Examen 3. Seminario de Estadística

Integrantes:

- Bonilla Alarcón Alejandro (420004750)
- Mejía Nájera Angel Josué (420003674)

Versión del examen: B

1	2

Tarea Examen 3

Versión B

Alejandro Bonilla Alarcón & Angel Josué Mejía Nájera

02-06-2023

Índice

1. Ejercicio 1

1.1. Datos

Se consideró la base de datos fat del paquete faraway, se usaron todas las variables, excepto siri, density y free. También se eliminó del análisis los casos con valores extraños en weight y height, así como valores cero en brozek. El objetivo del estudio es usar las variables clínicas observadas en los pacientes para predecir el porcentaje de grasa corporal en los hombres (var brozek).

Para calcular el poder de predicción se utiliza el método 5-CV

1.2. Modelos

En primer lugar se consideraron tres modelos lineales generalizados con función liga identidad y distribución Gausiana. El primer modelo tiene efectos principales, el segundo modelo tiene efectos principales e interacciones y el tercer modelo tiene efectos principales, interacciones y el cuadrado de las variables.

Luego, se añadio al proceso de entrenamiento la selección de variables usando el criterio BIC y el método de Lasso. Este nuevo proceso de entrenamiento se aplicó a los tres modelos anteriores.

Por último, se consideraron los tres modelos anteriores pero con función liga inversa y distribución Gamma. En cuanto al proceso de entrenamiento, se uso la selección de variables usando el método Lasso.

1.3. Resultados

En el Cuadro 1 se puede observar el poder predictivo de los modelos anteriormente descritos.

Cuadro 1: Poder predictivo de cada modelo, donde P indica componente principales, I indica interacciones y C indica variables al cuadrado. Por su parte, BIC y Lasso indican que método de selección de variables se utilizó.

	Poder.predictivo
Gaussian P	17.09
Gaussian P + I	56.64
Gaussian P + I + C	90.42
Gaussian P | Lasso	16.99
Gaussian P + I | Lasso	16.32
Gaussian P + I + C | Lasso	16.30
Gamma P | Lasso	32.96
Gamma P + I | Lasso	24.75
Gamma P + I + C | Lasso	24.16

Como se puede observar del Cuadro 1 el modelo con mejor poder predictivo es aquel con liga identidad, distribución Gausiana y componente lineal con efectos principales, interacciones y variables al cuadrado, y en el que se realiza selección de variables con el método Lasso.

Se observa que en los modelos sin selección de variables y con selección de variables usando el criterio BIC, el poder predictivo empeora conforme se vuelve más complejo el componente lineal. Por su parte, los modelos que usan la selección de variables con el método Lasso tienen un mejor poder predictivo conforme se vuelve más complejo el componente lineal. Esto puede deberse a que los modelos con componente lineal más complejo tienden a sobreajustar y no se desempeñan bien en nuevas observaciones. Sin embargo, con el método Lasso se tiene una buena selección de variables para poder predecir de buena forma las nuevas observaciones sin necesidad de tener un sobreajuste.

Finalmente, los modelos con liga identidad y distribucion Gausiana tuvieron un mejor poder predictivo que sus respectivos modelos con liga inversa y distribucion Gamma.

Las variables que más se seleccionaron en los modelos con el método Lasso se encuentran: age, height, abdom, writst, age:abdom, height:wrist.

