Prueba técnica: Analista de analítica e inteligencia artificial
Autora: Laura Camila Ballesteros

Documentación de la prueba técnica

Primer objetivo

El siguiente documento pretende documentar la prueba técnica que se llevo a cabo en el cuaderno de python llamado "prueba.ipynb".

El primer objetivo es analizar nueve csv presentes y generar una estructura de datos que relacione una tabla de países con los datos de cada uno de los indicadores de los diferentes archivos.

Para el cubrimiento de este objetivo, se leyeron todos los archivos en python, se exploraron las variables, se exploró si existían registros duplicados o nulos y si se debía realizar algún preprocesamiento a los datos.

La estructura de los nueve archivos csv se pudo agrupar en tres categorias, en tres archivos que son: "maternalMortalityRatio", "roadTrafficDeaths", "incedenceOfTuberculosis", estos archivos tienen una columna que hace referencia al nombre completo del indicador, al país, el año en el que se mide.

En el codigo se agrega una nueva columna que tiene el nombre del indicador y la medición de ese indicador, luego se usa el método concat outer de pandas, que fusiona estos archivos a tráves de el país y el año. Este método hace que si no se encuentra un registro con el mismo país y año se rellene con nans. Luego se crea una variable "Dim1" y se le asigna el valor de "Both sexes" ya que los indicadores no se clasificaron de acuerdo a sexos para estos archivos.

El siguiente grupo de archivos son: "30-70cancerChdEtc", "alcoholSubstanceAbuse", "crudeSuicideRates", "maternalMortalityRatio", "tobaccoAge15", La diferencia de este tipo de archivo respecto a los anteriores es la presencia de la variable "Dim1" que tiene información del sexo, por lo que estos archivos se unen mediante la función creada de python "concatenate_dfs_in_sequence" que itera sobre dos dataframes y los concatena por pares. Las claves a través de las cuales se unen son, país, género, año.

Solo queda pendiente un único archivo "airPollutionDeathRate" que también se concatena en el punto anterior pero añade una variable adicional ("Dim2") para el indicador: Tasa de mortalidad atribuible a la contaminación del aire ambiental y doméstica. "Dim2" indica la causa de esa tasa de mortalidad.

Como medida de preprocesamiento, aquellos datos que contienen tasas o odds se transforman con la función creada en python "split_interval_median" que recopila el intervalo de confianza en una columna y el valor esperado de ese intervalo en otra ecuación.

De esta manera se logra la estructura de datos presente en la figura 1, donde se recopila la información por país, año, genero, causa y todos los indicadores.

	Location	Period	Gender	Probability of dying 30-70	Crude suicide rates per 100 000	Infant mortality rate	Point estimate infant mortality rate	Prevalence of current tobacco smoking (15+)	Maternal mortality ratio per 100 000	Point estimate maternal mortality ratio per 100 000	Incidence of tuberculosis per 100 000	Point estimate incidence of tuberculosis	Road traffic death rate (per 100 000	Alcohol per capita (15+) consumption
0	Afghanistan	2016	Both sexes	29.8	0.0	44.36-58.38	51.32	NaN	457-1040	673.0	122-270	189.0	15.1	NaN
1	Afghanistan	2016	Both sexes	29.8	0.0	44.36-58.38	51.32	NaN	457-1040	673.0	122-270	189.0	15.1	NaN
2	Afghanistan	2016	Both sexes	29.8	0.0	44.36-58.38	51.32	NaN	457-1040	673.0	122-270	189.0	15.1	NaN
3	Afghanistan	2016	Both sexes	29.8	0.0	44.36-58.38	51.32	NaN	457-1040	673.0	122-270	189.0	15.1	NaN

Figure 1: Estructura de datos por países con indicadores

También se estructuró un diccionario con las variables que tiene la estructura y sus tipos. Este diccionario se muestra en la figura 2.

