Universidad Nacional Autónoma de México
Facultad de Ciencias

Examen 3 A

Realizado por

Cícero Álvarez Alicia Guadalupe 318010807
Hernández Alva Luis Ángel 315251674
Isunza Alvarez Marcos Guillermo 419002921
Regalado Urbina Brandon Imanol 317312878

Profesores

Gonzalo Pérez de la Cruz
Noe Eusebio Amador González
César Humberto Valle Márquez

Asignatura

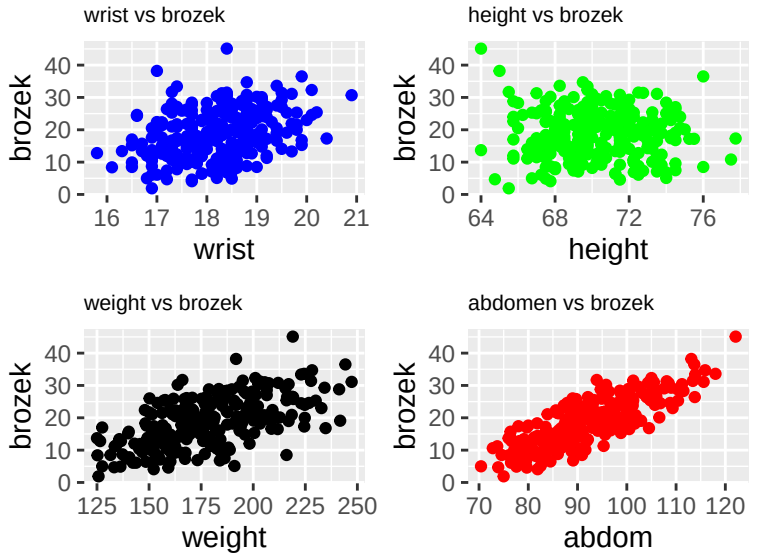
Seminario de Estadística I:

Aprendizaje Estadístico Automatizado

Lunes 11 de Diciembre de 2023

Prediccion del promedio de porcentaje de grasa corporal usando modelos lineales generalizados para datos continuos

Es de interes conocer cuales son las variables clínicas del conjunto de datos *fat* que predicen con presición el promedio de porcentaje de grasa corporal en hombres. Primeramente se realizó un análisis exploratorio de datos para determinar si existian valores extremos en los datos y se eliminaron aquellas observaciones que mostraban un peso superior a 250 lbs de la variable *weight*, una altura inferior a 60 pulgadas de la variable *height* o un valor de cero para la variable *brozek*. A continuación se muestran cuatro gráficas que muestran la relación entre distintas variables del conjunto de datos fat con la variable *brozek* después del procesamiento inicial [1].



En las gráficas presentadas anteriormente se puede observar que valores atípicos de height y weight han sido eliminados. Con el fin de prever el valor promedio del percentaje de grasa corporal en hombres se exploraron un total de 10 modelos lineales generalizados para datos continuos con distribución Gaussiana y función de enlace identidad. El objetivo es determinar su capacidad para predecir el promedio del porcentaje de grasa corporal en hombres, es decir determinar el poder predictivo de cada modelo usando las métricas de error cuadrático medio (MSE), media de la diferencia en valor absoluto (MAE) y el coeficiente de correlación entre y y $\hat{y}$. Para seleccionar las variables de los modelos, se usaron tres métodos diferentes de selección de variables, a saber, selección de variables stepwise usando el criterio BIC, selección de variables por el método Lasso y selección de variables por el método del mejor subconjunto. Los hallazgos se presentan resumidos en el siguiente **Cuadro 1**[2]:

Modelo	**MSE**	**MAE**	**Coeficiente de correlación**
Efectos principales	17.16308	3.410629	0.8631222
Interacciones	56.48963	5.46979	0.9177546
Variables al Cuadrado	18.56875	3.493047	0.8743873
EfectosPrincipales_BIC	17.04808	3.41743	0.8561274
Interacciones_BIC	71.25154	3.382046	0.8607871
VariablesalCuadrado_BIC	17.47172	3.445213	0.8619548
EfectosPrincipales_lasso	16.70562	3.383621	0.8568989
Interacciones_lasso	16.62958	3.382046	0.856508
VariablesalCuadrado_lasso	17.02291	3.408702	0.858146
MejorSubconjunto	17.17865	3.433325	0.8525693

Cuadro 1: Modelos ajustados con métricas MSE, MAE y coeficiente de correlación.

