UNIVERSIDAD DE EL SALVADOR
FACULTAD MULDISCIPLINARIA DE OCCIDENTE
DEPARTAMENTO DE MATEMÁTICA

LICENCIATURA EN ESTADÍSTICA

APLICACIÓN DE REGRESIÓN LOGÍSTICA MULTINOMIAL A RESULTADOS DEL ESTUDIO DE FRAMINGHAM

DOCENTE:

LICENCIADO. JAIME ISAAC PEÑA

PRESENTADO POR:

Darlyn Iveth Puentes Jimenez

Tabla de contenidos

1. Regresión Logística Binomial y Multinomial

La Regresión Logística Binomial es un método estadístico utilizado en análisis de datos para predecir la probabilidad de que un resultado pertenezca a una categoría específica o para estimar la probabilidad de un evento binario en función de una o más variables independientes.La salida de un modelo de regresión logística no es una predicción numérica en el sentido tradicional, sino una probabilidad. Generalmente, se establece un umbral (por ejemplo, 0.5) y se considera que si la probabilidad calculada es mayor que ese umbral, el evento es probable (1), y si es menor, el evento es improbable (0).

La Regresión Logística Multinomial es un método estadístico utilizado para modelar y predecir relaciones entre una variable dependiente categórica nominal (que tiene más de dos categorías) y una o más variables independientes, que pueden ser categóricas o continuas. El objetivo principal es entender cómo las variables independientes influyen en la probabilidad de pertenecer a una categoría específica de la variable dependiente. En términos más sencillos, la regresión logística multinomial se utiliza para predecir en cuál de varias categorías posibles caerá una observación en función de las variables explicativas. Las Probabilidades de cada alternativa en un modelo logit multinomial se expresan de la siguiente manera:

$$P_{ij} = P(Y_i = j) = \frac{e^{-(\beta_{1j}+\beta_{2j}X_{2i}+\ldots+\beta_{kj}X_{ki})}}{1+\sum_{g=1}^{J} e^{-(\beta_{1j}+\beta_{2j}X_{2i}+\ldots+\beta_{kj}X_{ki})}}, j = 1, 2, \ldots J$$

$$P_{i0} = P(Y_i = j) = \frac{1}{1+\sum_{g=1}^{J} e^{-(\beta_{1j}+\beta_{2j}X_{2i}+\ldots+\beta_{kj}X_{ki})}}$$

2. Aplicación de Regresión Logística Multinomial en R

2.1 Análisis Exploratorio Univariado en R

Antes de iniciar con el análisis de Regresión Logística conviene realizar un análisis descriptivo univariante para ver la relación entre la variable dependiente y las independientes por separado. Con aquellas variables métricas se realizarán una comparación de las medias de los grupos de la variable dependiente.

La base que se utilizará para realizar la aplicación de Regresión Logística Multinomial corresponde al famoso estudio de Framingham, también conocido como el Framingham Heart Study (FHS), el cual es uno de los estudios epidemiológicos más influyentes y largos en la historia de la investigación médica. Este estudio se lleva a cabo en la ciudad de Framingham, Massachusetts, Estados Unidos. El Framingham Heart Study se inició en 1948 y ha sido fundamental para mejorar la comprensión de las enfermedades cardiovasculares y sus factores de riesgo. El objetivo es analizar que factores de riesgo cardiovascular están relacionados con los diferentes resultados en el Estudio de Framingham. Las variables que se utilizarán en este caso están recogidas en el siguiente cuadro:

Tabla 1: Descripción de los datos.

Variable	Descripción	Codificación
Sexo	Sexo del Paciente	1=”Femenino”, 0= ”Masculino”
Colesterol	nivel de colesterol	Nivel continuo de Colesterol
Edad	Edad del Paciente	Años
IMC	índice de masa corporal	índice continuo
BPvar	Variabilidad sanguínea	Dato continuo
Frecuencia	Frecuencia cardíaca	Frecuencia continua
Glucosa	Nivel de glucosa	Glucosa Continua
Fuma	Si fuma o no	0=”No”, 1=”Si”
Resultado	Desenlace del paciente	0=”Sin evento”, 1=”Con hipertensión”, 2=”Fallecido”
Cigarrillos	Cigarrillos que fuma	Número de cigarrillos

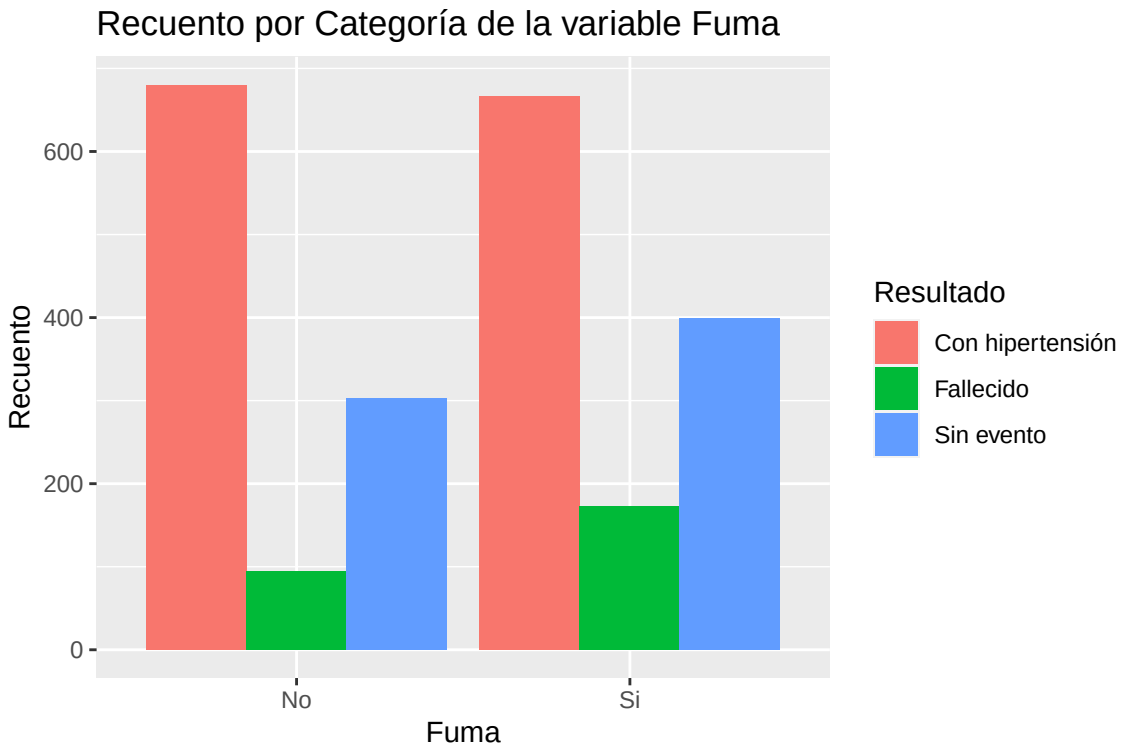
La data está contenida en un archivo de excel llamado *BaseRegresionMult*, para luego importarla a R. Para analizar se requiere que convertir a factores las variables categóricas para que R las tome como niveles, el siguiente código hace posible lo anterior:

```
library(readxl)
cardio= read_excel("BaseRegresionMult.xlsx")
cardio$Sexo=as.factor(cardio$Sexo)
cardio$Resultado=as.factor(cardio$Resultado)
cardio$Fuma=as.factor(cardio$Fuma)
summary(cardio)#estadísticas descriptivas del conjunto de datos
```

```
       Sexo        Colesterol         Edad            IMC
Femenino :1297   Min.   :113.0   Min.   :32.00   Min.   :15.54
Masculino:1019   1st Qu.:200.0   1st Qu.:41.00   1st Qu.:22.41
                 Median :227.0   Median :46.00   Median :24.50
                 Mean   :230.3   Mean   :47.43   Mean   :24.78
                 3rd Qu.:257.0   3rd Qu.:53.00   3rd Qu.:26.88
                 Max.   :464.0   Max.   :69.00   Max.   :39.94
    BPvar           Frecuencia        Glucosa       Fuma
Min.   :-43.250   Min.   : 44.00   Min.   : 40.00   No:1078
1st Qu.:-11.750   1st Qu.: 66.00   1st Qu.: 70.00   Si:1238
Median : -2.500   Median : 74.00   Median : 77.00
Mean   : -3.453   Mean   : 74.17   Mean   : 78.54
3rd Qu.:  5.312   3rd Qu.: 80.00   3rd Qu.: 85.00
Max.   : 19.500   Max.   :130.00   Max.   :163.00
          Resultado     Cigarrillos
Con hipertensión:1346   Min.   : 0.00
Fallecido       : 268   1st Qu.: 0.00
Sin evento      : 702   Median : 3.00
                        Mean   : 9.71
                        3rd Qu.:20.00
                        Max.   :60.00
```

