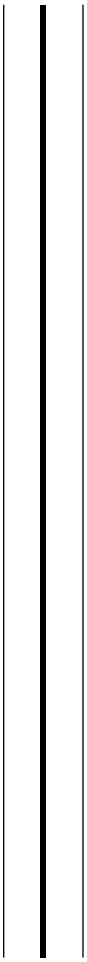
Universidad Nacional Autónoma
de México

Facultad de Ciencias



TAREA 4

Aprendizaje Supervisado

Aguirre Armada Guillermo

Figueroa Torres Ivan Emiliano

Luna Gutiérrez Yanely

Ortiz Silva Ana Beatriz

PROFESOR DE ASIGNATURA:
Guillermina Eslava

PROFESOR DE ADJUNTO:
Sofía Guzman

02 de diciembre de 2020

CIUDAD UNIVERSITARIA, CD. MX.

I.
En este ejercicio trabajamos con la base de datos `Pima.tr, Pima.te` del paquete `MASS`, la cual contiene información de 8 variables observadas en 532 mujeres mayores de 21 años que viven en el área de Phoenix, Arizon, EE.UU. a las cuales se les realizaron pruebas para detectar diabetes. Las variables estudiadas son:

- `npreg`: Número de embarazos.
- `glu`: Glucemia observada en una prubea oral de tolerancia de glucosa.
- `bp`: Presión arterial diastólica medida en mm Hg.
- `skin`: Espesor del pliegue de la piel del tríceps medido en mm.
- `bmi`: Índice de masa corporal en unidades de kg/m^2.
- `ped`: *Diabetes pedigree function.*
- `age`: Edad en años.
- `type`: Contiene el valor `Yes` cuando la persona fue diagnosticada con diabetes según el criterio de la Organización Mundial de la Salud y tiene el valor `No` en el otro caso. Los niveles fueron modificados como 0 ("No") y 1 ("Yes").

A excepción de la variable `type` que ya se encuentra registrada originalmente como factor, las demás variables son numéricas.
En este ejercicio se ajustarán modelos o métodos para clasificar la variable `type` basándonos en los valores de las otras siete variables numéricas.

a) Análisis del discriminante lineal.

Con el análisis del discriminante lineal se busca clasificar cada observación en uno de los dos niveles o categorías de la variable `type`. Este método supone que las variables predictoras condicionadas a que la observación pertenezce a una de las categorías de la variable respuesta siguen una distribución normal.

Dado que buscamos aplicar el método de análisis de discriminante lineal para poder evaluar su poder preditivo, incluiremos las siete variables predictoras con las que contamos.

Para aplicar el análisis de discriminante lineal se usó la función `lda()` de la paquetería `MASS`. Primero se aplicó el método usando todas las observaciones de la base y en la Tabla 4 se muestra la comparación de las predicciones hechas con este método y los valores observados de la variable respuesta, a partir de la cual se obtuvieron las tasas de error de clasificación aparente que se muestran en la Tabla 2.

	0	1
0	317	38
1	75	102

Table 1: **Matriz de confusión.** *En las columnas se encuentran las predicciones obtenidas.*

El grupo formado por las observaciones que pertenecen al nivel 1 de la variable `type` pero que fueron clasificadas como 0 (falsos negativos) le denominaremos **grupo 1**, mientras que al grupo de las observaciones que son del nivel 0 pero fueron clasificadas como 1 (falsos positivos) le denominaremos **grupo 0**.

Posteriormente, dividimos a nuestra base de datos en un conjunto de entrenamiento conformado por el ochenta por ciento de nuestras observaciones (426) y un conjunto de prueba formado por el veinte por ciento restante (106). Al conjunto de entrenamiento le aplicamos nuevamente el método de discriminante lineal y con el conjunto de prueba evaluamos el poder predictivo del método.

Las tasas de error de clasificación que se calcularon se muestran en la Tabla 2. Podemos notar que como era de esperarse, el error global en el conjunto de prueba es el más alto, al igual que para el grupo 1 y el grupo 0. Sin embargo, el grupo 1 presenta un error que es 2.5 veces mayor que error del grupo 0 en el conjunto de prueba, lo que nos dice que el análisis de discriminante lineal tiene significativamente más falsos negativos que falsos positivos.

	Global	Grupo 0	Grupo 1
Aparentes	21.2	10.7	42.4
Prueba	43.8	28.9	74.3

Table 2: **Tasas de error de clasificación** aparentes, de entrenamiento y de prueba, expresadas en porcentajes. Notamos que el error global en el conjunto de prueba es dos veces mayor al error aparente global.

Aún así, el método nos da errores globales que son menores al cincuenta por ciento, por lo que obtenemos una clasificación mejor a que si asignaramos los grupos al azar.

Ahora realizamos mil repeticiones del proceso anterior (entrenamiento y prueba) para obtener tasas de error de clasificación más precisas.