Tras analizar cuales son las variables que se incluyen con más frecuencia en aquellos modelos donde se implementó algún método de selección de variables, es decir, los últimos 7 modelos del **Cuadro 1**, se halló que la variable que más se repite es la variable *wrist* la cual se repite un total de 9 veces, seguida de la variable *abdom* la cual se repite un total de 8 veces, y después la variable *age* que se repite un total de 7 veces, *neck* aparece 5 veces y *biceps* aparece 4 veces. Por el contrario, las variables *knee* y *weight* no aparecen en ningun modelo, mientras que las variables *adipos*, *knee*, *thigh* y *forearm* aparecen todas ellas 2 veces. Esto indica que las variables *wrist*, *abdom* y *age* son las variables con mayor poder predictivo pues son las que aparecen con mayor frecuencia en los modelos explorados, esto quiere decir que las medidas de la muñeca y el abdomen, así como la edad predicen de manera efectiva el valor promedio del porcentaje de grasa corporal en los hombres, las medidas del cuello y biceps tienen un menor poder

predictivo para el promedio de porcentaje de grasa corporal y por otro lado las medidas de la *rodilla* y el *peso* no influyen en la predicción de la variable de interés.

Además, de entre todos los modelos explorados, aquel que pedice el valor promedio del porcentaje de grasa corporal de manera óptima es el modelo que se obtiene tras realizar una selección de variables por el método de mejor subconjunto [3], pues es el modelo que muestra un mejor rendimiento en las métricas consideradas: tiene tanto para la métrica MSE como para la métrica MAE un valor más bajo lo cual sugiere una mayor precisión y consistencia en las predicciones del modelo ajustado, además el valor de su coeficiente de correlación es 0.852569, cabe recordar que entre más se acerque este valor a 1 esto indica una mejor estimación de la variable respuesta, sin embargo el valor obtenida es razonable dado los valores obtenidos para el resto de modelos. Lo que ese modelo indica es que las variables que mejor predicen el valor promedio del porcentaje de grasa son *wrist*, *abdom* y *height*. La regla obtenida es la siguiente [3]:

$$\mathbb{E}(\hat{brozek}) = \beta_1 * height + \beta_2 * abdom + \beta_3 * wrist \tag{1}$$

Donde los valores que toman los coeficientes asociados a las variables son: $\beta_1 = -0.41845$, $\beta_2 = 0.72312$ y $\beta_3 = -1.48367$. Lo cual adquiere la siguiente interpretación: un aumento del 100 % en la medida de la altura *height* está asociado a una disminución del 41 % del promedio de porcentaje de grasa si se dejan el resto de variables fijas, un aumento del 100 % para la variable *abdom*, que representa la medida del abdomen, está asociado a un aumento del 72 % del valor del promedio de porcentaje de grasa y por último, al dejar al resto de variables fijas notamos que un aumento del 100 % de la variable wrist está asociado a una disminución de 148 % del promedio de porcentaje de grasa. Como se menciono anteriormente, en los modelos examinados, las variables *height*, *abdom* y *wrist* destacan por su frecuente aparición, lo que implica su significativa influencia en la predicción del promedio de grasa corporal. Así, se concluye que una mayor altura o un tamaño de muñeca más grande en una persona se asocia con una predicción de menores niveles de grasa corporal. Por el contrario, un abdomen de mayor tamaño indica que esa persona tendrá una proporción más alta de grasa corporal.

Referencias a los chunks

- **[1]** El código usado para implementar esto puede encontrarse en los chunks de código del archivo RMarkdown, específicamente los chunks Preprocesamiento(60) RemocionCasosExtraños(82).
- **[2]** Los respectivos modelos así como sus métodos de entrenamiento y validación se pueden encontrar en los siguientes chunks de código del archivo RMarkdown: PrimerosAjustes(línea de código 146), EfectosPrincipales_SeleccionBIC (línea de código 340), ModeloInteracciones_SeleccionBIC(línea de código 411), ModeloCuadratico_SeleccionBIC(línea de código 485), MetodoLasso_EfectosPrincipales(línea de código 474) MetodoLasso_Interacciones(línea de código 674) Lasso_VariablesAlCuadrado(línea de código 776) MejorSubconjunto(línea de código 873). Además de esto los summaries relevantes correspondientes a los modelos ajustados se pueden encontrar en las siguientes líneas de código: Ajuste_EfectosPrincipales (línea 158), Ajuste_Interacciones (línea 159), Ajuste_VariablesAlCuadrado (línea 160), EfectosPrincipales_BIC (línea 350), Interacciones_BIC (línea 422), Ajuste_Cuadrático (línea 512) EfectosPrincipales_Lasso (línea 603) , Interacciones_Lasso (línea 704), Ajuste_Cuadrático_Lasso (línea 804), AjusteMejor_subconjunto (línea 906).
- **[3]**. Consultar chunk de código *MejorSubconjunto* ubicado en la línea de código 873

Modelos de predicción para detectar la diabetes según variables clínicas.