La edad promedio de los pacientes es de 47 años y la persona mas joven sometido a esta prueba es de 32 años, del cual 1297 son personas del sexo femenino y 1019 del sexo masculino. A continuación se tiene un gráfico de barras agrupados por categoría de resultados. Es interesante observar que hay mayor número de personas fallecidas que si fuman y mayor número de personas que no poseen un resultado desfavorable aquellas que no fuman.

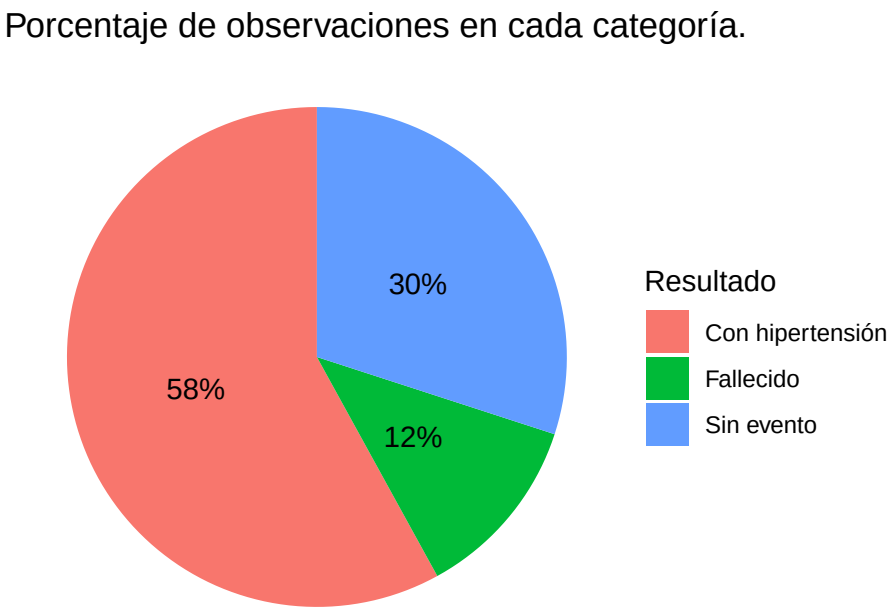
```
library(ggplot2)
ggplot(data = cardio, aes(x = Fuma, fill = Resultado)) +
  geom_bar(position = "dodge")+
  labs(title = "Recuento por Categoría de la variable Fuma",
       y = "Recuento")
```



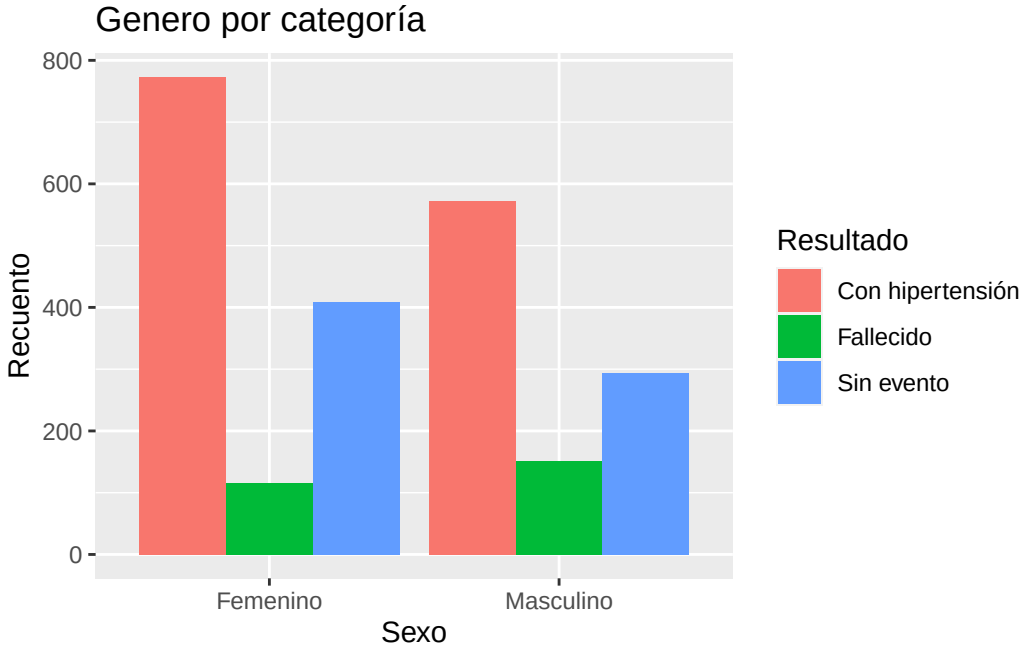

```
library(dplyr)

cardio %>%
  count(Resultado) %>%
  mutate(porcentaje = round(n/sum(n)*100), 1) %>%
  ggplot(mapping = aes(factor(1), y = porcentaje, fill = Resultado)) +
  geom_bar(stat = "identity") +
  coord_polar(theta = "y") +
  theme_void() +
  geom_text(aes(label = paste(porcentaje,"%", sep = "")),
            position = position_stack(vjust = 0.5))+
```

```
  labs(title = "Porcentaje de observaciones en cada categoría.")
```




```
library(ggplot2)
ggplot(data = cardio, aes(x = Sexo, fill = Resultado)) +
  geom_bar(position = "dodge")+labs(title = "Genero por categoría", y="Recuento")
```

