Universidad de Valparaíso
CHILE

Facultad de Ciencias
Instituto de Estadística
Ingeniería en Estadística

PROYECTO MODELOS LINEALES GENERALIZADOS

1. ANTECEDENTES

1. Título tentativo del Proyecto:	Proyecto accidentes cerebrovasculares
2. Nombre del Estudiante:	Caamaño.S, Collao.C, Suárez.F
3. Rut:	20.302.641-2, 19.136.697-2, 20.301.416-3
4. Número Matrícula:	COMPLETAR
5. Dirección:	COMPLETAR
6. Teléfono:	COMPLETAR
7. Correo Electrónico:	sebastian.caamano@alumnos.uv.cl, cristobal.collao@alumnos.uv.cl, america.suarez@alumnos.uv.cl
8. Profesor Guía Propuesto:	COMPLETAR. Ph. D.
9. Profesor Co-guía:	COMPLETAR (DE SER NECESARIO) (AFILIACIÓN)
10. Fecha Presentación Proyecto:	30 Noviembre 2021

AUTORIZACIONES (Firmas se consignan en copia escrita)

Director de programa *COMPLETAR*		**Fecha:**
Profesor Guía *COMPLETAR*		**Fecha:**
Estudiante *COMPLETAR*		**Fecha:**

Formulación General del Proyecto

Existen un sin fin de problemáticas que con el buen uso de los datos correctos se le puede dar una solución efectiva. Dentro de las problemáticas de la salud existen ciertas dolencias que representan una fuerte amenaza a la vida de las personas.
Por lo que en este trabajo se presenta un modelo de predicción y prevención de los derrames cerebrales a partir de una base de datos disponible en internet y con variables que son determinantes en el riesgo de padecimiento de una derrame cerebral.
En el contexto de la salud pública, los derrames cerebrales se presentan como la segunda causa de muerte responsable de aproximadamente un $11\,\%$ de las muertes en el mundo según la Organización Mundial de la Salud (OMS). Frente a esto es necesario crear modelos para intentar prever la ocurrencia de estos. Para esto utilizaremos el dataset que se presenta en el siguiente apartado, con la finalidad de suministrar un modelo utilizable en la prevención de los derrames cerebrales.

Presentación de los datos

El conjunto de datos que se usa para el desarrollo de este proyecto son con respecto a casos de accidentes cerebrovasculares.
La base de datos cuenta con 5110 observaciones con 12 atributos cada una. Los atributos que se encuentran son posibles factores de riesgo en accidentes cerebrovaculares, se presentan a continuación:

- **id**: Identificador.
- **gender**: Género: Femenino(Female), Masculino(Male).
- **age**: Edad del individuo.
- **hypertension**: Hipertentensión: 0 si el individuo no presenta hipertensión, 1 si presenta hipertensión.
- **heart_disease**: Enfermedad al corazón: 0 si no presenta enfermedad al corazón, 1 si presenta.
- **ever_married**: Alguna vez casado: sí, no.
- **work_type**: Tipo de trabajo: Niño(children), Trabajo de gobierno(Govt_jov), Nunca ha trabajado(Never_worked), Privado (Private), Independiente (Self-employed).
- **Residence_type**: Zona de residencia: Rural, Urbana(Urban).
- **avg_glucose_level**: Nivel medio de glucosa en la sangre.
- **bmi**: Indice de masa coorporal
- **smoking_status**: Estado fumador: anteriormente fumador(formerly smoked), nunca(never smoked),fuma(smokes), desconocido(unknown).
- **stroke**: Accidente cerebrovascular: 1 el individuo presento un accidente cerebrovascular, 0 no se presento.

De la cual un $59\,\%$ corresponde a mujeres y el $41\,\%$ a hombres. Y la edad de estas personas van desde $0{,}08$ años(menos de 1 mes de vida) hasta los 82 años. Se tiene la información de hipertensión, en donde un $90{,}26\,\%$ no posee esta atribución mientras que el $9{,}74\,\%$ si. Los pacientes que tienen alguna enfermedad al corazón son 4834 y los que no padecen son 276. Las personas que han estado casadas alguna vez conforman el $66\,\%$ y por otro lado las que nunca han estado casadas abarcan el $34\,\%$.
Respecto al tipo de trabajo se tiene que 2925 trabajan en el sector privado, 819 son independientes, 687 son niños, 657 tienen un trabajo de gobierno y solo 22 individuos nunca han trabajado. El $51\,\%$ vive en el sector urbano y el resto $49\,\%$ en el sector rural. El nivel medio de glucosa en la sangre es variado, pero la mayor frecuencia se encuentra en el intervalo $[76{,}78 - 98{,}44]$ en el cual se encuentran 1790 personas entre esos niveles. El promedio del índice masa corporal es de $28{,}9$.
La cantidad de personas sin accidente cerebrovascular es bastante alta 4861 abarcando el $95{,}13\,\%$ de los datos, por lo que se tomo la decisión de acotar la data con filtro la edad, es decir, individuos mayores a 49 años. Y así se obtienen 1981 personas sin ACV, por el contrario 229 personas con accidentes cerebrovasculares.