Para comenzar el análisis visualizamos las variables de interés distinguiendo por los grupos a clasificar “neg” y “pos”, tras una limpieza en los datos notamos la proporción de personas que no tienen diabetes es mucho mayor a quienes sí la tienen. Si consideramos variable por variable, para la mayoría el grupo “pos” presenta mayor variabilidad y una mediana superior, las variables que más parecen marcar diferencia por grupos son pregnant, glucose y age.

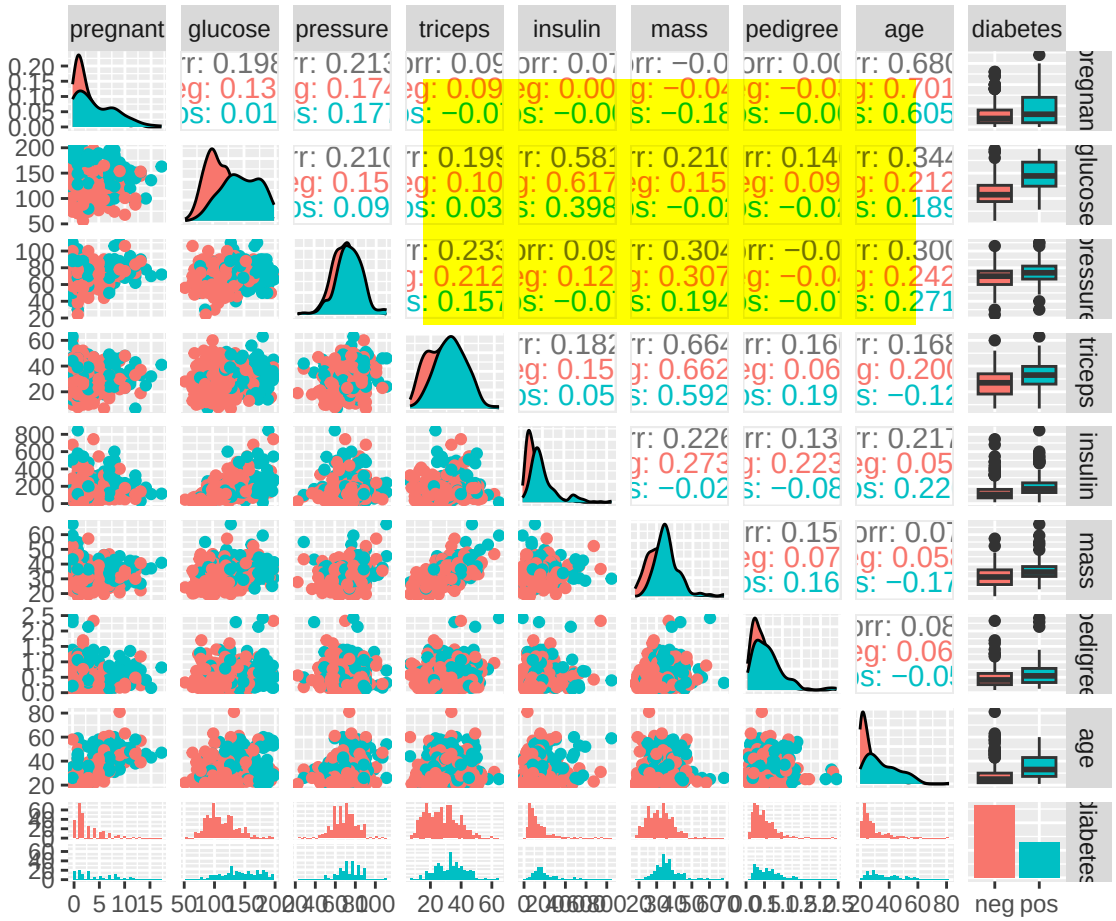


Figura 1: ggpairs variables explicativas por variable de respuesta

A fin de intuir un posible camino para el ajuste de modelos se obtuvieron las componente principales con prcomp (Chunk CompPrin), con 4 se recupera un 78 % de la varianza total y de manera individual aportan más del 10 % por lo que decidimos quedarnos con estas, para su interpretación revisamos las correlaciones con las variables originales y el resultado fue el siguiente:

1.- Para la primera componente, glucose, triceps y age son las de mayor peso con correlación mayor a .6, a excepción de pegigree las demás tienen correlacion de .5

2.- Para la segunda componente pregnant y edad en sentido negativo y mass con triceps son las de mayor peso, por encima de .5, es decir entre más embarazos y mayor edad menor es el valor en esta componente pero a mayor masa y mayor valor del pliegue en el triceps mayor es este componente, no se ve una clara interpretación.