El modelo con mejor poder predictivo tiene la siguiente forma:

$$E[Y] = \beta_0 + \beta_1 x_{abdom} + \beta_2 x_{height}^2 + \beta_3 x_{age} x_{adipos} + \beta_4 x_{age} x_{abdom} + \beta_5 x_{height} x_{wrist}$$

2. Ejercicio 2

2.1. Datos

La base de datos *PimaIndiansDiabetes*2 del paquete *mlbench* contiene información recopilada por el National Institute of Diabetes and Digestive and Kidney Diseases, la cual incluye datos de 768 pacientes. Sin embargo, se identificaron valores inusuales con registros de 0, los cuales se corrigieron y se convirtieron en valores NA (no disponibles). Debido a la presencia de NA's en los datos, se decidió eliminar todas aquellas observaciones que tuvieran al menos una variable con un registro NA. Como resultado, se redujo el conjunto de datos a un total de 392 observaciones, en comparación con las 768 observaciones originales. Dentro de este subconjunto, se encontraron 262 pacientes diagnosticados con diabetes y 130 pacientes sin esta condición.

El objetivo principal de este estudio es realizar una predicción óptima sobre la presencia o ausencia de diabetes en futuros pacientes. Para lograr esto, se utilizarán las 8 covariables clínicas presentadas en el estudio. Para determinar la regla de decisión óptima a utilizar, se hará uso de una selección de modelos de clasificación supervisada a través de la comparación de su poder predictivo. El preprocesamiento de los datos y la optimización de algunos hiperparámetros serán fundamentales para determinar la regla de decisión final. Además, se utilizará el criterio de máxima probabilidad para clasificar los datos, con corte en 0.5.

2.2. Análisis Descriptivo

Variable	Descripción
`pregnant`	Número de veces embarazada(Entero).
`glucose`	Concentración de glucosa en plasma(prueba de tolerancia a la glucosa)(Entero).
`pressure`	Tensión arterial diastólica(mm Hg)(Entero).
`triceps`	Espesor del pliegue cutáneo del tríceps(mm)(Entero).
`insulin`	Insulina sérica de 2 horas(mu U/ml)(Contínua).
`mass`	Indice de masa corporal (peso en kg/(altura en m)^2)(Contínua).
`pedigree`	Función pedigrí de la diabetes(Contínua).
`age`	Edad en años(Int).
`diabetes`	Variable binaria de dos niveles: (2 = Yes = Paciente con Diabetes, 1 = No = Paciente sin Diabetes).

Cuadro 2: Variables originales del DataFrame `PimaIndiansDiabetes2`, descripción y tipo de datos

A partir de la Figura 1 se observa que:

- El número de no diabéticos observados es casi el doble que el de diabéticos.
- Existen valores atípicos en el número de embarazos, llegando a presentar un paciente con 15 embarazos.
- Las personas con mayor numero de embarazos tienden a apresentar diabetes.
- Las personas con mayor edad presentan diabetes, mientras que los más jovenes no.
- Las personas que poseen diabetes concentran un mayor nivel de glucosa e insulina.

- Las variables pressure, insulin, mass y pedigreeno presentan muchas diferencias entre los grupos, por lo que se intuye, su poder predictivo sea bajo.

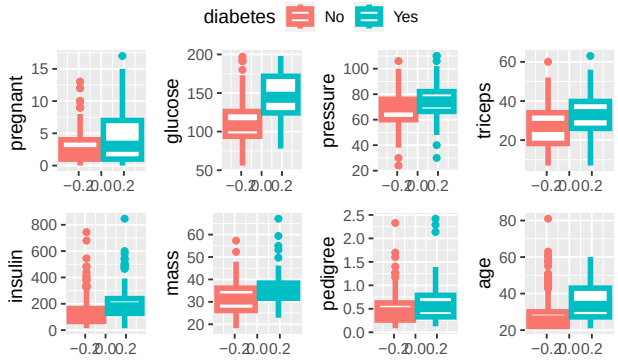

Figura 1: Conjunto de Boxplots por cada variable independiente y diferenciados según la presencia o ausencia de diabetes

A partir de la Figura 2 se observa que:

- En el Componente Principal 1, a medida que aumentan los valores de este componente, principalmente los valores de glucosa, edad, insulina, mass, triceps y pressure, también aumentan de manera positiva. Esto crea una partición de los datos, evidente en el 1° y 3° gráfico, entre aquellos que tienen diabetes y los que no la tienen. Los individuos con diabetes presentan valores más altos de glucosa, edad, insulina, etc., como era de esperar. Por otro lado, aquellos sin diabetes se encuentran en los valores negativos del Componente Principal 1, lo que indica que tienen niveles más bajos de las variables mencionadas anteriormente.
- En cuanto al Componente Principal 2, los datos se mezclan más y no hay una interpretación clara.
- En el Componente Principal 3, podemos observar que las variables más representativas son glucosa e insulina, y tienen una correlación negativa. Esto significa que a medida que disminuyen los valores de este componente, los valores de glucosa e insulina aumentan. Esto se corresponde con el hecho de que los pacientes sin diabetes se ubiquen en la parte positiva de dicho componente, lo que indica que tienen valores más bajos de insulina y glucosa. Sin embargo, esta conclusión no se puede generalizar debido a la dispersión de los datos para las personas con diabetes en este componente.

2.3. Algoritmos de clasificación

- 1. Regresión Logística (Solo efectos principales)
- 2. Regresión Logística (Efectos principales, interacciones y términos cuadrados)
- 3. Regresión Logística (Efectos principales + selección por pasos con criterio BIC)
- 4. Regresión Logística (Efectos principales, interacciones y términos cuadrados + selección por pasos con criterio BIC)
- 5. Regresión Logística (Efectos principales, interacciones y términos cuadrados + selección usando Lasso). Hiperparámetro tuneados: Lambda
- 6. Regresión Probit (Efectos principales, interacciones y términos cuadrados + selección por pasos con criterio BIC)
- 7. Método Naive
- 8. LDA contínuo (Análisis de Discriminante Lineal)
- 9. QDA contínuo (Análisis de Discriminante Cuadrático)
- 10. K vecinos mas cercanos. Hiperparámetro tuneados:K
- 11. Random Forest (200 arboles). Hiperparámetro tuneados: Mtry, NodeSize

Cuadro 3: Comparación de Modelos de Clasificación ordenados con base en el TCC, Tomando en cuenta el poder predictivo por clase e hiperparámetros tuneados.

Metodo	Especificidad	Sensibilidad	TCC	Tuneo
Modelo3	89.35	59.23	79.31	-
Modelo6	89.08	59.23	79.13	-
Modelo1	89.65	57.92	79.08	-
Modelo8	89.27	58.54	79.03	-
Modelo4	88.58	59.46	78.87	-
Modelo11	87.31	60.23	78.28	mtry, node_size
Modelo5	89.65	54.38	77.90	lambda
Modelo9	86.35	60.69	77.80	-
Modelo7	83.73	64.77	77.41	-
Modelo10	89.00	51.92	76.64	K
Modelo2	85.65	54.62	75.31	-

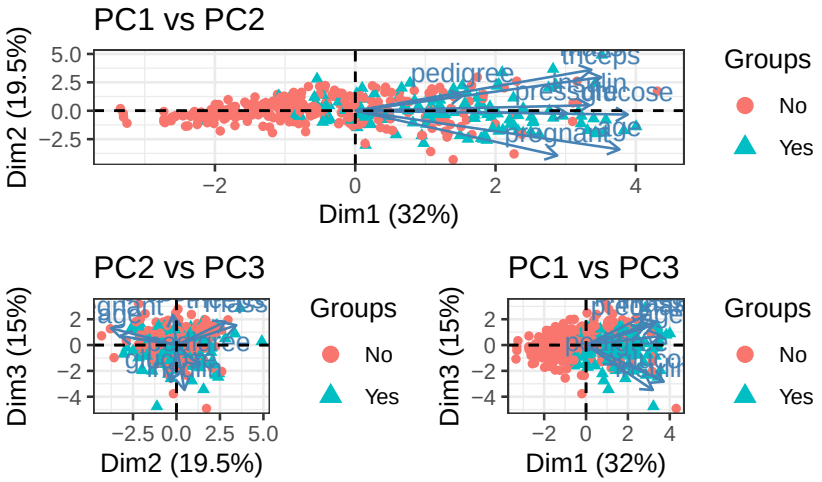

Figura 2: Gráfico de dispersión de los datos proyectados sobre los primeros 3 Componentes Principales. Cada observación se diferencía según la condición de diabetes.