Análisis descriptivo y visualización de los datos

Estudiando el conjunto de datos se observa una alta presencia de personas sin accidentes cerebrovasculares por lo que se tomo la decisión de realizar bootstrap para tener una nivelación de cantidad de ACV, por lo que la cantidad de accidentes cerebrovasculares no se mantiene como se expreso anteriormente, sino que ahora la cantidad personas sin y con ACV son la misma, es decir 500 y el total de individuos a estudiar es de 1000.
Y luego se realizó un resumen estadístico para los atributos numéricos, el cual se observa en el Cuadro 1. La edad mínima es de 50 años, como se explico anteriormente se acoto la base a personas con edad mayor o igual a 50 años, la edad máxima alcanza los 82 años y la media es de 67 años aproximadamente. El promedio de los niveles de glucosa es alrededor de 128 mientras que el nivel más bajo registrado es de 55.26, y el nivel máximo se encuentra al rededor de 272 aproximadamente. Observando el índice de masa corporal se aprecia que el mínimo es de 14 y el índice máximo es de 66.8 mientras que el promedio es de aproximadamente 31.

	Edad	**Glucosa**	**IMC**
Min	50	55.26	14.10
1st Qu	58.75	81.91	26.40
Mediana	68	102.11	29.50
Media	67.39	127.97	30.57
3rd Qu	77	190.14	33.60
Max	82	271.74	66.80

Cuadro 1: Resumen

Analizando los atributos gráficamente se tiene en la Figura 1 la densidad empírica de los derrames según la edad del individuo. Se destaca la alta presencia de accidentes cerebro vasculares en los rangos mayores de edad, es decir que se observa una tendencia al alza.

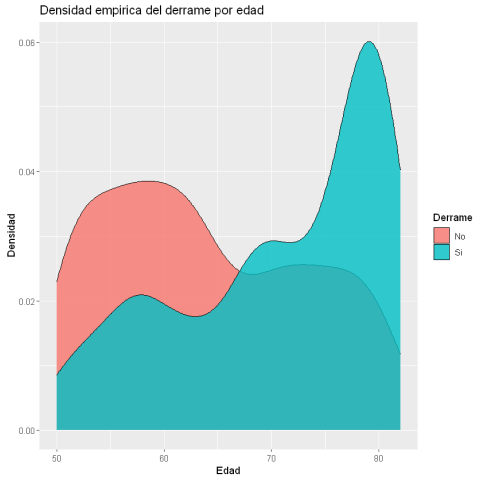

Figura 1: Densidad empírica del derrame por edad

Analizando la cantidad de personas que manifiestan ACV según su género, se obtiene que las mujeres son más propensas que los hombres a sufrir un accidente cerebro vascular.

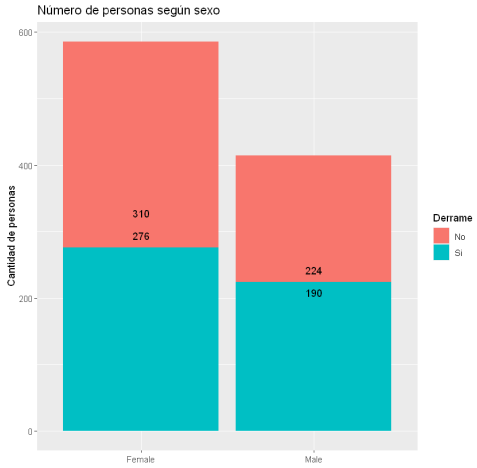

Figura 2: Número de personas según sexo

La cantidad de personas que presentaron accidentes cerebro vasculares es más alta en los individuos que no tienen enfermedades al corazón, resaltando que de la población son 856 los que no poseen enfermedades al corazón mientras que solo 144 los que si poseen.

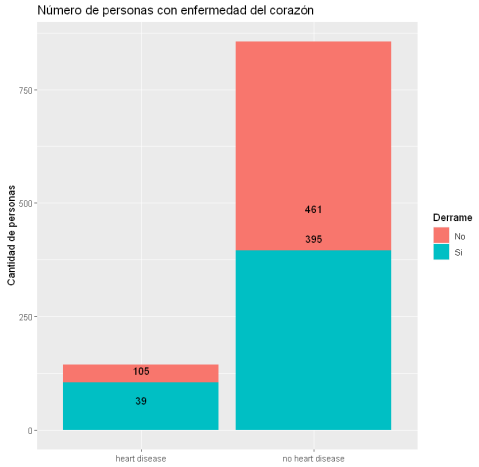

Figura 3: Número de personas con enfermedad del corazón

Analizando si la hipertensión es un factor determinante a la hora de que los individuos presenten ACV, al observar la Figura 4 notamos que 343 individuos presentan accidente cerebro vasculares pero no tienen hipertensión en cambio los que presentan hipertensión y ACV solo son 86, por lo que respecto a lo mencionado anteriormente se dice que la hipertensión no es influyente a simple vista como factor de riesgo.

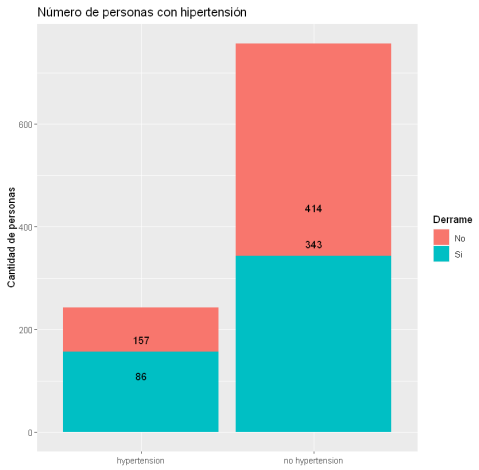

Figura 4: Número de personas con Hipertensión

¿El matrimonio o el haber estado casado/a alguna vez será un factor determinante a la hora de presentar un accidente cerebro vascular? Observando la Figura 5 se puede afirmar que es un factor determinante para los ACV, pero hay que considerar que no existe la misma cantidad de personas que han estado casadas con las que no, por lo que podremos analizar y corroborar esta aseveración más adelante en este proyecto.