VARIABLE	DESCRIPTION	TYPE
Location	Country	object
Period	Year of measurement of the indicator	int64
Gender	Gender: female, both sexes, male for the indicator	object
Probability of dying 30-70	Probability (%) of dying between age 30 and exact age 70 from any of cardiovascular disease, cancer, diabetes, or chronic respiratory disease	float64
Crude suicide rates per 100 000	Crude suicide rates (per 100 000 population)	float64
Infant mortality rate	Infant mortality rate (probability of dying between birth and age 1 per 1000 live births) Confidence interval	object
Median infant mortality rate	Infant mortality rate (probability of dying between birth and age 1 per 1000 live births) Median of the confidence interval	float64
Prevalence of current tobacco smoking (15+)	Age-standardized prevalence of current tobacco smoking among persons aged 15 years and older	float64
Maternal mortality ratio per 100 000	Maternal mortality ratio (per 100 000 live births) Confidence interval	object
Median maternal mortality ratio per 100 000	Maternal mortality ratio (per 100 000 live births) Median of the confidence interval	float64
Incidence of tuberculosis per 100 000	Incidence of tuberculosis (per 100 000 population per year) Confidence interval	object
Median incidence of tuberculosis	Incidence of tuberculosis (per 100 000 population per year) Median of the confidence interval	float64
Road traffic death rate (per 100 000	Estimated road traffic death rate (per 100 000 population)	float64
Alcohol per capita (15+) consumption	Total (recorded+unrecorded) alcohol per capita (15+) consumption	float64
Pollution death rate per 100 000 standarized	Ambient and household air pollution attributable death rate (per 100 000 population, age-standardized) Confidence interval	object
Median pollution death rate per 100 000 standarized	Ambient and household air pollution attributable death rate (per 100 000 population, age-standardized) Median of the confidence interval	float64
Pollution death rate per 100 000	Ambient and household air pollution attributable death rate (per 100 000 population) Confidence interval	object
Median pollution death rate per 100 000	Ambient and household air pollution attributable death rate (per 100 000 population) Median of the confidence interval	float64

Figure 2: Diccionario de variables

Segundo objetivo

Como segundo objetivo se tiene realizar un análisis descriptivo (univariado y multivariado) de las variables presentadas, interpretar los resultados y generar hipótesis o conclusiones a partir de los mismos.

Para el cumplimiento de este objetivo se propone cubrir cuatro indicadores y realizarles algunos análisis para aplicar de manera posterior pruebas de hipótesis para validar estos resultados.

Probability of dying between age 30 and exact age 70 from any of cardiovascular disease, cancer, diabetes, or chronic respiratory disease

Para este primer indicador se sacan los estadísticos descriptivos agrupados por media para los diez países donde hay mayor probabilidad de morir entre los 30 y 70 años por alguna enfermedad cardiovascular, cancer, diabetes o enfermedad respiratoria crónica. Los estadísticos se muestran en la figura 3.

Location	count	mean	std	min	25%	50%	75%	max
Kazakhstan	15.0	33.733333	10.478663	18.5	27.10	32.9	39.05	51.7
Mongolia	15.0	33.520000	7.332433	21.9	28.95	33.3	38.95	44.5
Fiji	15.0	33.113333	6.016509	24.0	29.30	32.9	36.90	42.5
Sierra Leone	15.0	32.540000	2.980364	27.6	30.25	32.6	34.50	36.9
Turkmenistan	15.0	32.286667	6.197911	22.9	28.05	31.0	37.10	42.6
Afghanistan	15.0	31.953333	2.788206	27.7	29.80	31.8	34.10	36.6
Yemen	15.0	31.813333	2.383235	28.2	30.30	31.9	33.25	35.7
Russian Federation	15.0	31.700000	11.295258	16.1	24.45	30.4	37.15	51.4
Guyana	15.0	31.286667	1.399932	28.8	30.55	31.3	32.35	33.7
Ukraine	15.0	30.500000	10.003999	16.2	24.00	29.3	36.10	48.4

Figure 3: 10 países con mayor probabilidad de morir entre los 30 y 70 años por alguna enfermedad cardiovascular, cancer, diabetes o enfermedad respiratoria crónica

Se puede observar que la mayor probabilidad de morir entre esas edades se encuentra en Kazakhstan con una probabilidad del 33.73% en el país de Kazakhstan y que los 10 países donde existe una probabilidad más elevada no existe una diferencia significativa entre sus medias.

La gráfica de barras que muestra este comportamiento se puede observar en la figura 4.

