);
pdf3 = mvnpdf(NaiveTest, mean3', cov3);

g3Enfermo = prior(2).*pdf3Enfermo./pdf3;

testdata.predictionDep(g3Enfermo>=0.5) = 1;
testdata.predictionDep(g3Enfermo<0.5) = 0;
```

Ilustración 19. Uso directo de teorema de Bayes a tres características.

```
wSano = -0.5*pinv(covSano);
wEnfermo = -0.5*pinv(covEnfermo);

wSano = pinv(covSano)*meanSano;
wEnfermo = pinv(covEnfermo)*meanEnfermo;

w0Sano = -0.5*meanSano'*pinv(covSano)*meanSano - 0.5*log(abs(det(covSano))) + log(prior(1));
w0Enfermo = -0.5*meanEnfermo'*pinv(covEnfermo)*meanEnfermo-0.5*log(abs(det(covEnfermo))) + log(prior
%%

cont = 0;
for i = 1:length(testdata.id)
    sano = NaiveTest(i,:)*wSano*NaiveTest(i,:)' + wSano'*NaiveTest(i,:)' + w0Sano;
    enfermo = NaiveTest(i,:)*wEnfermo*NaiveTest(i,:)' + wEnfermo'*NaiveTest(i,:)' + w0Enfermo;


    sano;
    enfermo;

    aiuda(i, [1 2]) = [sano enfermo];

    if sano > enfermo
        estado = 0;% Persona no está enferma
    else
        estado = 1;% Persona está enferma
    end

    if estado == testdata.cardio(i)
        cont = cont + 1;
    end
end
aciertosDep2 = cont/length(testdata.id)
```

Ilustración 20. Uso de función discriminante a tres características.

```
aciertosDep =

    0.6213


aciertosDep2 =

    0.6213
```

Ilustración 21. Tasa de aciertos para Naive Bayes (edad, altura y peso).

IV. Análisis

En el caso de Bayes univariable se ha visto cada una de las características por separado, donde se puede destacar el efecto de cada una de ellas a cada uno de los resultados obtenidos, resaltándose un efecto por parte de la edad y el peso, observándose un 59.7 y 57.2 porciento, respectivamente.

Debido a la simpleza del método, es conveniente el uso directo del teorema de Bayes sin la necesidad de la función discriminante, pues posee menos peso computacional y porque se está trabajando con tan solo dos clases, de donde la decisión se toma más fácil observando cuál clase tiene una probabilidad de más del 50 %.

Se obtuvo que el empleo directo del teorema de Bayes consume tan solo 0.021679 segundos para las tres características en total, mientras que la clasificación mediante la función discriminante consume 0.533113 segundos. En ambos casos se tuvo la misma cantidad de aciertos.

Sobre Bayes multivariable para dos características, así como con Bayes univariable, se obtuvo una predicción satisfactoria, en este caso de un 57.6 %, mientras se asume la correlación entre el peso y la altura. Este método presenta la desventaja de poseer un tiempo de ejecución para la estimación de 1.12 segundos aproximadamente, por lo que en mayores *datasets* esto podría presentar un inconveniente. Para la generación de los gráficos toma tan solo 0.632572 segundos.

Aunque parezca ser un buen partido, el uso de la combinación peso altura no se tiene gran ventaja frente al uso directo del peso, pues no presenta gran variación en cuanto la tasa de aciertos que justifique el aumento del tiempo de ejecución.

En el siguiente caso se introduce el clasificador de Naive Bayes, también conocido como Bayes ingenuo, en este caso se hace la asunción de que las características con las que se está trabajando (en este caso edad, peso y altura), son independientes entre sí.

Este algoritmo se destaca por ignorar las posibles dependencias, también llamadas correlaciones, entre las entradas y simplifica un problema multivariable a un grupo de problemas univariables.

Con el empleo de este método se observa una mejora importante en la eficiencia de la predicción, teniéndose una tasa de aciertos de 62.14 % en un tiempo de ejecución de 0.21 segundos.

Finalmente, se trabaja de nuevo con Bayes multivariable con tres características que se asumen correlacionadas entre sí, observándose cómo se destaca cada detalle que involucre la relación entre las características presentadas en los datos.

En este caso, mediante ambos métodos, se obtuvo un tiempo de ejecución de aproximadamente 0.2 segundos con una tasa de acierto de 62.13 %, lo que da a conocer que en muchos casos no es un sacrificio mayor el hecho de ignorar la correlación entre las variables a cambio de simplificar la implementación del clasificador.

Ergo se puede considerar que el método de Naive Bayes es un método simple y rápido de implementar en muchos casos donde se desconozca si las características están conectadas entre sí.

V. Referencias

[1] Ethem Alpaydin, *Introduction to Machine Learning*. 2014.

[2] S. Ulianova, «Cardiovascular Disease dataset», *www.kaggle.com*. https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.

[3] D. Stirzaker, *Elementary Probability*. Cambridge University Press, 1994.