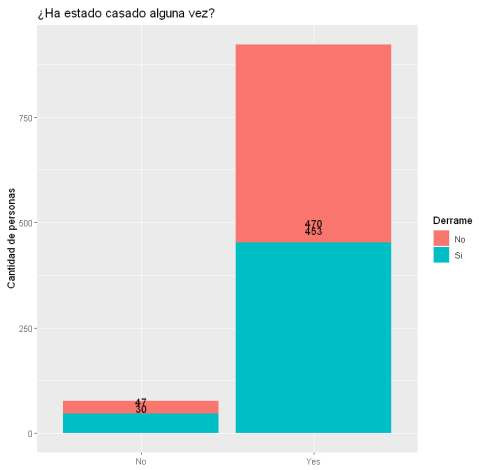

Figura 5: Número de personas casadas

Si se pregunta ¿a qué personas cree usted que es más probable que presentan un accidente cerebrovascular? intuitivamente se creería que a las personas con enfermedades de base, o enfermedades al corazón, personas casadas o fumadores. Pero como análisis preliminar se tiene en el Cuadro 2 los derrames según atributo, en el cual queda en evidencia la cantidad de derrames la cual casi no se ve afectada si la persona es fumadora o nunca ha fumado, pero también se observa que las personas sin hipertensión presentaron más accidentes cerebrovasculares, esto puede deberse a la cantidad de personas que presentan este atributo(n) ya que es más la cantidad de personas que no tiene hipertensión con respecto a las personas que si la poseen.

Atributo	**Proporción**	**n**	**Derrames**
Hipertensión	0.158	412	65
No hipertensión	0.091	1798	164
Enfermedad corazón	0.180	261	47
No enfermedad corazón	0.093	1949	182
Casado alguna vez	0.101	2038	205
No casado alguna vez	0.140	172	24
Anteriormente fumador	0.114	580	66
Nunca ha fumado	0.096	864	83
Fumador	0.104	355	37
Desconocido	0.105	411	43

Cuadro 2: Derrame según atributo

A continuación en el Cuadro 3 se observa la combinación de tipo de trabajo y la zona en la que reside el individuo, la cantidad de derrames no se ve afectada si la persona reside en una zona urbana o rural.

Tipo trabajo	Zona residencial	n	Proporción	Derrames
Gobierno	Rural	167	0.084	14
Gobierno	Urbano	194	0.093	18
Privado	Rural	603	0.1	60
Privado	Urbano	639	0.117	75
Independiente	Rural	294	0.105	31
Independiente	Urbano	313	0.099	31

Cuadro 3: Zonas residencial y actividad labora

Objetivo general

Construir un modelo que se ajuste a la predicción de los accidentes cerebro vasculares.

Objetivos específicos

1. Identificar las variables significativas en un modelo.
2. Comparar y evaluar distintos modelos realizados.

Hipótesis de Trabajo

Se pretende predecir la ocurrencia de accidentes cerebro vasculares respecto a los atributos presentes en la base de datos.Respecto la hipótesis general se tiene que las variables que pertenecen al conjunto de datos son factores de riesgo y son determinantes en la ocurrencia de accidentes cerebrovasculares.

Discusión Bibliográfica

Algunos autores (Ibañez y cols., 2009) de una tesis proponen ciertos factores asociados a los accidentes cerebrovasculares. En la cual la edad es el factor de riesgo más importante, ya que con el envejecimiento se producen numerosas alteraciones en el sistema vascular, que en conjunto con otros factores incrementan el riesgo de ACV. El riesgo de desarrollar un accidente cerebro vascular se duplica cada década desde los 55 años, ocurriendo más de la mitad de estos en mayores de 75 años.
Dado el género, se tiene que los hombres tienen una mayor incidencia de ACV que las mujeres, con un $25\,\%$ más, sin embargo la mortalidad es mayor en las mujeres.
El riesgo de desarrollar un ACV es doble en fumadores, este riesgo se relaciona con el número de cigarrillos fumados al día; cuanto más se fuma, mayor es el riesgo.
La edad se considera como el factor más importante. Con el envejecimiento se producen numerosas alteraciones en el sistema vascular, que junto a la actuación más prolongada en el tiempo de los otros factores de riesgo van a incrementar el riesgo de ACV. De hecho, el riesgo de desarrollar un accidente cerebro vascular se duplica cada década a partir de los 55 años, ocurriendo más de la mitad de los casos en pacientes mayores de 75 años.

Metodología

Dado la problemática y planteada la hipótesis, se lleva a cabo una investigación de tipo explicativa de corte cuantitativo que busca llegar a un resultado concluyente respecto de las variables que determinan la ocurrencia de los accidentes cerebrovasculares.
El proyecto ha sido realizado por medio de manipulación computacional del conjunto de datos a través del software y lenguaje de programación R.
El flujo de este proyecto ha sido primero, hacer limpieza de la base de datos y el preprocesamiento de los datos para alistarlos, luego hacer un análisis descriptivo de estos datos y presentar algunos gráficos, después empezar con el proceso de modelamiento estadístico probando y comparando distintos modelos y quitando las variables no significativas. Posteriormente de haber elegido el modelo, procedemos a evaluarlo y probar los resultados de su predicción.