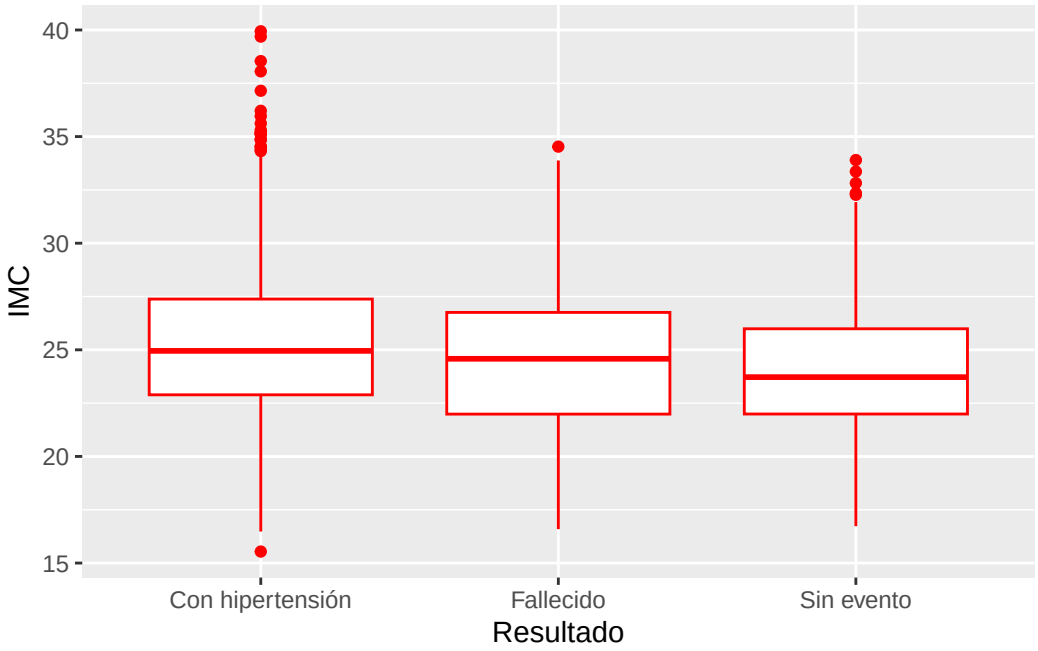


```
str(cardio)
```

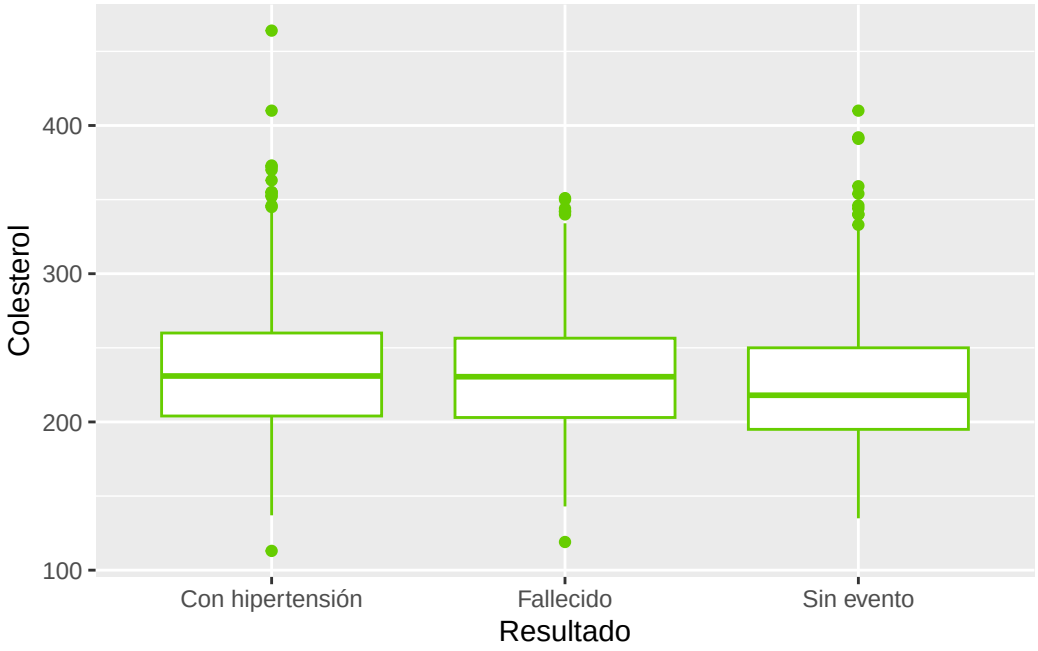
```
tibble [2,316 x 10] (S3: tbl_df/tbl/data.frame)
 $ Sexo       : Factor w/ 2 levels "Femenino","Masculino": 2 1 2 1 1 2 1 1 1 2 ...
 $ Colesterol : num [1:2316] 195 250 245 254 247 180 243 237 208 216 ...
 $ Edad       : num [1:2316] 39 46 48 50 43 36 43 41 49 51 ...
 $ IMC        : num [1:2316] 27 28.7 25.3 22.9 27.6 ...
 $ BPvar      : num [1:2316] -17 1.5 3.75 2.5 13.5 -11.5 -1.75 -1 -29 -18 ...
 $ Frecuencia : num [1:2316] 80 95 75 75 72 71 68 75 65 90 ...
 $ Glucosa    : num [1:2316] 77 76 70 76 61 80 78 74 98 95 ...
 $ Fuma       : Factor w/ 2 levels "No","Si": 1 1 2 1 1 1 2 2 1 1 ...
 $ Resultado  : Factor w/ 3 levels "Con hipertensión",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Cigarrillos: num [1:2316] 0 0 20 0 0 0 10 1 0 0 ...
```

Todas las variables contenidas de la data son numéricas, excepto *Fuma, Sexo* y *Resultado*

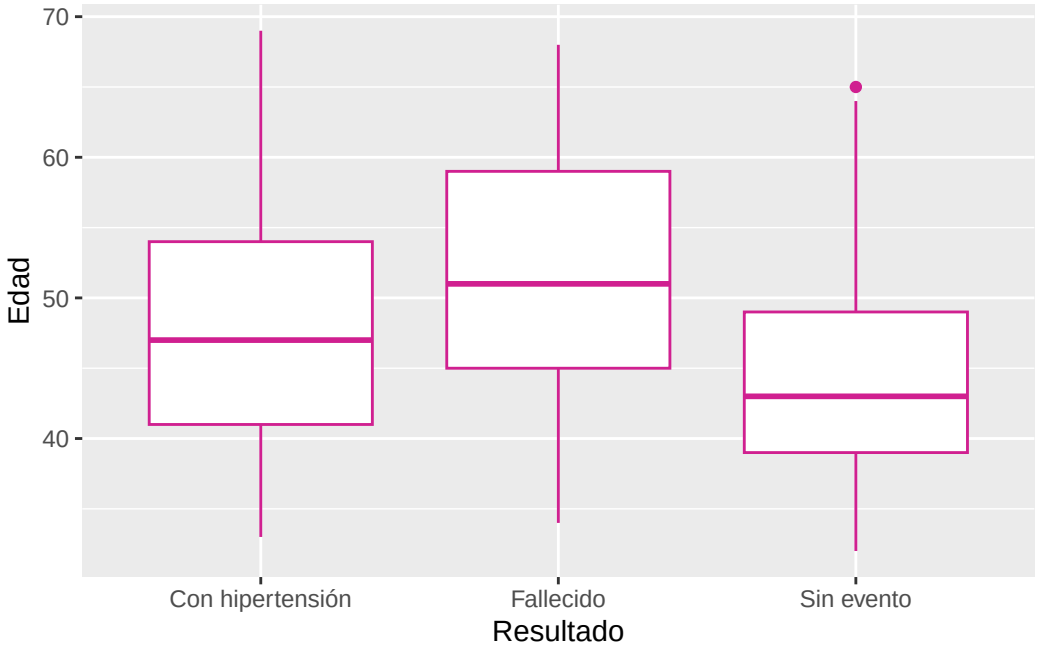
```
ggplot(data = cardio, aes(x = Resultado, y = IMC)) +
  geom_boxplot(col="red")
```



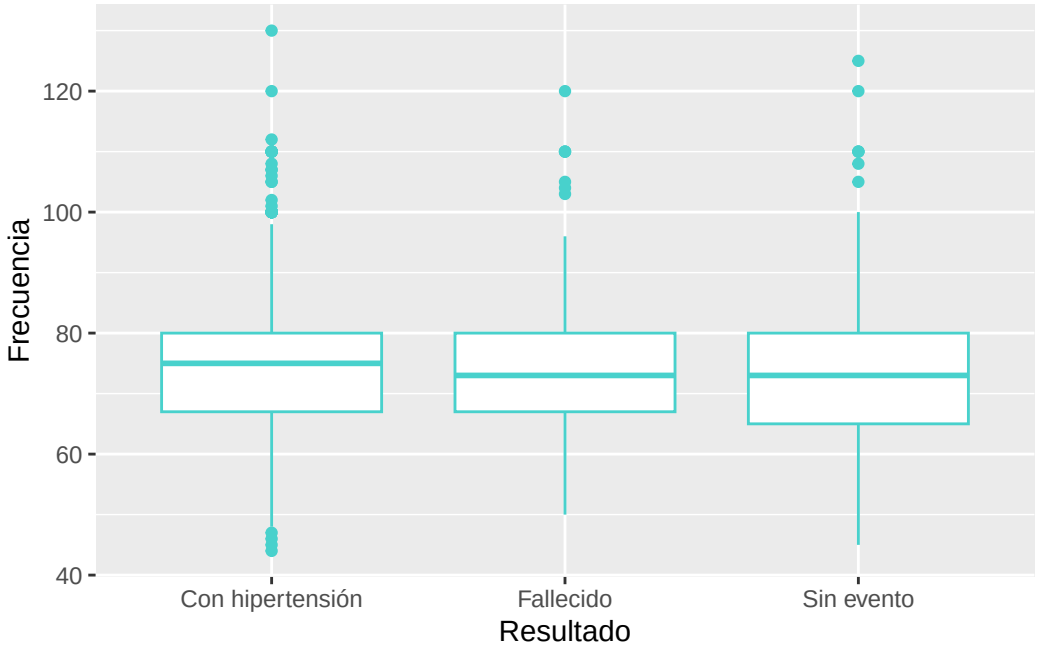

```
ggplot(data = cardio, aes(x = Resultado, y =Colesterol )) +
  geom_boxplot(col="chartreuse3")
```




```
ggplot(data = cardio, aes(x = Resultado, y =Edad )) +
  geom_boxplot(col="#D02090")
```




```
ggplot(data = cardio, aes(x = Resultado, y =`Frecuencia`)) +
  geom_boxplot(col="mediumturquoise")
```