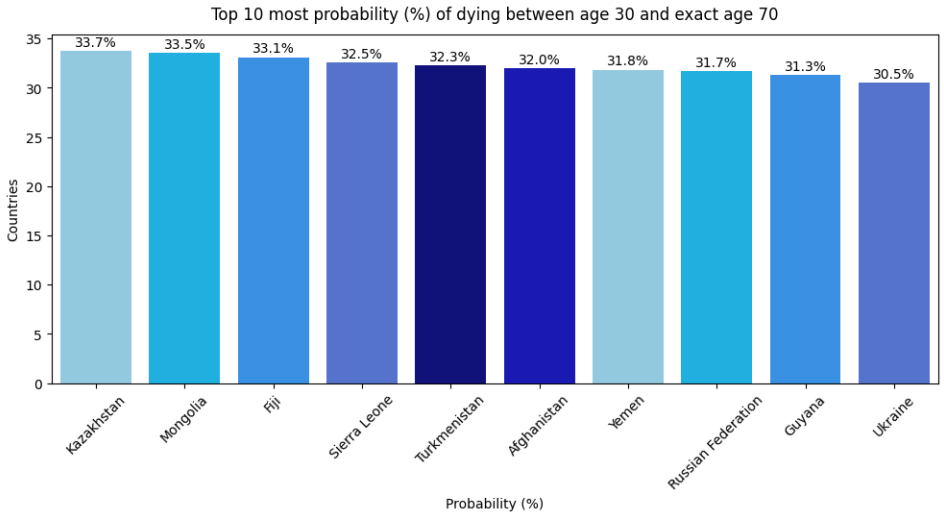

Figure 4: 10 países con mayor probabilidad de morir entre los 30 y 70 años por alguna enfermedad cardiovascular, cancer, diabetes o enfermedad respiratoria crónica

Además se realiza un análisis de los países donde existe una probabilidad menor de que suceda esto. El país donde se presenta una menor probabilidad es Japón. El tener los datos desagregados permitiría observar las características de los individuos de Japón que refleja esa probabilidad. La información suministrada se encuentre en la figura 5.

Location	count	mean	std	min	25%	50%	75%	max
Republic of Korea	15.0	11.386667	5.128334	4.7	8.05	10.6	14.15	22.9
Israel	15.0	11.386667	2.959698	7.5	9.65	11.6	12.45	17.8
Spain	15.0	11.333333	4.025041	6.4	7.65	10.7	13.75	19.1
Norway	15.0	11.246667	2.772278	7.6	9.25	10.9	12.80	17.6
Italy	15.0	10.993333	3.100338	7.2	8.75	10.3	12.85	17.9
Sweden	15.0	10.773333	2.275166	7.6	9.20	10.6	11.95	15.8
Iceland	15.0	10.660000	2.252554	7.7	9.20	9.9	11.80	15.4
Australia	15.0	10.560000	2.522697	7.2	8.95	10.0	11.75	16.1
Switzerland	15.0	10.133333	2.798129	6.6	8.20	9.7	11.50	16.4
Japan	15.0	9.693333	3.025479	5.7	7.20	9.5	11.45	15.5

Figure 5: 10 países con menor probabilidad de morir entre los 30 y 70 años por alguna enfermedad cardiovascular, cancer, diabetes o enfermedad respiratoria crónica

Para este indicador se agrupan los estadísticos descriptivos por país y genero, para considerar que diferencias existen. Los resultados se muestran en la figura 6.

Location	Dim1	count	mean	std	min	25%	50%	75%	max
Fiji	Both sexes	5.0	33.26	2.673574	30.6	30.8	32.9	35.6	36.4
	Female	5.0	26.56	2.680112	24.0	24.2	26.0	28.5	30.1
	Male	5.0	39.52	2.710535	36.8	37.0	39.2	42.1	42.5
Kazakhstan	Both sexes	5.0	33.28	5.714630	26.8	28.6	32.9	39.0	39.1
	Female	5.0	23.28	4.229303	18.5	19.8	23.0	27.4	27.7
	Male	5.0	44.64	6.944278	36.8	38.9	44.2	51.6	51.7
Mongolia	Both sexes	5.0	33.48	3.725185	30.2	30.4	32.3	35.6	38.9
	Female	5.0	25.92	4.734131	21.9	22.2	24.5	27.7	33.3
	Male	5.0	41.16	2.592875	38.8	39.0	40.2	43.3	44.5
Sierra Leone	Both sexes	5.0	32.58	2.752635	30.0	30.5	31.8	33.9	36.7
	Female	5.0	34.02	1.785217	32.2	32.6	33.7	35.1	36.5
	Male	5.0	31.02	3.833667	27.6	28.2	29.7	32.7	36.9
Turkmenistan	Both sexes	5.0	32.14	2.553037	29.5	30.6	31.0	34.0	35.6
	Female	5.0	25.56	2.438852	22.9	23.9	24.9	27.3	28.8
	Male	5.0	39.16	2.518531	36.7	37.5	38.0	41.0	42.6