Tratamiento de los datos

Para probar o desacreditar la hipótesis anteriormente mencionada al conjunto de datos se le realizaron una serie de análisis y manipulación de los datos. Primeramente se trataron los valores NA, al comienzo se propuso imputar los datos por la media de la variable, pero estos valores superan el $2\,\%$ de los datos de la variables no se realizó la imputación por la media. Por lo que se procedió a eliminar los valores NA con las respectivas variables correspondientes a la fila donde se encuentren estos NA. Se observo que la cantidad de personas sin accidentes cerebrovasculares correspondía al $95,12\,\%$ de los datos, por lo que se decidió acotar el conjunto de datos, a personas mayores de 50 años. Para realizar los análisis estadísticos correspondiente se balanceo el conjunto de datos, es decir, que se tomo en cuenta el $50\,\%$ de personas que no han presentado accidentes cerebro vasculares y el otro $50\,\%$ lo componen los individuos que si han tenido la ocurrencia de ACV.

Resultados

Tras la limpieza y preparación de los datos, se procede al análisis, modelamiento y evaluación de los distintos modelos que se barajaron en este trabajo. En primera instancia se aplicó un modelo que consideraba todas las variables de la base datos, el modelo extendido, los resultados de este modelo se muestran en la figura 6 :

```
Call:
glm(formula = stroke ~ gender + age + hypertension + heart_disease +
    ever_married + work_type + Residence_type + avg_glucose_level +
    bmi + smoking_status + stroke, family = binomial(link = "logit"),
    data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.45210  -0.96633  -0.07574   0.97228   2.02544

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -3.773148   0.867214  -4.351 1.36e-05 ***
genderMale                      0.044928   0.148002   0.304 0.761460
age                             0.068317   0.008177   8.354  < 2e-16 ***
hypertensionno hypertension    -0.614649   0.172513  -3.563 0.000367 ***
heart_diseaseno heart disease -0.599271   0.221775  -2.702 0.006889 **
ever_marriedYes                -0.613587   0.278730  -2.201 0.027710 *
work_typePrivate               -0.104210   0.214324  -0.486 0.626807
work_typeSelf-employed         -0.292335   0.239352  -1.221 0.221950
Residence_typeUrban            -0.238865   0.142571  -1.675 0.093854 .
avg_glucose_level               0.007914   0.001324   5.975 2.29e-09 ***
bmi                             0.001624   0.011911   0.136 0.891536
smoking_statusnever smoked     -0.270060   0.176220  -1.533 0.125395
smoking_statussmokes            0.210888   0.232944   0.905 0.365298
smoking_statusUnknown          -0.116966   0.222493  -0.526 0.599094
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.3  on 999  degrees of freedom
Residual deviance: 1174.7  on 986  degrees of freedom
AIC: 1202.7

Number of Fisher Scoring iterations: 4
```

Figura 6: Resumen estadístico modelo 1

Se puede ver que este modelo presenta muchas variables que resultan no ser significativas y que tienen un p-valor muy mayor a nuestro nivel de significancia de $5\,\%$. Se puede a simple vista pensar en no considerar algunas de las variables como el tipo de trabajo, el índice de masa corporal o si el sujeto fuma o no. Este modelo da una idea de como continuar el proceso de modelamiento.
En la siguiente Figura 7 se aplicó el modelo reducido dejando solamente las variables que resultaron significantes:

```
Call:
glm(formula = stroke ~ age + hypertension + heart_disease + ever_married +
    avg_glucose_level, family = binomial(link = "logit"), data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.30030  -0.95831  -0.09314   0.98180   1.92502

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -3.753948   0.650750  -5.769 7.99e-09 ***
age                           0.063157   0.007382   8.556  < 2e-16 ***
hypertensionno hypertension  -0.566261   0.167696  -3.377 0.000734 ***
heart_diseaseno heart disease -0.696492   0.215447  -3.233 0.001226 **
ever_marriedYes              -0.523434   0.273110  -1.917 0.055293 .
avg_glucose_level             0.007955   0.001245   6.388 1.68e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.3  on 999  degrees of freedom
Residual deviance: 1184.9  on 994  degrees of freedom
AIC: 1196.9

Number of Fisher Scoring iterations: 4
```

Figura 7: Resumen estadístico modelo 2

En este modelo se puede ver que la variables edad, presencia de hipertensión, problemas cardiaco, el haberse casado y los niveles de glucosa en la sangre serían las variables a determinar el riesgo de la ocurrencia de un accidente cerebro vascular. Se observa como los indicadores mejoran, como el AIC que se muestra menor al modelo extendido.
A continuación se presentan modelos que se probaron, en estos modelos se probó la interacción entre algunas variables.

```
Call:
glm(formula = stroke ~ age + hypertension * ever_married + avg_glucose_level +
    heart_disease, family = binomial(link = "logit"), data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2839  -0.9585  -0.0909   0.9826   1.9206

Coefficients:
                                          Estimate Std. Error z value
(Intercept)                              -3.526631   0.720294  -4.896
age                                       0.063294   0.007389   8.566
hypertensionno hypertension              -0.960342   0.545007  -1.762
ever_marriedYes                          -0.797253   0.458243  -1.740
avg_glucose_level                         0.007931   0.001245   6.369
heart_diseaseno heart disease            -0.688157   0.215761  -3.189
hypertensionno hypertension:ever_marriedYes  0.436524   0.572606   0.762
                                         Pr(>|z|)
(Intercept)                              9.78e-07 ***
age                                       < 2e-16 ***
hypertensionno hypertension               0.07806 .
ever_marriedYes                           0.08189 .
avg_glucose_level                        1.90e-10 ***
heart_diseaseno heart disease             0.00143 **
hypertensionno hypertension:ever_marriedYes  0.44585
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.3  on 999  degrees of freedom
Residual deviance: 1184.3  on 993  degrees of freedom
AIC: 1198.3

Number of Fisher Scoring iterations: 4
```