Se observa que para las 3 categorías el promedio del nivel de colesterol es igual, también para la variable Edad y Frecuencia Cardíaca, esto puede dar indicios de no hay un efecto significativo de la variable categórica en la variable numérica. Para asegurar esto, hay que ver si estas variables contribuyen a la predicción de la variable dependiente realizando el modelo de regresión.

2.2 Estimación del Modelo de Regresión

Para efectos de este estudio se tomará como referencia a la categoría "Sin evento", pues el objetivo es realizar comparaciones con las categorías que presentan resultados desfavorables de aquellos pacientes de acuerdo a características dadas respecto de la categoría que no posee ningún evento. Para encontrar la probabilidad de que una persona pertenezca a alguno de los resultados respecto de los distintos factores asociados, como el nivel de Glucosa, Colesterol, Masa Muscular y otros, se estima el modelo de regresión multinomial y su significatividad global a través de la librería VGAM.

```
library(VGAM)
```

```
Warning: package 'VGAM' was built under R version 4.3.1
Loading required package: stats4
Loading required package: splines
```

```
#Modelo Nulo
model0<-vglm(Resultado~1,data=cardio,family=multinomial(refLevel="Sin evento"))

#Modelo con todos los predictores
model1=vglm(Resultado~.,data=cardio,family=multinomial(refLevel="Sin evento")
            ,model=TRUE)
#model1

lrtest(model0,model1)
```

```
Likelihood ratio test

Model 1: Resultado ~ 1
Model 2: Resultado ~ .
   #Df  LogLik  Df  Chisq Pr(>Chisq)
1 4630 -2146.4
2 4612 -1894.8 -18 503.27  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El primer modelo que se observa contiene únicamente por los interceptos β_0, es decir que las variables independientes no están dentro de este modelo y sus coeficientes son nulos, por otro lado, el segundo modelo contiene a todas las variables independientes. El objetivo con realizar estos dos modelos es calcular el ratio de verosimilitud entre ambos(deviance), si el del segundo modelo es más pequeño entonces se dice que existe alguna variable que ejerce influencia significativa en la variable dependiente. De la salida anterior se observa que el p valor de correspondiente al de ji-cuadrado es significativo, recordando que para saber si la diferencia es significativa, se debe calcular la diferencia entre las máximas verosimilitudes de los modelos y distribuirla como una χ^2 con grados de libertad igual a la diferencia en el número de parámetros de los modelos $(k_M - k_0)$. Por tanto existe algún factor (β distinto de cero) que influye en el Resultado, estos es, el modelo completo se ajusta significativamente mejor que el modelo reducido.

2.3 Efecto de cada predictor

En la sección anterior se verifica que al menos un predictor es distinto de cero, y efectivamente al menos uno es distinto de cero, ahora se evaluará para identificar cuales de las variables son las que aportan en la explicación de la variable Resultado.

```
summary(model1)
```

```
Call:
vglm(formula = Resultado ~ ., family = multinomial(refLevel = "Sin evento"),
    data = cardio, model = TRUE)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept):1   -3.022454   0.704859  -4.288 1.80e-05 ***
(Intercept):2   -7.548877   1.060125  -7.121 1.07e-12 ***
SexoMasculino:1 -0.431568   0.115022  -3.752 0.000175 ***
SexoMasculino:2  0.047102   0.170385   0.276 0.782209
Colesterol:1    -0.001456   0.001286  -1.132 0.257720
Colesterol:2    -0.001965   0.001902  -1.033 0.301686
Edad:1           0.055027   0.007236   7.605 2.86e-14 ***
Edad:2           0.116141   0.010123  11.473  < 2e-16 ***
IMC:1            0.062394   0.016697   3.737 0.000186 ***
IMC:2            0.016036   0.024892   0.644 0.519438
BPvar:1          0.067725   0.004967  13.636  < 2e-16 ***
BPvar:2          0.031260   0.007115   4.393 1.12e-05 ***
Frecuencia:1     0.001721   0.004741   0.363 0.716606
Frecuencia:2     0.009780   0.007017   1.394 0.163436
Glucosa:1        0.002982   0.004251   0.702 0.482907
Glucosa:2        0.001200   0.006136   0.196 0.844899
FumaSi:1        -0.313644   0.161656  -1.940 0.052356 .
FumaSi:2         0.174896   0.238277   0.734 0.462947
Cigarrillos:1    0.023601   0.007334   3.218 0.001291 **
Cigarrillos:2    0.031078   0.009805   3.170 0.001527 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Residual deviance: 3789.532 on 4612 degrees of freedom

Log-likelihood: -1894.766 on 4612 degrees of freedom

Number of Fisher scoring iterations: 5

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):2'


Reference group is level  3  of the response
```

La salida anterior ofrece los resultados de los coeficientes estimados, la hipótesis que se verifica para estos coeficientes es análoga al contraste de significatividad global anterior. Como se tiene $k-1$ modelos de regresión, se observa que para cuando se hace la comparación de la categoría de referencia "Sin evento" con "Con hipertensión" resulta significativos los coeficientes de sexo masculino, Edad, Masa Muscular, BPvar y Cigarrillos por día. esto es: Los hombres en comparación con las mujeres tienen una menor probabilidad de pertenecer a la categoría de referencia en lugar de "Con hipertensión" o "Fallecido". Por otro lado si se tiene un aumento de masa muscular es más probable que esta persona esté en el grupo de los que tienen hipertensión, además entre más cigarrillos consuma su riego de poseer hipertensión es mayor en lugar de no tener un resultado negativo. En la siguiente comparación "Sin evento" con "Fallecido" resultan que los coeficientes que contribuyen al modelo son: La Edad, BPvar y cigarrillos por día, es decir que a menor edad es menos probable fallecer en comparación de obtener un resultado negativo, y también si la persona posee

un aumento en la variabilidad de presión arterial es más probable que esta persona fallezca en lugar de no poseer ningún resultado desfavorable, finalmente si la persona fuma una pequeña cantidad de cigarrillo es más probable que no posea ningún resultado desfavorable en lugar de fallecer.

2.4 Pronósticos de la Categoría de Resultado

Suponiendo que un paciente tiene las siguientes características:Masculino, con Colesterol de 195, 39 años de edad, 26 de masa corporal, variabilidad sanguínea de -17, frecuencia cardíaca de 80, Glucosa de 77, No fuma y no fuma ningún cigarrillo. cual es la probabilidad de que esta persona no posea ningún evento:

```
e1=exp(-3.02-0.43*0-0.001*195+0.055*39+0.062*26+0.067*-17+0.001*80+0.002*77
       -0.31*0+0.02*0)
e2=exp(-7.54+0.047*0-0.0019*195+0.116*39+0.016*26+0.03*-17+0.009*80
       +0.001*77
       +0.17*0+ 0.03*0)

pse = 1 / (1 + e1 + e2);pse
```

```
[1] 0.5669223
```

```
pch =e1 / (1 + e1 + e2);pch
```

```
[1] 0.3943435
```

```
pf = e2 / (1 + e1 + e2);pf
```

```
[1] 0.03873418
```

Para este caso esta observación tiene un 56.7% de probabilidad de pertenecer a la categoría *Sin evento*, 39.4% a *Con hipertensión* y 3.9% a *Fallecido*, es decir, que esta persona de acuerdo a las características dadas tiene mayor probabilidad de ser una persona *Con hipertensión* pues es la probabilidad más alta. En un modelo logit multinomial, la probabilidad de cada categoría se calcula en relación con las demás categorías, y la suma de las probabilidades para todas las categorías debe ser igual a 1.

2.5 Obtención de los *Risk Ratios*

La contribución relativa sin tener que preocuparse por las unidades de medidas de cada variable independiente se solventa a través de los coeficientes estimados estandarizados, es decir los *risk ratios*[1], que permiten interpretar el efecto de una variable independiente en términos de probabilidad de pertenecer a una categoría en comparación con la categoría de referencia.

[1] La interpretación se brinda en la Sección 4 de este documento

```
round(exp(coef(model1)),4)
```

```
(Intercept):1 (Intercept):2 SexoMasculino:1 SexoMasculino:2    Colesterol:1
       0.0487        0.0005          0.6495          1.0482          0.9985
 Colesterol:2         Edad:1          Edad:2           IMC:1           IMC:2
       0.9980        1.0566          1.1232          1.0644          1.0162
      BPvar:1       BPvar:2    Frecuencia:1    Frecuencia:2       Glucosa:1
       1.0701        1.0318          1.0017          1.0098          1.0030
    Glucosa:2       FumaSi:1        FumaSi:2   Cigarrillos:1   Cigarrillos:2
       1.0012        0.7308          1.1911          1.0239          1.0316
```

2.6 Evaluación del Ajuste del modelo estimado

Para determinar cuan bueno es el modelo, se hace la comparación entre los siguientes coeficientes: Como se observa los R^2 son relativamente bajos, lo que índica que el modelo no explica una gran proporción de la variabilidad en los datos. El R^2 de CoxSnell es un poco más alto, indica que el modelo explica un 19.53% de la variabilidad en la variable respuesta en relación a la categoría de referencia "Sin evento". También el R^2 de Nagelkerke, establece que el modelo es capaz de explicar un 23.16% de los cambios en la variable respuesta en comparación con el modelo nulo.

```
library(DescTools)

PseudoR2(model1, which = "all")
```

```
      McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
     0.1172368       0.1079189       0.1953143       0.2316019       0.1785122
VeallZimmermann           Efron McKelveyZavoina            Tjur          logLik
     0.2748208       0.2696445       0.5294955              NA   -1894.7657547
       logLik0              G2
 -2146.4031993     503.2748892
```

3. Aplicación de Regresión Logística Multinomial en Python

3.1 Análisis Exploratorio en Python

```
import pandas as pd
cardio1 = pd.read_excel(r"BaseRegresionMult.xlsx")
cardio1.head(20)
```

	Sexo	Colesterol	Edad	IMC	...	Glucosa	Fuma	Resultado	Cigarrillos
0	Masculino	195	39	26.97	...	77	No	Sin evento	0
1	Femenino	250	46	28.73	...	76	No	Sin evento	0
2	Masculino	245	48	25.34	...	70	Si	Sin evento	20
3	Femenino	254	50	22.91	...	76	No	Sin evento	0
4	Femenino	247	43	27.64	...	61	No	Sin evento	0
5	Masculino	180	36	27.78	...	80	No	Sin evento	0
6	Femenino	243	43	26.87	...	78	Si	Sin evento	10
7	Femenino	237	41	23.28	...	74	Si	Sin evento	1
8	Femenino	208	49	20.68	...	98	No	Sin evento	0
9	Masculino	216	51	23.47	...	95	No	Sin evento	0
10	Masculino	223	38	23.01	...	78	Si	Sin evento	20
11	Femenino	302	52	23.51	...	87	No	Sin evento	0
12	Masculino	175	42	28.61	...	95	Si	Sin evento	10
13	Femenino	213	44	21.16	...	89	Si	Sin evento	20
14	Femenino	268	45	20.68	...	71	Si	Sin evento	9
15	Femenino	173	42	23.25	...	99	Si	Sin evento	20
16	Femenino	185	37	18.38	...	72	Si	Sin evento	5
17	Masculino	210	36	21.93	...	77	No	Sin evento	0
18	Femenino	241	39	26.12	...	87	No	Sin evento	0
19	Femenino	237	46	20.20	...	62	Si	Sin evento	3

[20 rows x 10 columns]

Estadísticas descriptivas:

```
cardio1.describe()
```

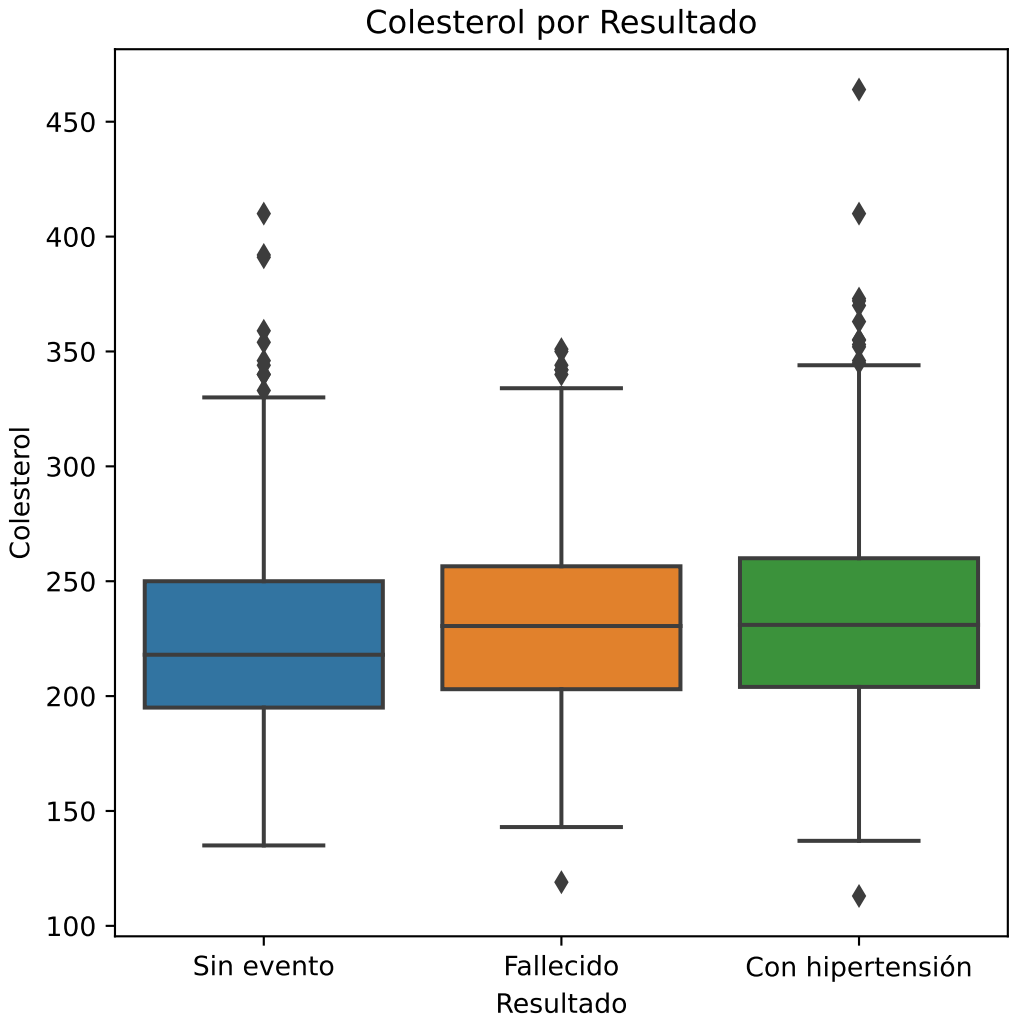
	Colesterol	Edad	...	Glucosa	Cigarrillos
count	2316.000000	2316.000000	...	2316.000000	2316.000000
mean	230.335924	47.431347	...	78.535838	9.710276
std	42.137760	8.120424	...	12.250940	11.796229
min	113.000000	32.000000	...	40.000000	0.000000
25%	200.000000	41.000000	...	70.000000	0.000000
50%	227.000000	46.000000	...	77.000000	3.000000
75%	257.000000	53.000000	...	85.000000	20.000000
max	464.000000	69.000000	...	163.000000	60.000000

[8 rows x 7 columns]

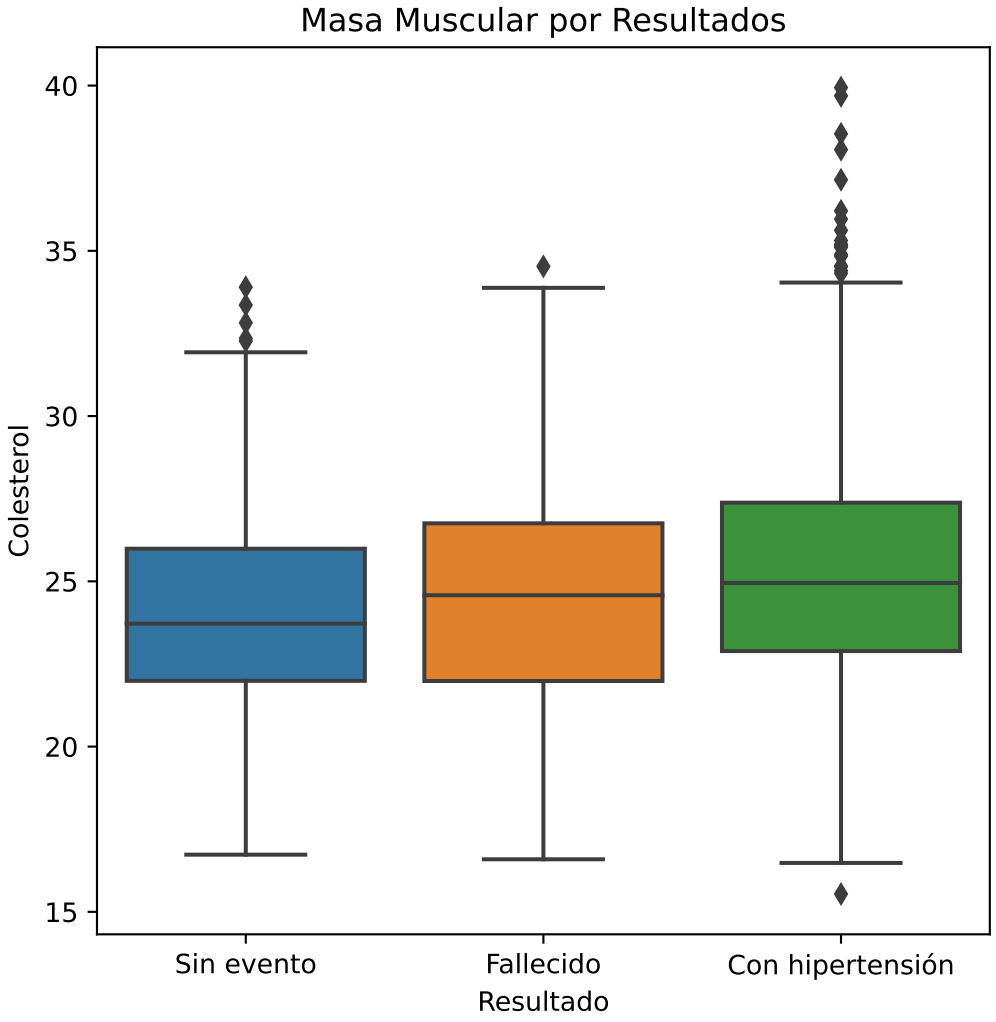
```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(6, 6))
sns.boxplot(data=cardio1, x='Resultado', y='Colesterol')
plt.title('Colesterol por Resultado')
```

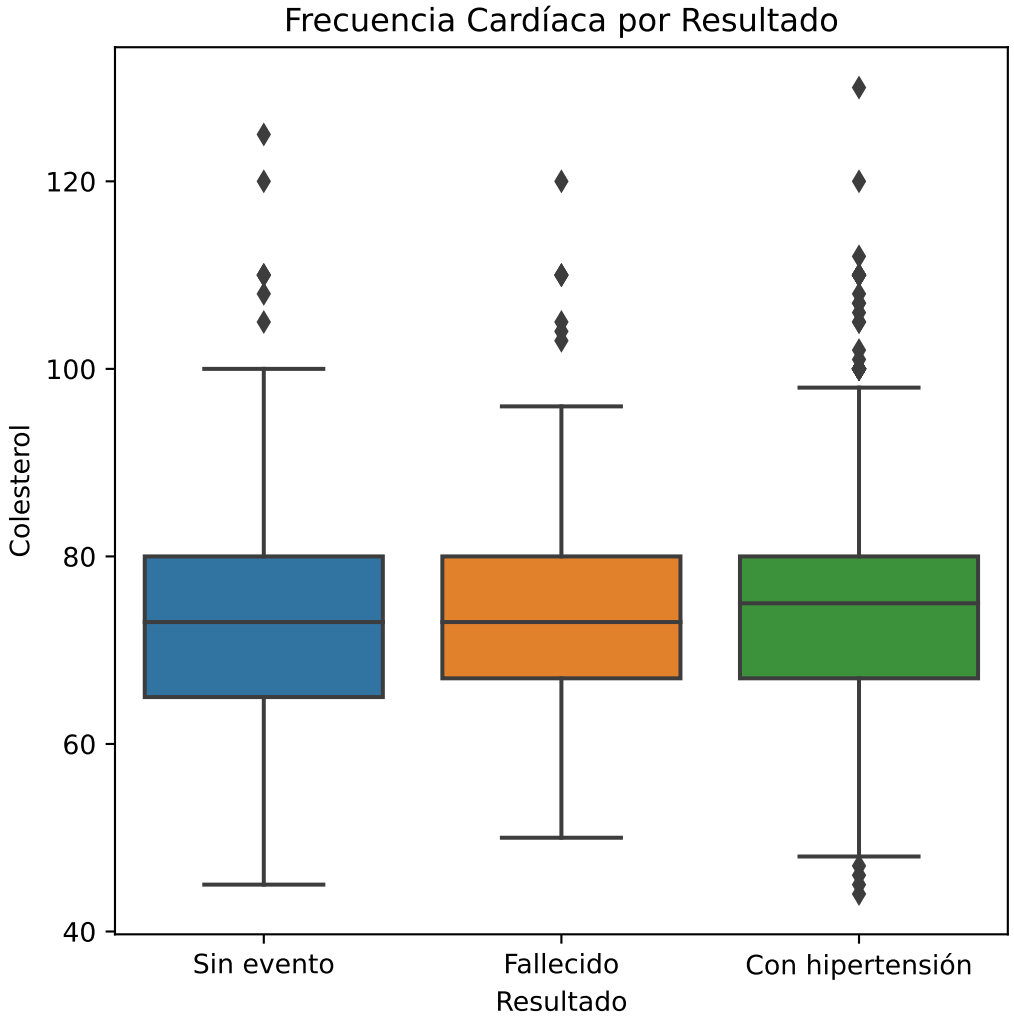
```
plt.xlabel('Resultado')
plt.ylabel('Colesterol')
plt.show()
```




```
plt.figure(figsize=(6, 6))
sns.boxplot(data=cardio1, x='Resultado', y='IMC')
plt.title('Masa Muscular por Resultados')
plt.xlabel('Resultado')
plt.ylabel('Colesterol')
plt.show()
```




```
plt.figure(figsize=(6, 6))
sns.boxplot(data=cardio1, x='Resultado', y='Frecuencia')
plt.title('Frecuencia Cardíaca por Resultado')
plt.xlabel('Resultado')
plt.ylabel('Colesterol')
plt.show()
```



3.2 Regresión Logística utilizando Usando Sci-Kit Learn

A continuación se muestran las bibliotecas generales de aprendizaje automático de Python para realizar el análisis de Regresión Logística en los resultados obtenidos. Algunas funciones de estas librerías son: `LogisticRegression` para el modelo de regresión logística multinomial. `train_test_split` para dividir los datos en conjuntos de entrenamiento y prueba. `accuracy_score` y `classification_report` para evaluar el rendimiento del modelo.

