1.EXECUTIVE SUMMARY

El propósito del siguiente informe se basará en la creación de un modelo de regresión logística con los datos, que son una encuesta de condiciones de vida en 2016, realizada por el INE.

La cual, se realiza desde 2004. Basada en criterios armonizados para todos los países de la Unión Europea, su objetivo fundamental es disponer de una fuente de referencia sobre estadísticas comparativas de la distribución de ingresos y la exclusión social en el ámbito europeo.

La realización de la ECV permite poner a disposición de la Comisión Europea un instrumento estadístico de primer orden para el estudio de la pobreza y desigualdad, el seguimiento de la cohesión social en el territorio de su ámbito, el estudio de las necesidades de la población y del impacto de las políticas sociales y económicas sobre los hogares y las personas, así como para el diseño de nuevas políticas, de ahí su importancia.[[1]](#endnote-1)

Por tanto, lo que se buscará en el siguiente análisis será realizar un modelo de regresión logística con dichos datos, utilizando como variable explicada “Hogar en riesgo de pobreza”, que es de tipo categórico.

A través de estos modelos podemos utilizar para los mismos variables cualitativas o categóricas como es el caso de nuestra variable, fragmentaremos el total de las informaciones en 60% que será la parte de entrenamiento, y un 40% lo de dedicaremos a la parte de testeo.

Después de esto realizaremos el test ANOVA, como sus propias siglas indican es un análisis de la varianza, y por tanto nos aporta la información necesaria para conocer que variables son estadísticamente significativas, AyudaFamilias (\*), VacacionesOutdoor (\*\*\*), CapacidadAfrontar (\*\*\*), LlegarFinMes (\*\*\*), Miembros (\*), HogaresSemanales (\*\*\*) y finalmente ActMayor (\*).

Con toda esta información podremos realizar el Pseudo R2 de McFadden (1979)[[2]](#endnote-2), el cual busca una reducción proporcional en la varianza del error. En nuestro análisis nos arroja un resultado del 37.89%.

Recordemos que un modelo de clasificación es aquel capaz de predecir a qué clase va a pertenecer una nueva instancia, basándose en lo aprendido en instancias anteriores, por tanto, realizaremos una matriz de confusión para observar si se han clasificado bien las variables o no si existen falsos positivos o falsos negativos. Y finalmente para evaluar este modelo podríamos simplemente calcular su **precisión** (**“accuracy”**), como la proporción entre las predicciones correctas que ha hecho el modelo y el total de predicciones.[[3]](#endnote-3)

2.INTRODUCCION

EI modelo de regresión logística permite, consecuentemente, relacionar una variable dependiente dicotómica con una o más variables independientes, las cuales pueden ser dicotómicas, politómicas o continuas.

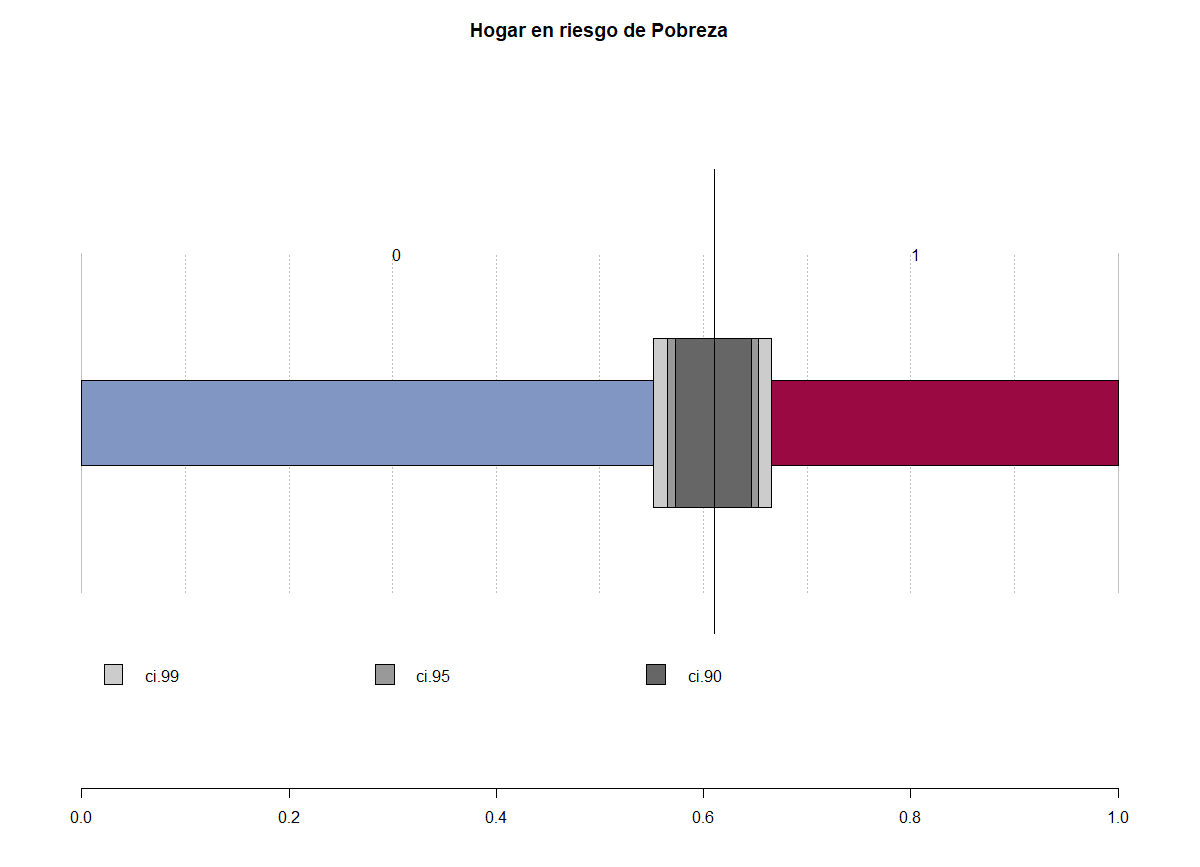
Este forma parte del aprendizaje supervisado (Machine Learning) necesitaremos una parte de prueba y otra de conjunto de pruebas. Primero, entrenamos al clasificador en el conjunto de entrenamiento, en nuestro caso (0.6) será el tamaño de nuestra muestra de entrenamiento como hemos indicado anteriormente.

