Aprendizaje estadístico automatizado

Profesor: Gonzalo Pérez de la Cruz.

Ayudantes: Leonardo Daniel de la Cruz Cuaxiloa, Fernando Raúl Garay Araujo.

Estudiante: Carlos Iván Canto Varela.
Número de cuenta: 315649888.
Correo: civancv@ciencias.unam.mx.

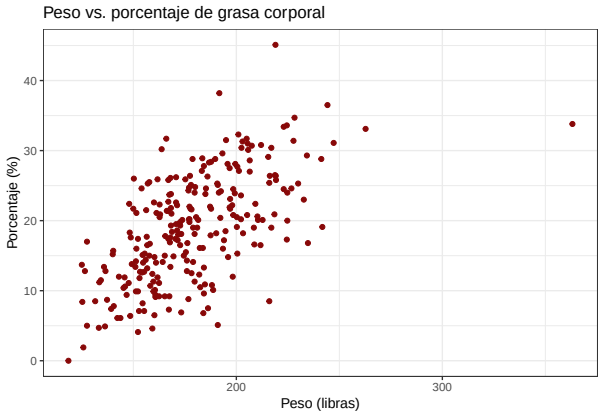
Examen 3. Predicción: versión A

1. Predicción en el caso continuo

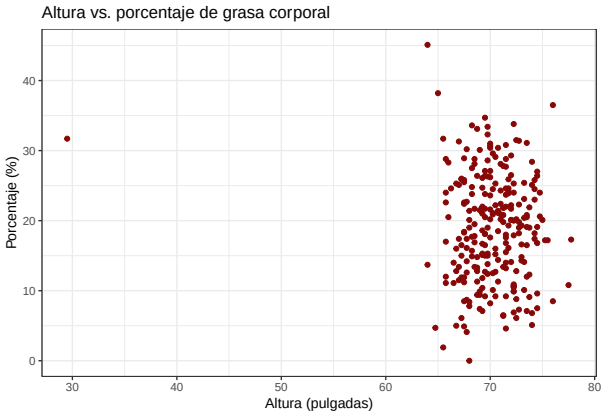
Se requiere predecir el porcentaje de grasa corporal: un valor continuo. Para esto, el método de evaluación del poder predictivo fue *repeated holdout* con 50 repeticiones. Sin embargo, los datos no se pueden tomar crudos.

Dado que la base de datos *fat* contiene algunas incongruencias, se procedió a eliminar las observaciones extrañas; estas incluyeron a gente con un porcentaje de grasa corporal nulo. Tampoco se usaron las variables *siri, density* y *free*. Con los gráficos de dispersión bidimensionales se pueden identificar otros valores atípicos:

(a) Porcentaje de grasa corporal



(b) Altura



Diagramas de dispersión por covariable

De aquí que se desechen a aquellos con alturas menores a 30 pulgadas o pesos mayores a 350 libras. Posteriormente, la partición de los datos a utilizar fue fijada en 80-20 para entrenamiento y prueba y estratificada por cuartiles para una mejor representación y —consiguientemente— un cálculo del poder predictivo más certero.

i) MLG gaussiano con liga identidad

Un modelo linear generalizado gaussiano con liga identidad es equivalente a un modelo de regresión lineal múltiple, por lo que se utilizó el método "lm" para realizar estos ajustes. Se siguieron dos corrientes de acuerdo a las covariables tomadas en cuenta.

Modelo simple

En este caso, *brozek* fue descrito únicamente con los efectos principales de las covariables disponibles.

Modelo comprehensivo

Al modelo anterior se le añadieron las interacciones por pares y los cuadrados de todas las medidas para tener un total de 119 covariables.

ii) Selección de variables con el criterio BIC

Para este apartado se realizó la selección por pasos bidireccionales. Inició en el modelo simple (sólo efectos principales); el límite inferior fue uno con sólo el intercepto y el superior el comprehensivo (efectos principales, interacciones y cuadrados). Se llegó a un modelo con las variables *age, adipos, chest, abdom, hip, wrist,* hip^2 y *age:wrist*.

iii) Selección de variables por método lasso

Para aplicar el método lasso eficientemente se tuneó el valor de λ con 5-CV por medio de la función "cv.glmnet".

Modelo simple

Para el modelo simple se llegó a la conclusión que utilizar todo menos el peso, el índice de adiposidad y la medida de rodilla sería lo mejor.