Figure 6: 10 países con menor probabilidad de morir entre los 30 y 70 años por alguna enfermedad cardiovascular, cancer, diabetes o enfermedad respiratoria crónica de acuerdo al género

Se puede observar que existen diferencias notables en algunos países en la probabilidad de morir por las causas mencionadas entre mujeres y hombres como en Mongolia, en otros países como Sierra Leona no existe una diferencia de 10%.

Por último se realiza un histograma para observar el comportamiento de la distribución según el género para este indicador. Los resultados se muestran en la figura 7.

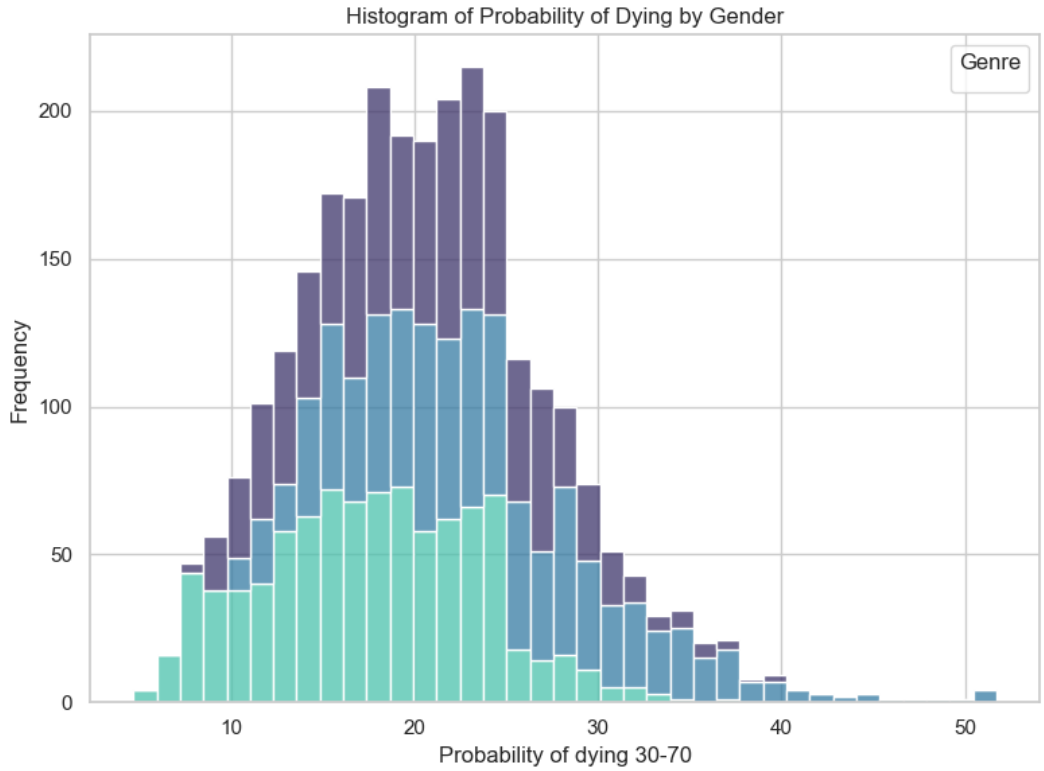

Figure 7: Histogramas probabilidad de morir entre los 30 y 70 años por alguna enfermedad cardiovascular, cancer, diabetes o enfermedad respiratoria crónica de acuerdo al género

Los análisis se realizan para los indicadores: "Consumo total (registrado+no registrado) de alcohol per cápita (15+)", "Tasas brutas de suicidio por 100 000 habitantes", "Prevalencia estandarizada por edad del tabaquismo actual entre personas de 15 años o más" siguiendo la misma secuencia, los pasos realizados y las conclusiones se encuentran en el notebook.

Tercer objetivo

Como tener objetivo se tiene plantear hipótesis referente a lo observado en el análisis.