Primeramente, realizamos el análisis exploratorio de los datos, el cual, para realizar este modelo no nos ha llevado demasiado tiempo, hemos analizado las variables que posee nuestra base de datos, desde la estructura, la dimensión…

Nuestra base de datos se compone de 477 observaciones distribuidas en 18 variables, de las cuales excluiremos la mayoría, las que no nos aporten información o que nos aporten cierta información que nos pueda distorsionar nuestro análisis.

Las variables que hemos eliminado han sido “El id del Hogar”, “Si tienen o no TVcolor”, “La Renta Total del ejercicio anterior”, “La región”, “Personas que tienen menos de 16 y reciben renta” y “Sexo mayor”, hemos reducido seis variables.

Más tarde depuramos cada variable, es decir, las métricas las convertimos en numéricas, las categóricas las convertimos en tipo factor es decir, cero y uno en función de sus valores, en concreto “VacacionesOutdoor”, “CapacidadAfrontar”, “Ordenador”, “LlegarFinMes”, “RegimenTenencia”, “SexoMayor” y “ActMayor”.

Observamos como de nivelada está nuestra variable dependiente, en el siguiente gráfico.

Se observa que se encuentra un poco desnivelada, simplemente es un análisis exploratorio de los datos.

3.MODELO DE REGRESIÓN LOGÍSTICA

Realizaremos un modelo de regresión logística sobre “Hogar en riesgo de pobreza”, para más tarde intentar predecir si el hogar está en riesgo de pobreza o no(en función de nuestra variable dependiente, la cual es dicotómica, si el hogar está en riesgo de pobreza o no) con la información del resto de variables.

Primeramente, definimos la cantidad de datos que necesitamos tanto la parte de entrenamiento, que el 60% de los datos,286 observaciones y la parte de test, la cual será el 40%, es decir, 191 observaciones.

Una vez creado el modelo GLM, con distribución binomial debido al carácter dicotómico de nuestra variable y analizándolo con la función summary, realizamos el test ANOVA, el cual nos arroja que las variables que son estadísticamente significativas son las siguientes AyudaFamilias(\*), VacacionesOutdoor(\*\*\*), CapacidadAfrontar(\*\*\*), LlegarFinMes(\*\*\*), Miembros(\*), HogaresSemanales(\*\*\*), ActMayor(\*).

Ahora analizaremos la bondad de ajuste através del pseudo de McFadden, nos arroja un resultado de 0.3789. Este estadístico en palabras de Mcfadden un número entre 0.2 y 0.4 es un resultado excelente, por tanto, podemos deducir que nuestro modelo muestra excelentes resultados, por ahora[[4]](#endnote-4).

En este momento se realiza la predicción sobre la parte de testeo , y establecemos que los valores superiores a 0.68 le daremos el valor 1 y si son menores de 0.68 un 0.

Por tanto una vez realizada la predicción, comparamos los resultados de nuestra predicción con la parte que teníamos, que es lo que se conoce como matriz de confusión, definida anteriormente y nos muestra los siguientes resultados.

|  |  |  |
| --- | --- | --- |
|  | Predicted |  |
| Actual | 0 | 1 |
| 0 | 102 | 7 |
| 1 | 42 | 40 |

Por tanto lo que nos muestra la matriz de confusión es que se han predicho bien 142 observaciones y 49 se han predicho mal. Por último mediremos el “acccuracy” de nuestro modelo, el cual, es la proporción de las predicciones generadas por el modelo, será el resultado de restarle 1 a al ratio de error de la unidad, en definitiva es la proporción de aciertos en la clasificación dada.

De manera que cuanto mayor sea la dispersión o varianza menor será nuestra precisión, compara el número de predicciones correctas y las no correctas y sacan un porcentaje, en nuestro modelo el porcentaje arrojado es de 74.34%.

4.Conclusiones

Como podemos observar en nuestro informe, hemos realizado un modelo de regresión logística, con los datos de una encuesta de condiciones de vida en 2016, realizada por el INE, los resultados, utilizando como variable dependiente “Hogar en riesgo de pobreza”, y 12 variables independientes que son las que explican la variable dependiente.

Se han realizado pruebas previas como el test ANOVA, y el test Mcfadden, una vez realizada tanto la limpieza como la depuración de los datos que nos arrojaban resultados correctos, que nos permitían realizar el modelo de regresión logística, y que mas tarde realizando la predicción de si un hogar estará en riesgo de pobreza obtenemos buenos resultados debido a que obtenemos un accuracy del 74,8%.

1. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\_C&cid=1254736176807&menu=ultiDatos&idp=1254735976608 [↑](#endnote-ref-1)
2. https://statisticalhorizons.com/r2logistic [↑](#endnote-ref-2)
3. https://data-speaks.luca-d3.com/2018/01/ML-a-tu-alcance-matriz-confusion.html [↑](#endnote-ref-3)
4. McFadden, D. (1974) "Análisis logit condicional del comportamiento de elección cualitativa". Pp. 105-142 en P. Zarembka (ed.), Frontiers in Econometrics. Prensa Académica. <http://eml.berkeley.edu/~mcfadden/travel.html> [↑](#endnote-ref-4)