Figura 8: Resumen estadístico modelo 3

```
Call:
glm(formula = stroke ~ age + hypertension * avg_glucose_level +
    heart_disease + ever_married, family = binomial(link = "logit"),
    data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.29612  -0.95836  -0.09172   0.97962   1.92593

Coefficients:
                                            Estimate Std. Error z value
(Intercept)                                -3.7388847  0.7072548  -5.286
age                                         0.0631418  0.0073874   8.547
hypertensionno hypertension                -0.5859232  0.3986776  -1.470
avg_glucose_level                           0.0078440  0.0023919   3.279
heart_diseaseno heart disease              -0.6957920  0.2158456  -3.224
ever_marriedYes                            -0.5233799  0.2731233  -1.916
hypertensionno hypertension:avg_glucose_level  0.0001517  0.0027895   0.054
                                           Pr(>|z|)
(Intercept)                                1.25e-07 ***
age                                         < 2e-16 ***
hypertensionno hypertension                 0.14165
avg_glucose_level                           0.00104 **
heart_diseaseno heart disease               0.00127 **
ever_marriedYes                             0.05533 .
hypertensionno hypertension:avg_glucose_level  0.95664
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.3  on 999  degrees of freedom
Residual deviance: 1184.9  on 993  degrees of freedom
AIC: 1198.9
```

Figura 9: Resumen estadístico modelo 4

```
Call:
glm(formula = stroke ~ age + hypertension + avg_glucose_level +
    heart_disease * ever_married, family = binomial(link = "logit"),
    data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.29100  -0.95979  -0.09771   0.98880   1.92301

Coefficients:
                                             Estimate Std. Error z value
(Intercept)                                -3.323181   1.206940  -2.753
age                                         0.063137   0.007380   8.556
hypertensionno hypertension                -0.561705   0.168067  -3.342
avg_glucose_level                           0.007946   0.001246   6.378
heart_diseaseno heart disease              -1.159974   1.102428  -1.052
ever_marriedYes                            -0.973614   1.089852  -0.893
heart_diseaseno heart disease:ever_marriedYes  0.484580   1.123336   0.431
                                           Pr(>|z|)
(Intercept)                                0.005898 **
age                                         < 2e-16 ***
hypertensionno hypertension                0.000831 ***
avg_glucose_level                          1.79e-10 ***
heart_diseaseno heart disease              0.292708
ever_marriedYes                            0.371672
heart_diseaseno heart disease:ever_marriedYes 0.666195
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.3  on 999  degrees of freedom
Residual deviance: 1184.7  on 993  degrees of freedom
AIC: 1198.7

Number of Fisher Scoring iterations: 4
```

Figura 10: Resumen estadístico modelo 5

```
Call:
zeroinfl(formula = ifelse(as.numeric(data$stroke) == 1, 0, 1) ~ gender +
    age + hypertension + heart_disease + ever_married + avg_glucose_level +
    bmi, data = data, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.9036 -0.3210 -0.2347 -0.1830  6.6099

Count model coefficients (negbin with log link):
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -5.166890   0.837053  -6.173 6.71e-10 ***
genderMale                    0.023238   0.148068   0.157   0.8753
age                           0.051320   0.008027   6.393 1.62e-10 ***
hypertensionno hypertension  -0.386450   0.161653  -2.391   0.0168 *
heart_diseaseno heart disease -0.396234   0.183891  -2.155   0.0312 *
ever_marriedYes              -0.299442   0.243675  -1.229   0.2191
avg_glucose_level             0.005712   0.001241   4.603 4.16e-06 ***
bmi                          -0.015877   0.013001  -1.221   0.2220
Log(theta)                   16.694397  67.195480   0.248   0.8038

Zero-inflation model coefficients (binomial with logit link):
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                    66.315   3310.873   0.020    0.984
genderMale                     94.770    395.076   0.240    0.810
age                             9.918     42.611   0.233    0.816
hypertensionno hypertension   209.409   3394.553   0.062    0.951
heart_diseaseno heart disease -104.441    426.428  -0.245    0.807
ever_marriedYes               -267.749   1122.626  -0.239    0.811
avg_glucose_level                3.851     16.336   0.236    0.814
bmi                            -62.409    261.982  -0.238    0.812
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 17794438.5345
Number of iterations in BFGS optimization: 91
Log-likelihood: -591.9 on 17 Df
```

Figura 11: Resumen estadístico modelo 6

Comparando los modelos anteriormente mencionados se presenta la Figura 12 para saber y escoger cual modelo predice mejor por lo que se utiliza el criterio de AIC, el cual esta basado en el deviance y el número de parámetros del modelo. El modelo que presenta el menor AIC es el de los datos originales(Figura 7).

Modelos Planteados

Datos Originales

	0	1	Exactitud	Deviance	AIC
0	1898	1	0.9086124	1165.914	1193.914
1	190	1			

Modelo Balanceado, con variables significativas

	0	1	Exactitud	Deviance	AIC
0	348	152	0.71	1184.876	1196.876
1	138	362			

Balanceado Hipertension Casado

	0	1	Exactitud	Deviance	AIC
0	347	153	0.708	1184.283	1198.283
1	139	361			

Balanceado Hipertension Glucosa

	0	1	Exactitud	Deviance	AIC
0	349	151	0.711	1184.873	1198.873
1	138	362			

Balanceado Problema Cardiaco y Casado

	0	1	Exactitud	Deviance	AIC
0	348	152	0.71	1184.671	1198.671
1	138	362			

Figura 12: Comparación de modelos

Teniendo las variables que conformarán el modelo ya elegidas, se procede a revisar la distancia de Cook, con el fin de obtener los datos que más influían en el modelo, los resultados se muestran en la figura 13