Modelo comprehensivo

En este caso se prefirió la medida del abdomen, los interceptos de edad con pierna, tobillo y biceps, los de altura con cuello y muñeca y el de cuello con muñeca.

iv) MLG gamma de liga identidad y selección por pasos

De manera similar al inciso **ii** se realizó un modelo por pasos bidireccionales guiados por el criterio BIC. Partió en los efectos principales y se experimentó con modelos desde ninguna variable hasta el comprehensivo, esta vez con distribución gamma y liga identidad. La regla óptima se determinó con *height, neck, abdom, biceps, wrist,* age^2 y *height*biceps*.

iv) Resultados

Dados los modelos finales, los predictores que aparecen más veces en los métodos de selección son *age, wrist* y *height* en orden descendiente y al tomar en cuenta interacciones y cuadrados. En menor medida, se pueden incluir *abdom, neck* y *biceps*. Estas son las variables con mayor poder predictivo.

Esquema de entrenamiento			**MSE**
RLM	Simple		17.1932
	Comprehensivo		113.9358
	Seleccionado por pasos		18.1462
	Lasso	Simple	17.1
		Comprehensivo	17.0413
MLG gamma sel. por pasos			18.6737

Poder predictivo encontrado por esquema de entrenamiento

v) Conclusiones

Con un MSE de 17.04, se considera que el mejor esquema de entrenamiento es el de regresión lineal múltiple comprehensivo con selección de variables lasso tuneada en cuanto a λ con validación cruzada de 5 cinco. Para los datos disponibles, la regla final se determinó como:

$$\widehat{brozek} \approx -2.13_{*10} + 7.36_{*10^{-1}} abdom - 9.02_{*10^{-4}} age * thigh$$
$$+1.12_{*10^{-3}} age * ankle + 2.48_{*10^{-3}} age * biceps - 4.12_{*10^{-4}} height * neck$$
$$-1.49_{*10^{-2}} height * wrist - 1.45_{*10^{-2}} neck * wrist.$$

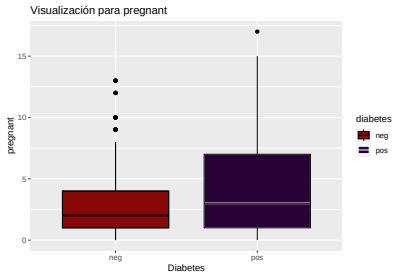
2. Clasificación supervisada

Nótese que el desenlace a predecir es binario: si el paciente tiene diabetes o no. Dicho esto, todas las variables son continuas por lo que se puede visualizar el panorama mediante diagramas de caja y aplicar componentes principales.

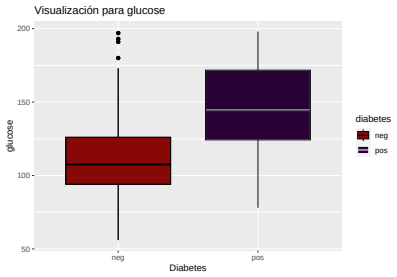
i) Análisis descriptivo

Una vista rápida de los predictores se presenta a continuación:

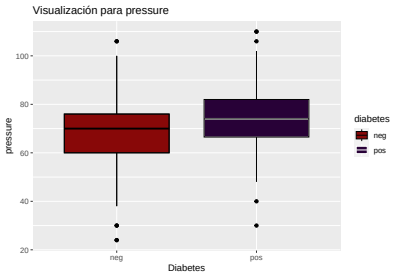
(c) Número de embarazos



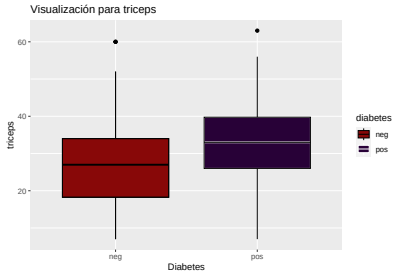
(d) Concentración de glucosa



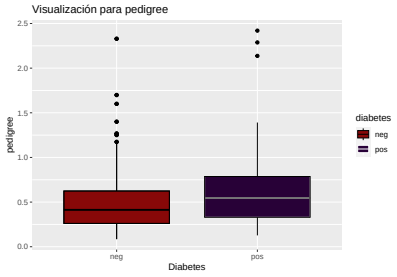
(e) Presión diastólica



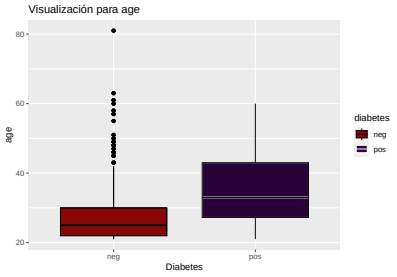
(f) Grosor de la piel del tríceps



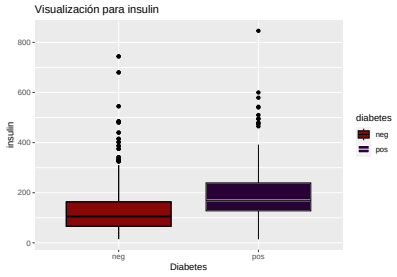
(g) Pedigrí de diabetes



(h) Edad



(i) Suero de insulina a dos horas



(j) Índice de masa corporal

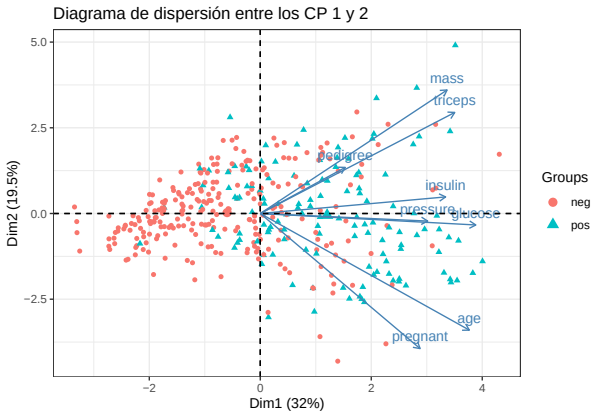
Visualizaciones de los predictores

Las medias tienen un contraste más aparente en ambos grupos para las variables de glucosa y edad. Esto podría indicar que estas variables tienen mayores papeles en la determinación de la diabetes que las demás. Por otra parte, los cuartiles que difieren en mayor parte para ambos grupos se encuentran en *pregnant, glucose, age* y *mass*.

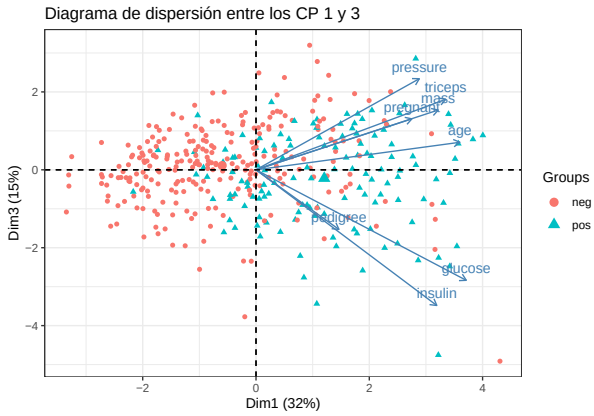
ii) Componentes principales

Un análisis deriva de los componentes principales, donde el total de la varianza explicada por los primeros tres CP es de 66.5 %. La primera CP se ve influenciada positivamente por la glucosa, edad y grosor de la piel del tríceps en mayor parte. La segunda ve cambios con el índice de masa corporal (+) y la edad (-); mientras que la tercera es afectada inversamente por la insulina y glucosa.

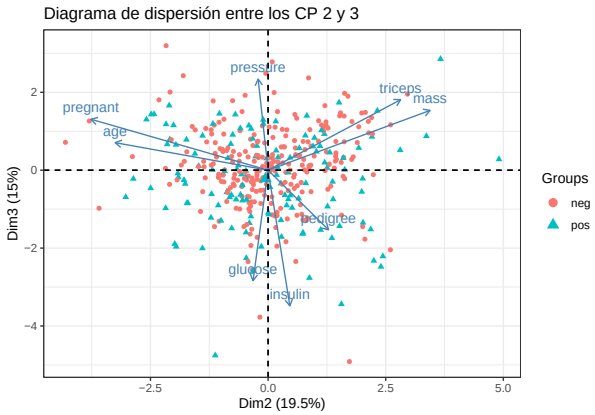
(k) Componentes uno y dos



(l) Componentes uno y tres



(m) Componentes dos y tres



Diagramas de dispersión entre los componentes principales

Debido a la presentación de los diagramas, se puede entender que la componente principal 1 determina en mejor medida a la segregación entre los individuos con y sin diabetes. Como en el análisis descriptivo, la glucosa y edad muestran más potencial de clasificación, por lo que se espera que estén presentes en los métodos a utilizar.

iii) MLG binario con liga logit

Se consideraron modelos lineales generalizados binarios con liga logit pero con distintas covariables.