	Global	Grupo 0	Grupo 1
Valor estimado	43	28	72
Error estándar	2.50	1.42	2.87

Table 3: **Tasas de error de clasificación.** *Usando training-test con B=1000.*

b) *Naive Bayes.*

Utilizando el método Naive Bayes podemos modelar a nuestra variable respuesta `type` de acuerdo a el "Teorema de Bayes".

$$P(A|B) = \frac{(P(A) * P(B|A))}{P(B)}$$

En este modelo asumimos que las variables son independientes y calculamos la probabilidad de clasificación en los niveles de `type` $= \{0(\text{"}No\text{"}), 1(\text{"}Yes\text{"})\}$ acuerdo a una probabilidad a priori.

Usando la paquetería `e1071` ajustamos el modelo `model` usando la función `NaiveBayes()`

Vemos las probabilidades iniciales de pertenecer a cada grupo.

0	1
0.66	0.33

Table 4: **Probabilidades a Priori** *Se observan las probabilidades iniciales de no tener o tener diabetes.*

Aparte de estas primeras probabilidades obtenemos las medias y desviaciones estándar de los valores de las 7 variables condicionas a su clasificación según type.

Variable	0.mean	1.mean	0.sd	1.sd
npreg	2.92	4.70	2.78	3.91
glu	110.01	143.11	24.28	31.26
bp	69.91	74.70	11.90	12.52
skin	27.29	32.97	10.08	10.39
bmi	31.42	35.81	6.54	6.61
ped	0.44	0.61	0.29	0.39
npreg	29.22	36.41	9.90	10.83

Table 5: **Medidas condicionadas** *Se observan las medias y desviaciones estándar de los valores de cada variable tomando de referencia si tienen diabetes o no. Podemos notar, por ejemplo, que las mujeres que no se diagnosticaron con diabetes tienen una media mayor de números embarazos.*

Con la función `predic()` predecimos la clasificación de cada observación a "0" y "1". En la tabla 6 vemos un ejemplo de esta predicción según la probabilidad a posteriori calculada.

Observación	0	1	npreg	glu	bp	skin	bmi	ped	age	type
1	0.98	0.01	5	86	68	28	30.2	0.36	24	0
2	0.01	0.99	7	195	70	33	25.1	0.16	55	1
3	0.83	0.16	5	77	82	41	35.8	0.15	35	0
4	0.10	0.89	0	165	76	43	47.9	0.25	26	0
5	0.99	0.01	0	107	60	25	26.4	0.133	23	0
6	0.69	0.30	5	97	76	27	35.6	0.378	52	1

Table 6: **Probabilidades a posteriori de clasificasión** Se observan las probabilidades de las primeras 6 observaciones.

Vemos la matriz de confusión del modelo

	0	1
0	289	66
1	61	116

Table 7: **Matriz de confusión Naive Bayes.** *Se predice más erroneamente a mujeres con diabetes que a mujeres que no*

Encontramos los errores de clasificación aparentes y utilizamos técnica de remuestreo Training- test con 1000 simulaciones.

	Global	0	1
Aparentes NaiveB	23.87	18.59	34.46
Entrenamiento B=1000	23.06	17.66	33.93
Prueba B=1000	38.88	16.21	83.87

Table 8: **Errores de clasificación.** *Los errores están expresados en porcentaje. Tenemos que el error global en entrenamiento es ligeramente menor en el conjunto de entrenamiento pero aumenta considerablemente en el conjunto prueba.*

Habiendo usado usado un cuarto de la muestra para el conjunto prueba obtenemos que 0.61 del modelo tiene un buen valor predictivo.

c) Regresión Logística.

Otra forma de modelar la variable respuesta `type` es ajustando una regresión logística. Para determinar qué variables incluir en este modelo primero ajustamos una regresión logística tomando en cuenta las siete variables predictoras con las que contamos. Nos referiremos a este modelo como **modelo 1**.

	Valor estimado	p-value
(Intercept)	-9.55	0.00
`npreg`	0.12	0.01
`glu`	0.04	0.00
`bp`	-0.01	0.46
`skin`	0.01	0.65
`bmi`	0.08	0.00
`ped`	1.31	0.00
`age`	0.03	0.06

Table 9: Resumen del modelo 1. *Los p valores que son menores a 0.001 están registrados como 0.00*

En la Tabla 9 se observa que solo dos variables (`bp` y `skin`) no son significativas a un nivel de 0.05, por lo que ajustamos primero un modelo sin la variable `bp` (**modelo 2**). Usando la función `anova()` para comparar la significancia de los coeficientes del modelo 1 y el modelo 2 notamos que es mejor (en el sentido de que la variable que quitamos no es significativa) quedarnos con el modelo 2. Posterioremente ajustamos otro modelo sin ninguna de estas dos variables (**modelo 3**) y nuevamente comprobamos que es mejor quedarnos con el modelo con menos variables.

Notamos que la variable `age` tiene un valor muy cercano a 0.05, por lo que probamos con un modelo que no incluye a esta variable y a ninguna de las dos antes mencionadas (**modelo 4**) y notamos que la variable `age` resulta no ser significativa.