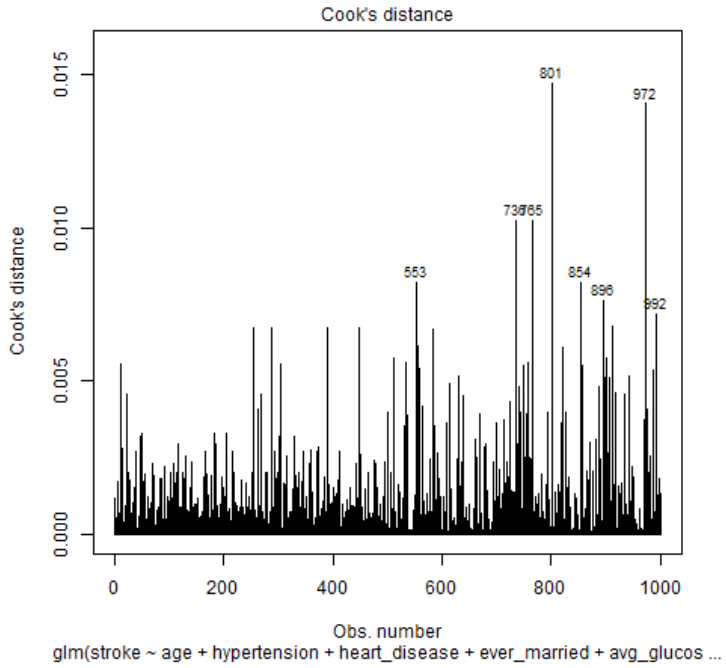

Figura 13: Gráfico distancia de cook

Luego de revisar estos datos influyentes obtuvimos datos de individuos que resultaron tener características no deseables para el modelo, como personas con niveles de glucosa muy altos, además de problemas cardiacos y hipertensión que no presentaron ACV, o lo contrario, sujetos que presentaron ACV sin tener factores de riesgos. Por lo tanto, para mejorar el modelo, se tomo la decisión de quitar estas observaciones.

Los resultados del modelo con las filas eliminadas es el siguiente:

```
Call:
glm(formula = stroke ~ age + hypertension + heart_disease + ever_married +
    avg_glucose_level, family = binomial(link = "logit"), data = newdata1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1698  -0.9235   0.2928   0.9424   1.9920

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -3.980534   0.677971  -5.871 4.33e-09 ***
age                           0.067704   0.007643   8.859  < 2e-16 ***
hypertensionno hypertension  -0.619899   0.173979  -3.563 0.000367 ***
heart_diseaseno heart disease -0.978546   0.232287  -4.213 2.52e-05 ***
ever_marriedYes              -0.438288   0.287595  -1.524 0.127515
avg_glucose_level             0.009239   0.001303   7.088 1.36e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1354.4  on 976  degrees of freedom
Residual deviance: 1121.3  on 971  degrees of freedom
AIC: 1133.3

Number of Fisher Scoring iterations: 4
```

Figura 14: Modelo finalizado

Se puede observar que el AIC disminuye, por lo que este será el modelo final. Lo siguiente será evaluar el modelo.
Para tener una noción visual de los efectos que las variables tienen en la variable respuesta.

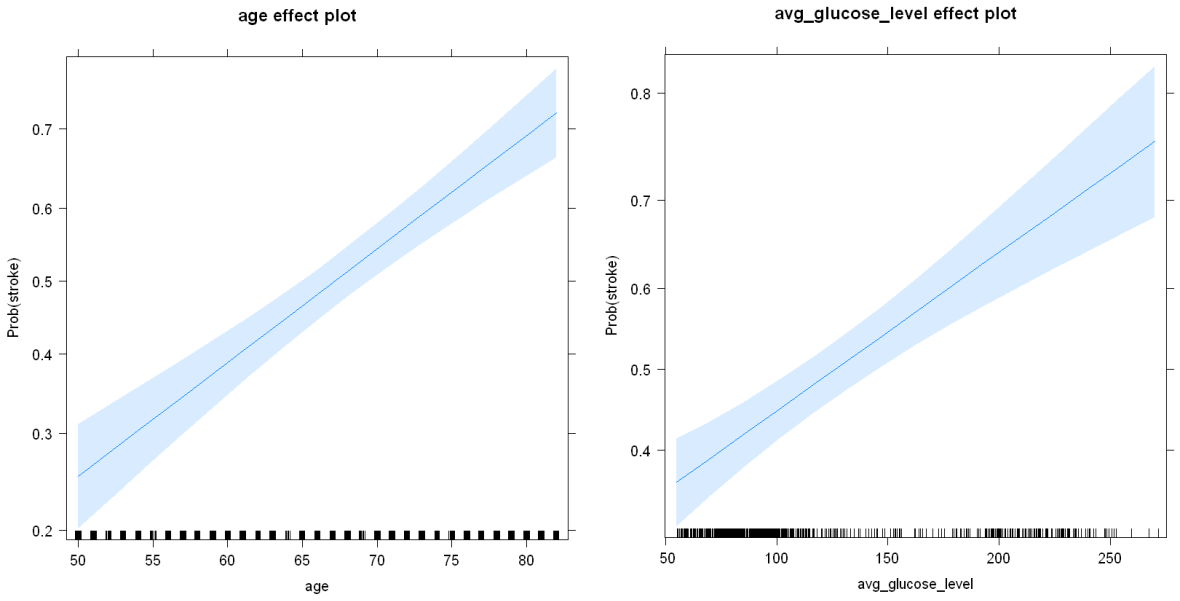

Figura 15: Efectos edad y glucosa

Se observa en este gráfico (ver Figura 15) como claramente la probabilidad de ocurrencia de un ACV aumenta a medida que aumentan los años y aumenta el nivel promedio de glucosa.