Modelo simple

Las probabilidades de diabetes fueron descritas con los efectos principales disponibles sólamente.

Modelo comprehensivo

Mediante efectos principales, interacciones y cuadrados, un número mayor de predictores se utilizó en este modelo.

iv) Selección de variables por pasos y criterio BIC

Con una penalización logarítmica, la selección por pasos bidireccionales del método "stepAIC" inició en el modelo simple y experimentó entre un ajuste del mero intercepto y del comprehensivo. Las variables incluidas fueron *glucose, mass, pedigree, age* y age^2.

v) MLG con liga probit y selección lasso tuneada por 5-CV

Este modelo binario usó una liga probit y la selección de variables —que incluyó las posibilidades que el modelo comprehensivo ofrece— fue tuneada con *5-cross validation*. Se escogió la glucosa, IMC, pedigrí diabético, los cuadrados de glucosa, insulina y edad y los interceptos entre el número de embarazos con la insulina, pedigrí y medida del tríceps, entre glucosa y tríceps, entre presión diastólica y edad, entre insulina con pedigrí y edad y entre IMC y edad.

vi) *Naive classifier*, ADL, ADC y K-NN

Naive classifier, ADL y ADC

Aunque el método ingenuo se considera no paramétrico, el análisis de discriminante —tanto lineal como cuadrático— puede verificar los supuestos de normalidad para las variables continuas para argumentar la efectividad de la asignación; sin embargo, en este caso sólo se usó con el propósito de predicción.

K-NN

Se realizó con la versión para variables continuas. Nuevamente, la K del método fue tuneada por validación cruzada (5) y una malla de posibles valores para K, donde el mejor fue 16.

vii) *Random forest* con tuneo de "mtry"

Con una malla de valores y el método 5-CV, se encontró que el mejor valor de "mtry" es tres para esta versión de bosques aleatorios.

viii) Resultados

Los entrenamientos fueron evaluados de acuerdo a las métricas de precisión (*accuracy*), exhaustividad (*recall*) y especificidad (*specificity*).

Esquema de entrenamiento	**Precisión**	**Exhaustividad**	**Especificidad**
MLG binario logit simple	0.7757	0.5692	0.8784
MLG binario logit comprensivo	0.7323	0.5538	0.8211
MLG logit sel. por pasos	0.7833	0.6154	0.8668
MLG binario probit lasso	0.7935	0.5615	0.9087
Naive classifier	0.7578	0.6231	0.8248
ADL	0.7757	0.5615	0.8821
ADC	0.7783	0.6308	0.8515
K-NN	0.7733	0.5462	0.8858
Bosque aleatorio	0.7629	0.5769	0.8554

Poder de clasificación encontrado por esquema de entrenamiento

Para los modelos con selección de variables, el predictor que se presentó más veces —ya sea por si sólo, en cuadrado o en interacción— fue la edad, seguido por la glucosa y el pedigrí en misma proporción; esto da a entender que tienen un poder predictivo mayor sobre los demás.

En general, la clase de esquemas de entrenamiento por modelos lineales generalizados muestra valores más altos de precisión y especificidad, mientras que la no paramétrica (junto con *naive*) gana en cuanto a la exhaustividad. Se puede decir que la primera mencionada tiene mejor poder predictivo con base en las precisiones alcanzadas.

ix) Conclusiones

Con las métricas presentadas, el que se considera como mejor modelo es el modelo lineal generalizado binomial con liga probit que utilizó selección de variables por lasso tuneado por validación cruzada de 5 ya que tuvo la mejor precisión con 0.7935. Su expresión, con los datos presentes, es:

$$\begin{gathered}\Phi^{-1}(\mu) \approx -4.12 + 3.18_{*10^{-3}} glucose + 2.84_{*10^{-2}} mass + 1.13 pedigree + 6.24_{*10^{-5}} glucose^2 \\ -1.85_{*10^{-6}} insulin^2 - 2.07_{*10^{-6}} age^2 + 3.03_{*10^{-4}} pregnant * insulin \\ +7.32_{*10^{-3}} pregnant * pedigree + 8.95_{*10^{-5}} glucose * triceps + 3.66_{*10^{-5}} pressure * age \\ -2.77_{*10^{-3}} insulin * pedigree + 4.39_{*10^{-5}} insulin * age + 1.38_{*10^{-4}} mass * age\end{gathered}$$

para $\mu \equiv \mathrm{E}[diabetes]$ y Φ^{-1} la inversa de la distribución acumulada de la variable aleatoria normal estándar: función asociada a la liga probit.