Para cumplir este objetivo se realizaron cuatro hipótesis, dos referentes al indicador de probabilidad de morir entre los 30 y los 70 años exactos por cualquier enfermedad cardiovascular, cáncer, diabetes o enfermedad respiratoria crónica y se guiarón por la pregunta propuesta de: ¿Varía la probabilidad de mortalidad entre las edades de 30 y 70 años debido a enfermedades en función del país?

Por lo que se proponen inicialmente la hipótesis nula y alternativa que permitirán concluir acerca de la pregunta.

Hipotesis nula: No existe diferencia entre la probabilidad de morir entre 30 y 70 años por enfermedad cardiovascular, cancer, diabetes, enfermedad respiratoria crónica entre los individuos de Kazakhstan respecto a Mongolia, Fiji, Sierra Leona.

Hipotesis alternativa: Existe diferencia entre la probabilidad de morir entre 30 y 70 años por enfermedad cardiovascular, cancer, diabetes, enfermedad respiratoria crónica entre los individuos de Kazakhstan respecto a Mongolia, Fiji, Sierra Leona.

Para probar las hipótesis debe validar si la distribución de probabilidad normal, para esto se aplica el test de Shapiro Wilk.

El resultado de la prueba muestra que no se trata de una distribución normal por lo que se usan test no paramétricos, la variable por la que se agrupan es independiente (Locación) y se tienen más de dos grupos de comparación se aplica, características necesarias para aplicar el test Kruskall-Wallis.

Los resultados del test pueden verse en la figura 8.

```
Comparación Kazakhstan vs. Mongolia
Estadística de prueba de Kruskal-Wallis: 0.01549421193232426
Valor p: 0.9009386010250762
No se puede rechazar la hipótesis nula: no hay evidencia suficiente para concluir que hay diferencias significativas.

Comparación Kazakhstan vs. Sierra Leone
Estadística de prueba de Kruskal-Wallis: 0.021075268817199344
Valor p: 0.8845740956336217
No se puede rechazar la hipótesis nula: no hay evidencia suficiente para concluir que hay diferencias significativas.

Comparación Kazakhstan vs. Fiji
Estadística de prueba de Kruskal-Wallis: 0.0004302989836037336
Valor p: 0.9834501494461406
No se puede rechazar la hipótesis nula: no hay evidencia suficiente para concluir que hay diferencias significativas.
```

Figure 8: Aplicación de test Kruskall-Wallis

Estos resultados muestran que como los valores de la prueba no son superiores a 0.05 esta falla y en consecuencia no se puede rechazar la hipótesis nula. Esto implica que para los países seleccionados no existe diferencia probabilística por pertenecer a un país de incrementar esa posibilidad. Esta conclusión no aplica para el resto de países.

Luego se plantearon las siguientes hipótesis para el mismo indicador por sexo Hipotesis nula: No existe diferencia entre la probabilidad de morir entre 30 y 70 años por enfermedad cardiovascular, cancer, diabetes, enfermedad respiratoria crónica entre los individuos de Kazakhstan del sexo femenino y masculino.

Hipotesis alternativa: Existe diferencia entre la probabilidad de morir entre 30 y 70 años por enfermedad cardiovascular, cancer, diabetes, enfermedad respiratoria crónica entre los individuos de Kazakhstan del sexo femenino y masculino.

De los razonamientos anteriores se concluye que se debe aplicar test no parámetricos. Y como se tiene una comparación solo con dos grupos independientes se aplica el test Mann Withney U . Los resultados se muestran en la figura 9.

```
Comparación Male vs. Female in Kazakhstan
Estadística de prueba de Mann-Whitney-U: 25.0
Valor p: 0.007936507936507936
Se rechaza la hipótesis nula
```

Figure 9: Aplicación de test Mann Withney U

Se puede concluir que para el país escogido si existe una diferencia probabilistica de morir asociada al género.

Además se realizaron hipótesis para el indicador de: Prevalencia estandarizada por edad del tabaquismo actual entre personas de 15 años o más. Las hipótesis son:

Hipotesis nula: No existe diferencia entre la prevalencia estandarizada por edad del tabaquismo actual entre personas de 15 años o más entre los individuos de Myanmar respecto a Chile, Lao People's Democratic Republic.

Hipotesis alternativa: Existe diferencia entre la prevalencia estandarizada por edad del tabaquismo actual entre personas de 15 años o más entre los individuos de Myanmar respecto a Chile, Lao People's Democratic Republic.