3.- Para la tercer componente la glucosa y la insulina son las únicas mayores a .5, recordemos el cuerpo convierte los alimentos en azúcar y los envía a la sangre, luego, la insulina ayuda a trasladar el azúcar (glucosa) de la sangre a las células. Esta componente podría referirse a este proceso conjunto.

4.- En cuanto la cuarta componente, esta se la lleva practicamente pedigree, lo cuál tiene sentido pues la descendencia es una variable muy significativa en cuanto a las enfermedades.

A continuación proyectamos los grupos de interes sobre las componentes principales donde podemos notar una división por grupos a exepción de la última gráfica, además parece haber un comportamiento lineal:

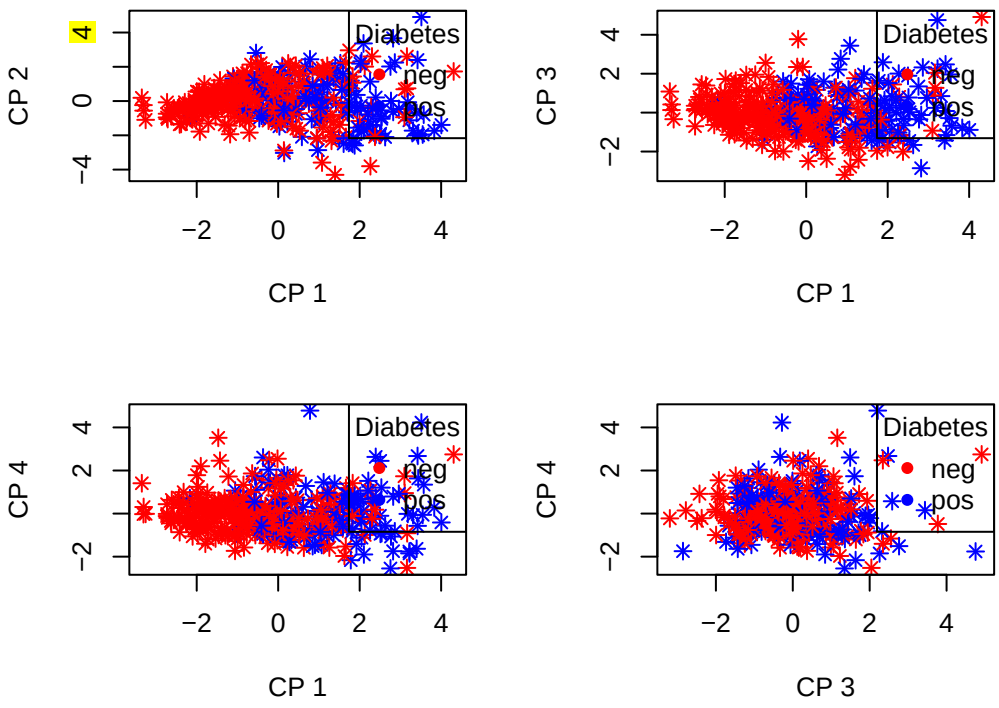


Figura 2: Componente principales por grupos a clasificar

En el equipo de trabajo hicimos busqueda de diferentes modelos para la predicción del padecimiento de diabetes en los pacientes. Para nuestro análisis consideramos los siguientes modelos:

-Regresión logit, efectos principales y predecir probabilidades con punto de corte 0.5 (Chunk EfectPrincip)-Regresión logit, efectos principales con selección de variables, método mejor subconjunto y predecir probabilidades con punto de corte 0.5 (Chunk EfecPrinMS)
-Regresión logit, con interacciones $.^2$ y selección de variables, método mejor subconjunto y predecir probabilidades con punto de corte 0.5 (Chunk InteraccionMS)
-Regresión logit, selección método por pasos both y predecir probabilidades con punto de corte 0.5 y predecir probabilidades con punto de corte 0.5 (Chunk StepBothEP)
-Regresión logit, con interacciones $.^2$, selección por Fordware y predecir probabilidades con punto de corte 0.5 (Chunk fprdwardInteraccion2)
-Regresión logit, con interacciones $.^2$ más las variables originales al cuadrado y selección tipo lasso con lambda tuneado por CV, se elige lambda.min y se asigna a la clase de mayor probabilidad con punto de corte 0.5 (Chunk LassoMasCompleto)
-Naive Classifier (Chunk Naive)
-LDA y QDA asignando a la clase de mayor probabilidad (Chunk´s LDA y QDA)
-KNN, con tunning con 5 CV (Chunk Knn)
-Random Forest, tuneando el hiperparámetro mtry con CV y 200 árboles (Chunk Random Forest)
-Regresión probit, efectos principales y asignando la probabilidad con punto de corte 0.5 (Chunk glmprobit)

Presentamos una tabla que resume el modelo, la regla y la metrica:

Como podemos apreciar, el modelo con mayor precisión global es la Regresión logit, con interacciones $.^2$ más las variables originales al cuadrado y selección tipo lasso con lambda tuneado por CV.
Sin embargo, nosotros queremos clasificar, detectar la enfermedad y accuracy no es buena opción porque puede estar sesgado ya que hay muchos "negativo" y pocos "positivo". Pero, este modelo sigue siendo el mejor en cuanto a especificidad. Es decir, si quisiéramos reducir la mayor cantidad de falsos negativos podríamos usar este modelo sin ningún problema, ya que si futuras observaciones son clasificadas como "negativo" bajo este modelo, el 91 % de las veces habremos clasificado de manera correcta. Es decir, si llega un paciente nuevo y lo ponemos bajo este modelo y arroja un resultado negativo, lo más probable es que este nuevo paciente no tenga diabetes. Si sale positivo podemos considerar mejor otro modelo:
Respecto a la mejor sensibilidad tenemos que el modelo Naive Classifier es el ganador por mucho. Es decir, para nuestra mala fortuna, todos los demás modelos están a un 50 %, por lo que no ayuda, y en este modelo es en el único que sobrepasamos el 60 %.

Como nos interesa saber si un nuevo paciente es positivo a diabetes, no recomendaríamos usar sólo un modelo.

Cuadro 1: Esquemas de entrenamiento explorados

Modelo	acurracy	recall	specifity
Regresión logit, Efectos Principales	0.77	0.55	0.87
Regresión logit, Efectos Principales, Selección de Variables, Mejor Subconjunto	0.77	0.56	0.87
Regresión logit, interacciones, Selección de Variables, Mejor Subconjunto	0.77	0.55	0.88
Regresión logit, selección método por pasos both	0.77	0.56	0.88
Regresión logit, interacciones, selección por Fordware	0.77	0.55	0.88
Regresión logit, interacciones, selección lasso, lambda tuneado por CV	0.78	0.48	0.91
Naive Classifier	0.77	0.63	0.84
LDA	0.77	0.54	0.88
QDA	0.76	0.58	0.84
KNN	0.74	0.49	0.87
Random Forest, tuneado el hiperparámetro mtry con CV	0.74	0.55	0.85
Regresión probit, Efectos Principales	0.77	0.55	0.88

Los siguientes pasos son importantes para el uso efectivo de estos modelos de predicción para detectar la diabetes:
1. Utilizar el modelo de regresión logit, con interacciones $.^2$ más las variables originales al cuadrado y selección tipo lasso con lambda tuneado por CV.
2. Aquí tenemos dos situaciones:
-Si sale negativo podemos estar bastante seguros para descartar la diabetes. Se puede considerar otro modelo para estar más seguros y evitar preocuparnos demasiado.
-Si sale positivo, usaremos el modelo de Naive Classifier. Si nuevamente sale positivo es bastante probable que el paciente tenga diabetes, por lo que la empresa debería proseguir como sea más conveniente. Es decir, comenzar tratamientos, gastar en realizar estudios de confirmación más certeros, prevenir un avance, entre otros.

Finalmente, observando los coefficientes de todos los modelos planteados, pudimos observar que las variable que más efecto tiene en el diagnóstico de diabetes es, por mucho, la genética (pedigree), y en segundo lugar pudimos notar a la glucosa, la insulina y la edad.

En conclusión, es sumamente importante tener en cuenta estos factores de los pacientes. Lamentablemente, del lado de la genética se tienen las manos atadas, pero cuidar la glucosa y la insulina quizás pueda ayudar a que futuros pacientes tengan menor probabilidad de ser detectados con diabetes.