```
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
```

```
import  matplotlib.pyplot as plt
from sklearn.metrics import accuracy_score, classification_report
```

3.3 Codificación de variables categóricas

Antes de aplicar cualquier modelo de aprendizaje automático, hay que codificar aquellas variables que son categóricas, en este caso esta característica las cumplen las variables: *Resultado*, *Sexo* y *Fuma*. A continuación se muestra el mapeo utilizado para realizar el cambio en la codificación de las variables, este paso se realiza con el fin de que el modelo pueda comprender y utilizar esta información en el proceso de entrenamiento.

```
resultados2={"Sin evento":1,"Con hipertensión":2, "Fallecido":3}
cardio1["Resultado"]=cardio1["Resultado"].map(resultados2)# Recodificación de la Resultado

sexo={"Masculino":0,"Femenino":1} # Recodificación de la variable Sexo
cardio1["Sexo"]=cardio1["Sexo"].map(sexo)


fuma={"No":0,"Si":1}            # Recodificación de la variable Fuma
cardio1["Fuma"]=cardio1["Fuma"].map(fuma)
```

3.4 Estimación del Modelo de Regresión

Antes de estimar el modelo, hay que realizar la división de los datos en conjuntos de entrenamiento y prueba. Esto se hace para evaluar el rendimiento del modelo de manera adecuada, con la función train_test_split de scikit-learn:

```
X = cardio1.drop('Resultado', axis=1)# en x las variables independientes
#y borramos la dependiente de la data
y=cardio1['Resultado']# asignamos en la variable dependiente la variable Resultado
```

División de la data en train(70%) y testing (30%):

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
#División de datos
#en conjuntos de entrenamiento y prueba.

#instanciamos el modelo
modelLog = LogisticRegression(multi_class='multinomial', solver='lbfgs', max_iter=10000)
# entrenamiento del modelo
modelLog.fit(X_train, y_train)
```

```
LogisticRegression(max_iter=10000, multi_class='multinomial')
```

El modelo ”aprende” a partir de los datos de entrenamiento y estima los coeficientes que se utilizarán para hacer predicciones.

```
# Predecir los valores en el conjunto de prueba
y_pred = modelLog.predict(X_test)

#y_pred-> arroja las categorías que pertenecen a cada observación.
```

3.5 Precisión del modelo y Matriz de Confusión

Una vez entrenado el modelo, se puede evaluar su rendimiento en los datos de prueba utilizando métricas adecuadas para problemas de clasificación multinomial, como la precisión(accuracy), la matriz de confusión o cualquier otra métrica relevante.

```
precision = accuracy_score(y_test, y_pred)
round(precision,4)
```

```
0.6187
```

La precisión del modelo es de 0.60 lo que resulta aceptable.

```
from sklearn.metrics import confusion_matrix
confusion_matrix = confusion_matrix(y_test, y_pred)
print(confusion_matrix)
```

```
[[ 85 126   1]
 [ 57 340   2]
 [ 18  61   5]]
```

La matriz de confusión indica en su diagonal principal como el modelo ha clasificado correctamente y por debajo de esta, están los falsos positivos, es decir aquellos que fueron clasificados en esa categoría cuando en realidad no lo son, y por encima de la diagonal los falsos negativos, en donde le modelo ha clasificado a las observaciones no pertenecientes a esa categoría cuando en verdad si pertenecen.

Para pronosticar un nuevo caso con el modelo se tiene una observación con las caracteristicas siguientes: suponiendo que la persona es del sexo masculino, con Colesterol de 195, 39 años de edad, 26 de masa corporal, variabilidad sanguínea de -17, frecuencia cardíaca de 80, Glucosa de 77, No fuma y no fuma ningún cigarrillo.

```
pres=modelLog.predict([[0, 195, 39, 26.97, -17, 80, 77, 0, 0]])
print(pres)
```

```
[1]
```

Estos datos corresponde a la primera observación de la data que esta clasificado que no tiene resultado desfavorable, al realizar la predicción con el modelo, efectivamente lo clasifica en la categoría 1.

3.6 Probabilidades calculadas

```
df=pd.DataFrame( modelLog.predict_proba(X_test),columns=modelLog.classes_)
df['Suma'] = df.sum(axis=1)
df["Resultado Predicho"]=y_pred
df["Resultado observado"]=y_test.to_frame().reset_index().drop(columns="index")

df.head(15)
```

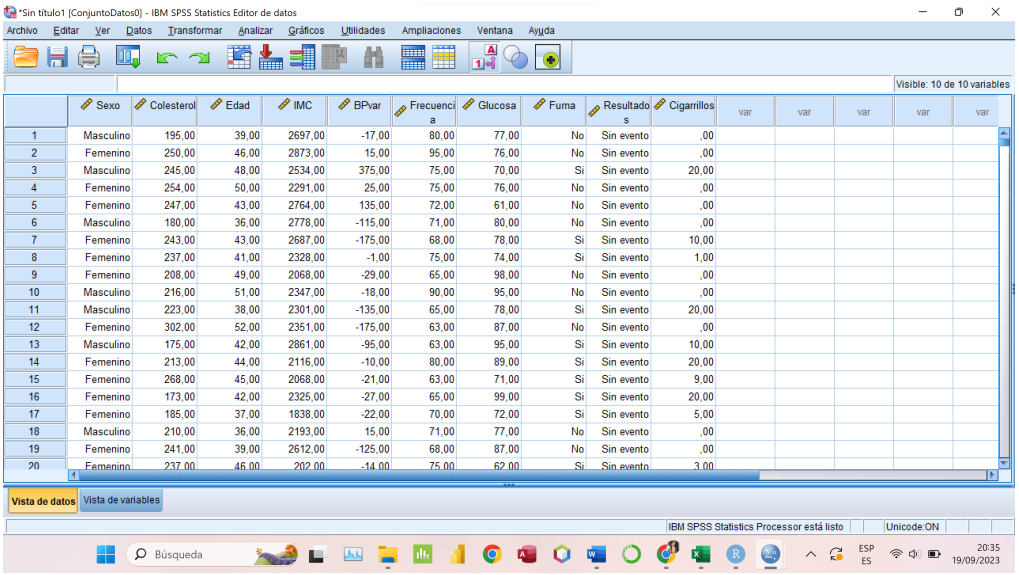
	1	2	3	Suma	Resultado Predicho	Resultado observado
0	0.266256	0.608618	0.125126	1.0	2	2
1	0.387702	0.551187	0.061112	1.0	2	3
2	0.343617	0.489944	0.166439	1.0	2	1
3	0.173904	0.765611	0.060486	1.0	2	1
4	0.167181	0.457074	0.375745	1.0	2	3
5	0.283136	0.682605	0.034259	1.0	2	2
6	0.774910	0.201082	0.024008	1.0	1	2
7	0.274706	0.456319	0.268975	1.0	2	1
8	0.160391	0.521419	0.318190	1.0	2	1
9	0.484203	0.478162	0.037634	1.0	1	2
10	0.173887	0.635222	0.190891	1.0	2	2
11	0.214856	0.723143	0.062001	1.0	2	3
12	0.313628	0.528844	0.157529	1.0	2	2
13	0.162018	0.627848	0.210134	1.0	2	3
14	0.098469	0.864153	0.037378	1.0	2	2

La clase con mayor probabilidad es el resultado de la clase predicha. Además, esta la columna en donde está la predicción y el resultado observado, y se evidencia que en la categoría 2 acierta en clasificación el modelo mientras que cuando el valor observado es la categoria 3 el modelo no acierta en la clasificación, luego esta la suma de las probabilidades que es igual a 1.