Para probar las hipótesis debe validar si la distribución de probabilidad normal, para esto se aplica el test de Shapiro Wilk. Los resultados de la aplicación de esta prueba se puede ver en la figura 10.

```
Estadística de prueba: 0.9680908918380737
Valor p: 5.315137525243037e-29
Los datos no siguen una distribución normal.
```

Figure 10: Aplicación de test Shapiro wilk

Se puede observar que los datos no siguen una distribución normal y como se los grupos con los que se comparan son dos, se usa la prueba Mann Withney U. Los resultados pueden observarse en la figura 11.

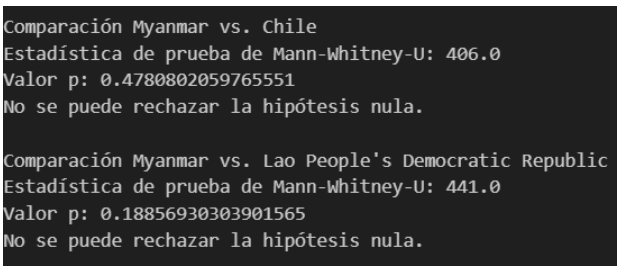

Figure 11: Aplicación de test Mann Withney U

Para el caso de estas hipótesis se puede observar que si existe una diferencia. Los pasos seguidos en el proceso de inferencia estadística fueron guiados por el libro: Biostatistics Manual for Health Research

Cuarto objetivo

Como último objetivo se tiene la construcción de un modelo que permita determinar los factores que podrían explicar la aparición de los tipos de cáncer presentados en los datos.

La construcción del modelo inicia en los datos de "airPollutionDeathRate" que contiene la causa de muerte del indicador en la variable "Dim2" en una se encuentra por cáncer de pulmón. En el código de python se fija está condición y las filas donde se encuentra se pone el valor de 1 el resto de cero para aplicar un modelo de clasificación binaria que trasmita explicabilidad de los factores que influyen sobre el cáncer de pulmón.

Por esta razón y porque existen variables categóricas se emplea el modelo de regresión logística. Los resultados asociados y el código se muestra a continuación, fue desarrolado en R.

```
##
## Call:
## glm(formula = Target ~ `Probability of dying 30-70` + `Point estimate infant mortality rate` +
##     `Prevalence of current tobacco smoking (15+)` + `Point estimate maternal mortality ratio per 100
##     `Point estimate incidence of tuberculosis` + Gender + df_model$Location,
##     family = binomial, data = df_model)
##
## Coefficients:
##                                                   Estimate Std. Error
## (Intercept)                                     -1.386e+00  1.986e-01
## `Probability of dying 30-70`                    -7.419e-16  8.525e-03
## `Point estimate infant mortality rate`           8.442e-17  2.847e-03
## `Prevalence of current tobacco smoking (15+)`    8.385e-17  3.338e-03
## `Point estimate maternal mortality ratio per 100 000` -5.329e-17  4.166e-04
## `Point estimate incidence of tuberculosis`       9.054e-18  6.073e-04
## Gender                                          -4.468e-15  7.463e-02
## df_model$Location                                2.315e-17  9.351e-04
##                                                 z value Pr(>|z|)
## (Intercept)                                      -6.981 2.92e-12 ***
## `Probability of dying 30-70`                      0.000        1
## `Point estimate infant mortality rate`            0.000        1
## `Prevalence of current tobacco smoking (15+)`     0.000        1
## `Point estimate maternal mortality ratio per 100 000`   0.000        1
## `Point estimate incidence of tuberculosis`        0.000        1
## Gender                                            0.000        1
## df_model$Location                                 0.000        1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2747.2  on 2744  degrees of freedom
## Residual deviance: 2747.2  on 2737  degrees of freedom
```

```
## AIC: 2763.2
##
## Number of Fisher Scoring iterations: 4
```

Los resultados del modelo muestran que no hay significancia de los factores escogidos sobre la aparición del cancer, esto puede deberse a que un tamaño de muestra pequeño en relación con el número de variables predictoras.

Además en el cruce de las bases muchos de los indicadores para el año seleccionado no se miden.

La carpeta con nombre "Biostatistics-analisis" contiene una carpeta llamada archivos-r que contiene el archivo empleado para plantear el modelo. La carpeta "estructura-datos" contiene los datos que fueron el insumo del modelo y la estructura de datos planteada para el punto 1.