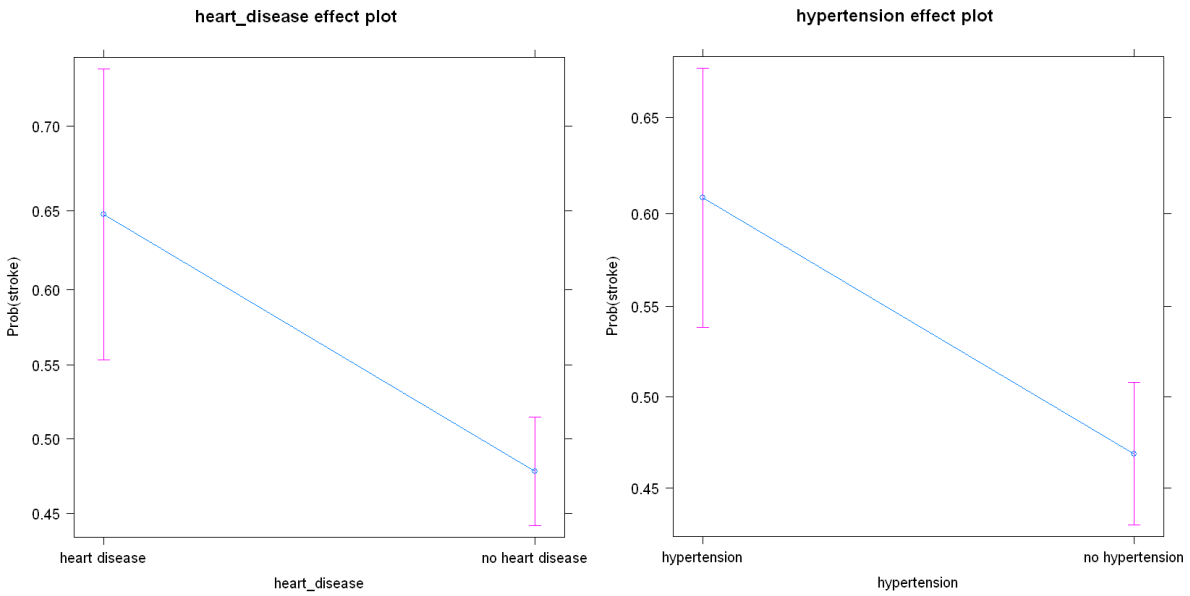

Figura 16: Efectos problemas cardiacos y hipertensión

También se puede ver el efecto que tiene la presencia de hipertensión y problemas cardiacos en la ocurrencia de un ACV.

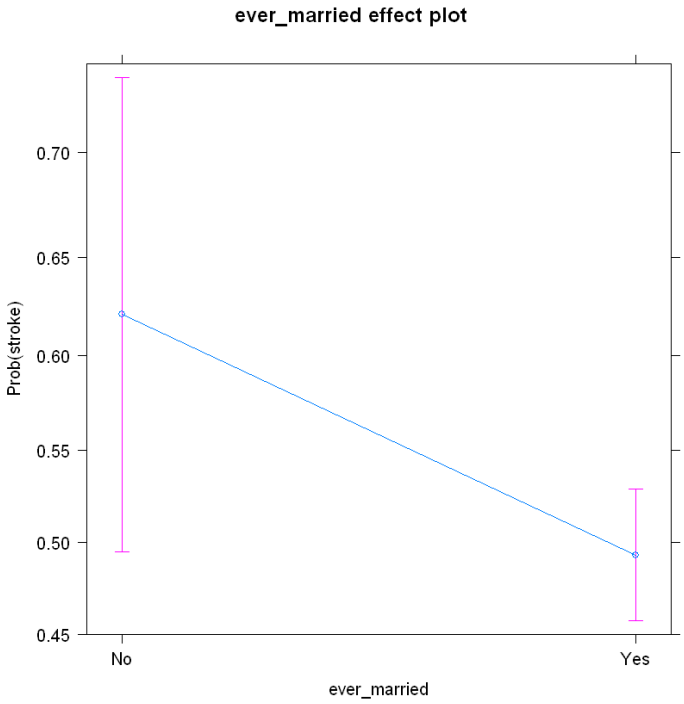

Figura 17: Efecto de haber estado casado

Finalmente se aprecia el efecto de haber estado casado o no en la ocurrencia del ACV.
Ahora se procede a testear el modelo con el test Pearson χ^2 de bondad de ajuste que se explica de la siguiente manera:

$$\chi^2 = \sum_{j=1}^{J} r_j^2$$

Donde los r^2 son los residuos de Pearson. El resultado del test resulto ser un valor 0, con lo que se explica un buen ajuste del modelo.
Lo siguiente será evaluar la capacidad predictiva del modelo, para esto se realiza una tabla de confusión donde los casos que queden en la diagonal serían los casos en el que el modelo predijo bien.

	Predicciones	
Observaciones	0	1
0	343	142
1	136	356

Cuadro 4: Matriz de confusión

Podemos decir que el modelo, de los datos, predice con un $71{,}54\,\%$ de precisión.
Para tener otra noción más gráfica de la capacidad predictiva del modelo presentamos el ROC.