4. Aplicación de Regresión Logística en SPSS

SPSS es un software que facilita mucho la aplicación de esta técnica, puesto que es mas gráfico en comparación que cuando se realiza en R y Python. Para dar inicio, se especifican las variables que estarán en la aplicación y se proporcionan os datos, como sigue:

Figura 1: Vista de datos en SPSS



En la parte superior selecciona *Analizar>Logística>Regresión Multinomial.* luego se especifica la variable dependiente y covariables, para este caso en el modelo no se evaluaran interacciones de variables, por lo que se deja por defecto, Efectos Principales. En la parte de *Estadíscos* se selecciona *Pseudo R cuadrado, Informacion de Ajuste* y *criterios de Información.* Los Resultados que brinda el software son los siguientes:
1. Resumen de Procesamiento de casos el cual indica que la base esta conformada por un 30.3% de personas que no tienen ningún resultado negativo, 58.1% con hipertensión y finalmente un 11.6% han fallecido.

2. Información de ajuste del modelo: el siguiente cuadro se observa como el modelo final se ajusta significativamente mejor que el modelo nulo (solo con interceptos)

Figura 2: Ajuste del Modelo

Información de ajuste de los modelos

	Criterios de ajuste de modelo	Pruebas de la razón de verosimilitud		
Modelo	Logaritmo de la verosimilitud -2	Chi-cuadrado	gl	Sig.
Sólo intersección	4292,806			
Final	3789,532	503,275	18	,000

Figura 3: Bondad de Ajuste del Modelo

Bondad de ajuste

	Chi-cuadrado	gl	Sig.
Pearson	4584,322	4612	,611
Desvianza	3789,532	4612	1,000

Figura 4: Estimación de Parámetros

Estimaciones de parámetro

Resultado[a]		B	Desv. Error	Wald	gl	Sig.	Exp(B)	95% de intervalo de confianza para Exp(B) Límite inferior	Límite superior
Con hipertensión	Intersección	-3,336	0,706	22,321	1	0,000			
	Colesterol	-0,001	0,001	1,281	1	0,258	,999	,996	1,001
	Edad	0,055	0,007	57,829	1	0,000	1,057	1,042	1,072
	IMC	0,062	0,017	13,964	1	0,000	1,064	1,030	1,100
	BParv	0,068	0,005	185,931	1	0,000	1,070	1,060	1,081
	Frecuencia	0,002	0,005	,132	1	0,717	1,002	,992	1,011
	Glucosa	0,003	0,004	,492	1	0,483	1,003	,995	1,011
	Cigarrillos	0,024	0,007	10,355	1	0,001	1,024	1,009	1,039
	[Sexo=0]	-0,432	0,115	14,078	1	0,000	,649	,518	,814
	[Sexo=1]	0[b]	.	.	0	0.	.	.	.
	[Fuma=0]	0,314	0,162	3,764	1	0,052	1,368	,997	1,879
	[Fuma=1]	0[b]	.	.	0	.	.	.	.
Fallecido	Intersección	-7,374	1,052	49,113	1	0,000			
	Colesterol	-0,002	0,002	1,067	1	0,302	,998	,994	1,002
	Edad	,116	0,010	131,619	1	0,000	1,123	1,101	1,146
	IMC	,016	0,025	,415	1	0,519	1,016	,968	1,067
	Bparv	,031	0,007	19,302	1	0,000	1,032	1,017	1,046
	Frecuencia	,010	0,007	1,942	1	0,163	1,010	,996	1,024
	Glucosa	,001	0,006	,038	1	0,845	1,001	,989	1,013
	Cigarrillos	,031	0,010	10,046	1	0,002	1,032	1,012	1,052
	[Sexo=0]	,047	0,170	,076	1	0,782	1,048	,751	1,464
	[Sexo=1]	0[b]	.	.	0	.	.	.	.
	[Fuma=0]	-,175	0,238	,539	1	0,463	,840	,526	1,339
	[Fuma=1]	0[b]	.	.	0	.	.	.	.

a. La categoría de referencia es: Sin evento.

En la Figura 4 se obtienen las estimaciones de los parámetros del modelo, en la primera columna luego de las variables, esta los coeficientes del modelo B, luego el error estándar, el test de Wald asociado a cada coeficiente, grados de libertad, el p valor y finalmente el *risk ratio*.

En la columna del p valor, como antes se mencionaba que para cuando las categorías son Sin evento y Con Hipertensión las variables que contribuyen a la explicación de esta combinación son: Edad, IMC, BParv, cigarrillos y Sexo y de la misma manera cuando la comparación de Sin evento y Fallecido son las mismas variables que contribuyen excepto IMC.

En los risk ratios, un aumento de una unidad en la variable Edad se asocia con un aumento del 5.7% en las probabilidades de tener hipertensión en lugar de no tener un resultado desfavorable, manteniendo todas las demás variables constantes en el modelo. Además, un incremento en el índice de masa muscular

indica que es más probable que el paciente tenga hipertensión, luego si la variabilidad sanguínea disminuye en una unidad es 1.070 veces menos probable pertenecer al grupo de personas con hipertensión. Los hombres tienen un 36% (*risk ratio* de 0.64) menos de probabilidades de tener hipertensión en lugar de no tener evento en comparación con las mujeres.

Cuando se cambia a la categoría de fallecido con la de referencia ("Sin evento") las interpretaciones son análogas, es decir que un aumento en la edad existe un 12.3% de fallecer en comparación de un resultado desfavorable, así mismo un aumento en la variabilidad sanguínea se asocia con un 3.2% de probabilidad de fallecer en lugar de no tener ningún evento desfavorable y entre más cigarrillos consuma por día existe la probabilidad de un 3.2% de que fallezca en lugar de no tener ningún evento de salud.

3. Clasificación de los casos: muestra como fue hecha la clasificación(Figura 5), en la diagonal principal se muestran los casos que acertaron en la clasificación, en *Sin evento* se tienen 294 casos bien clasificados, 178 se clasificaron en esta misma siendo en realidad, observaciones con resultado que tiene hipertensión, y 45 que son fallecidos se clasificaron en la categoría *Sin evento*, y así se puede continuar con la clasificación con las demás categorías. Entre más valores existan en la diagonal mejor es la precisión del modelo.

Figura 5: Clasificación General

		Pronosticado		
Observado	Sin evento	Con hipertensión	Fallecido	Porcentaje correcto
Sin evento	294	402	6	41,9%
Con hipertensión	178	1164	4	86,5%
Fallecido	45	212	11	4,1%
Porcentaje global	22,3%	76,8%	0,9%	63,4%

5. Conclusión

Las enfermedades cardiovasculares, también conocidas como enfermedades del corazón, son un grupo de trastornos que afectan el corazón y los vasos sanguíneos. Estas enfermedades pueden ser potencialmente graves y, en algunos casos, mortales.

Las variables que resultan que tienen peso para determinar el desenlace de un paciente son la Edad, IMC, BParv, cigarrillos y Sexo. Entre más edad posea el paciente es más probable que éste, al cabo del tiempo desarrolle hipertensión o fallezca debido a una enfermedad cardiovascular, además si presente un aumento en la Masa muscular y fume un número alto de cigarrillos por día existe la probabilidad de que se categorice en el grupo de hipertensión o fallezca.

En general, la regresión logística es una herramienta estadística que puede ayudar en los diferentes campos de la ciencia, en este trabajo se evidencia como esta es útil para predecir cual es el resultado de acuerdo con las características presentaron los pacientes que se encuentran en la base de Datos de Framingham. Es muy accesible su aplicación, como se ha presentado en este documento, en R, Python y SPSS poseen esta herramienta y arrojan resultados similares con algunas diferencias muy pequeñas, si se desea una manera mas gráfica SPSS es el que encabeza la lista. Luego en Python, tenemos el aprendizaje automático el cual tiene la capacidad de aprender de datos y extraer patrones complejos lo convierte en una herramienta poderosa para la toma de decisiones, finalmente R lo hace de una forma más pausada y analítica.