2.4. Resultados

Para evaluar el poder predictivo de cada uno de los esquemas de clasificación presentados en 2.3 se utilizó el método de Repeated Holdout Method, donde se generaron 50 muestras aleatorias estratificadas según la condición de diabetes (B = 50), de la cual se obtuvieron los conjuntos para entrenar los modelos (Train) y para evaluar el poder predictivo (Test), en una poderación de 80 %-20 %, es decir, dada una muestra obtenida, 80 % se utilizó para el conjunto Train y el 20 %, para el conjunto Test.

En el Cuadro 3 se presentan los resultados obtenidos, tanto la tasa de Especificidad (Verdaderos negativos/Falsos negativos), la tasa de Sensibilidad (Verdaderos Positivos/Falsos Positivos) y la precisión o TCC, que mide el número de clasificaciones acertadas totales.

Se observa que el modelo con mejor poder predictivo fue el modelo 3, que con base en 2.3, corresponde a un modelo de regresión logística que solo toma en cuenta efectos principales y se le aplica una selección de variables vía Step usando como criterio el BIC. Los siguientes modelos con mayor desempeño y no tan alejados del modelo 3, fueron el 6 (Regresión Probit con efectos principales, interacciones, términos cuadráticos y selección step por BIC) y el 1 (Regresión Logít con solo efectos principales).

Cuadro 4: 10 Covariables Clínicas con mayor frecuencia en los ajustes simulados en los conjuntos Train

.	Freq
glucose	171
mass	158
age	119
pedigree	75
I(age^2)	74
I(pregnant^2)	67
I(glucose^2)	51
glucose:triceps	41
pregnant:insulin	39
I(insulin^2)	27

Cuadro 5: Coeficientes del modelo con mayor poder de predicción

	x
(Intercept)	-10.0920
glucose	0.0362
age	0.0530
mass	0.0744
pedigree	1.0871

El modelo peor clasificado fue el modelo 2, una Regresión Logit con Efectos principales, interacciones y términos cuadrados sin ningún tipo de selección de variables. Por otro lado, los modelos con hiperparámetros tuneados tuvieron un rendimiento medio. El modelo de K-vecinos más cercanos tuvo un rendimiento pésimo aún haciendo un escalado a las variables predictoras.

Nótese la gran diferencia de rendimiento entre el modelo con discriminante lineal(modelo 8) y el de discriminante cuadrático (modelo 9).

Podemos observar que, en general, los modelo presentan poca sensibilidad, es decir, dado que los pacientes presentan diabetes, nuestro modelo predice que no la tienen. Dicho error es preocupante y se podría realizar un ajuste para disminuir la tasa de especificidad a cambio de aumentar la sensibilidad.

Para la determinación de las covariables con mayor poder predictivo, se guardaron las covariables utilizadas en cada uno de los entrenamientos de aquellos modelos que utilizaron alguna selección de variables (Modelos 3,4,5 y 6). Así, en la tabla 4, se aprecia que las covariables más utilizadas y que por tanto poseen mayor poder predictivo son: “glucose”, “mass” y “age”, las cuales se corresponden con lo analizado en el apartado 2.2.

2.5. Regla final

El modelo final con mayor poder predictivo es es el modelo 3, el cual consiste en una regresión logística sobre los Efectos Principales más Selección por Pasos mediante criterio BIC. No se realizó ningún tuneo sobre hiperparámetros ni se escalaron las covariables.

Se presentan los coeficientes que determinan la regla final en la tabla 5