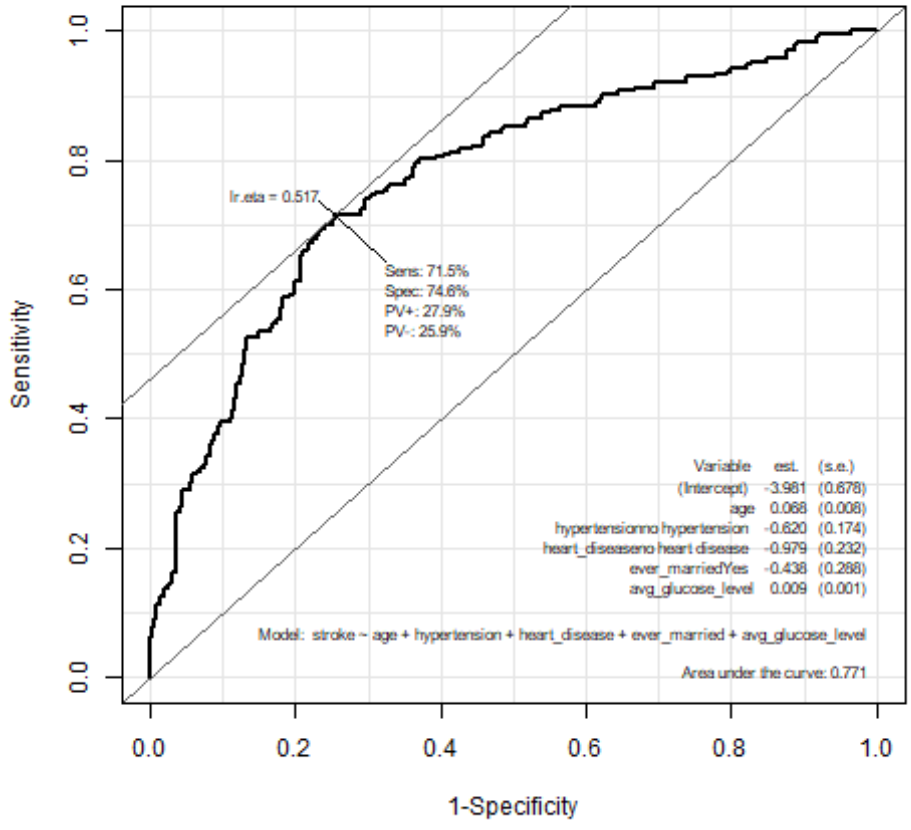

Figura 18: Gráfico ROC

El ROC (Receiver Operating Characteristic), es una curva y un estadístico que hace una relación entre los casos en que el modelo predice bien y las veces que el modelo predice mal, este estadístico para este caso es de 0.771, con este estadístico se puede decir que el modelo discrimina de forma adecuada los casos de especificidad y sensitividad respecto de los falsos positivos y los falsos negativos. Por ende el modelo resulta tener una buena capacidad predictiva adecuada.

Conclusiones

Para finalizar este trabajo, se barajaron numerosos modelos para terminar eligiendo uno con las variables significativas y quitando los datos que resultaron apoyarían al modelo, se termino con un modelo con una capacidad predicativa adecuada.
El modelo utilizó las variables de nivel de glucosa en la sangre, edad, presencia de hipertensión, presencia de problemas cardiacos y si alguna vez ha estado casado. Las variables que resultaron ser más significativas y que por ende resultan ser más de riesgo son la edad, los niveles de glucosa en la sangre y los problemas cardiacos, por lo que se puede decir que estos factores resultan ser factores de riesgo a tener en cuenta para la prevención de ACV.
Al contrario de como se encontró en la discusión bibliográfica, el tabaquismo, para este modelo no resultó ser un factor de riesgo, esto se puede explicar por la alta presencia de valores perdidos (que estaban expresados como categoría en los datos), y a pesar de que el modelo no lo consideró, es sabido que el factor de tabaquismo es determinante y factor de alto riesgo, que es necesario tener en consideración para futuros estudios.
Finalizando, en el transcurso del flujo de trabajo, encontramos indicios de que los datos no eran los idóneos para la resolución de la problemática, sin embargo, igualmente se logró obtener un modelo que dilucidara los factores de riesgo en la prevención de los ACV, por lo que el modelo sería adecuado para predecir y, sobretodo, para prevenir la ocurrencia de accidentes cerebrovasculares, sin descartar que puedan existir modelos mejor ajustados, con datos y variables distintas.

Plan de Trabajo

Se presentan las etapas y actividades para cada una de las semanas de ejecución del proyecto.

Carta Gantt	Mes 1				Mes 2			
Semana	1	2	3	4	1	2	3	4
Manipulación Computacional								
Visualización de los datos								
Presentación avance 1								
Avance 2								
Presentación avance 2								
Ejecución proyecto								
Presentación del final								

Referencias

Coalition, O. A. (2015). Entendiendo la prediabetes y el exceso de peso.

Durbán, M. (2014). *Modelos lineales generalizados.*

Escalante, R., Lourido, M., Melcón, C., y Curatolo, L. (2003). Accidente cerebrovascular en la policlínica bancaria: Registro de 1699 eventos consecutivos. *Revista Neurológica Argentina*, *28*(2), 91–95.

Fedesoriano. (26-01-2021). Stroke prediction dataset.

Ibañez, A., Rodrigo, A., Mercado, B., Pamela, J., Molina, M., Daniela, A., ... others (2009). Factores asociados al accidente cerebrovascular en usuarios hospitalizados en el servicio de medicina del hchm de chillán año 2008.

James W. Hardin, J. M. H. (2018). *Generalized linear models and extensions* (Fourth. ed.).

P. McCullagh, J. A. N. (1989). *Generalized linear models.* (Second. ed.).

Vespucio, C. (15-11-2021). Imc: Calculadora Índice de masa corporal.