Por último, basándonos en la Tabla 23 (en el anexo del análisis exploratorio) ajustamos un modelo que incluye la interacción entre la variable `age` y la variable `npreg` (**modelo 5**). Sin embargo, al compararlo con el modelo 1 (con todas las variables predictoras) notamos que el coeficiente de la interacción no es significativa.

Modelo	Fórmula	AIC	BIC
1	`npreg` + `glu` + `bp` + `skin` + `bmi` + `ped` + `age`	482.3	516.5
2	`npreg` + `glu` + `skin` + `bmi` + `ped` + `age`	480.5	510.5
3	`npreg` + `glu` + `bmi` + `ped` + `age`	479.1	504.7
4	`npreg` + `glu` + `bmi` + `ped`	480.3	501.7
5	`npreg` + `glu` + `bp` + `skin` + `bmi` + `ped` + `age` + `npreg:age`	481.5	520.0

Table 10: **Modelos propuestos.** *Podemos ver que el modelo 3 tiene el AIC más bajo mientras que el modelo 4 tiene el menor BIC.*

Una vez ajustados los 5 modelos descritos, comparamos el AIC y BIC de cada modelo (Tabla 10) y las tasas de error de clasificación aparente. Posteriormente evaluamos el poder predictivo de cada modelo. Para esto usamos primero un conjunto entrenamiento del 80% del total de los datos y con éste ajustamos los 5 modelos. Con los datos que no están en este conjunto (el conjunto de prueba) evaluamos las predicciones.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
Eror Global aparente	21.2	20.7	20.3	**20.1**	21.2
Error Grupo 0 aparente	10.7	10.4	10.1	**9.6**	11.0
Error Grupo 1 aparente	42.4	41.2	**40.7**	41.2	41.8
Error Global prueba	**26.4**	27.4	**26.4**	28.3	**26.4**
Error Grupo 0 prueba	19.7	19.7	**21.1**	21.1	**21.1**
Error Grupo 1 prueba	40.0	42.9	**37.1**	42.9	**37.1**

Table 11: **Comparación de modelos.** *Tasas de error de clasificación en porcentajes.*

En la Tabla 11 notamos que en general, el modelo 4 tienen mejores tasas de error de clasificación aparentes, pero en el conjunto de prueba los errores superan a los de los otros modelo (son en promedio dos veces mayores a las aparentes), por lo que podemos decir que en realidad este modelo está sobreajustado a los datos. Por otro lado, el modelo 3 tiene un buen desempeño en el conjunto prueba, al igual que el modelo 5, sin embargo, el modelo 3 tiene menos variables, por lo que decidimos quedarnos con este modelo.

Para validar las tasas de error de clasificación del modelo elegido, repetimos el proceso de *training-test* 1000 veces y obtuvimos los resultados que se muestran en la Tabla 12. Notamos que estas tasas solo difieren en promedio en un 1.5% de las tasas aparentes, lo que nos dice que tenemos un modelo que realiza un buena predicción.

	Global	Grupo 0	Grupo 1
Valor estimado	21.38	10.90	41.98
Error estándar	3.691	4.089	8.388

Table 12: **Tasas de error de clasificación.** *Usando training-test con B=1000.*

d) *SVM.*

Por último utilizamos el método que Máquina de soporte de vectores que construye hiperplanos que separan de mejor forma los grupos de clasificación de la variable predictora, `type`.

Con las paqueterías `MASS` y `e1071`, usamos la función `svm()` para ajustar 3 modelos con diferente kernel: lineal, poligonal y radial.Obteniendo los modelos `svm.fit$lineal`, `svm.fit$polinomial` y `svm.fit$radial`.

Realizando los summaries, vemos la distribución de los vectores en la tabla siguiente.

Kernel	Costo	Num. Vectores de soporte	Num. V.S 0	Num. V.S 1
Lineal	100	255	128	127
Polinomial	100	245	127	118
Radial	100	257	143	114

Table 13: **Calculo de Vectores de soporte.** *Se separan los valores de la tabla según el kernel del modelo y su clasificación de type*

Se observa que en el modelo lineal, el número de vectores del grupo 0 es prácticamente igual al del grupo 1. El modelo polinomial tiene menor número de vectores y estos estan dividimos en los grupos 0 y 1 casi iguales. En cambio el modelo radial tiene más número de vectores y su clasificación va más acorde a las muestra.

Sacamos la matriz de errores aparentes de los tres modelos

Kernel	svm.global	svm.0	svm.1
Lineal	21.05	10.42	42.37
Polinomial	15.04	4.789	35.59
Radial	4.323	0.8451	11.3

Table 14: **Tasas de errores aparentes SVM.** *Se separan los valores de la tabla según el kernel del modelo y su clasificación de type. Tenemos que los errores del modelo radial son mucho menores*

Vemos las matrices de confusión de los tres modelos.

	Li. 0	Li. 1	Pol. 0	Pol. 1	Rad. 0	Rad. 1
0	318	37	338	17	352	3
1	75	102	63	114	20	157

Table 15: **Matriz de Confusión SVM.** *Se separan los valores de la tabla según el kernel del modelo y su clasificación type. Vemos que la confusión es mayor en el modelo lineal*

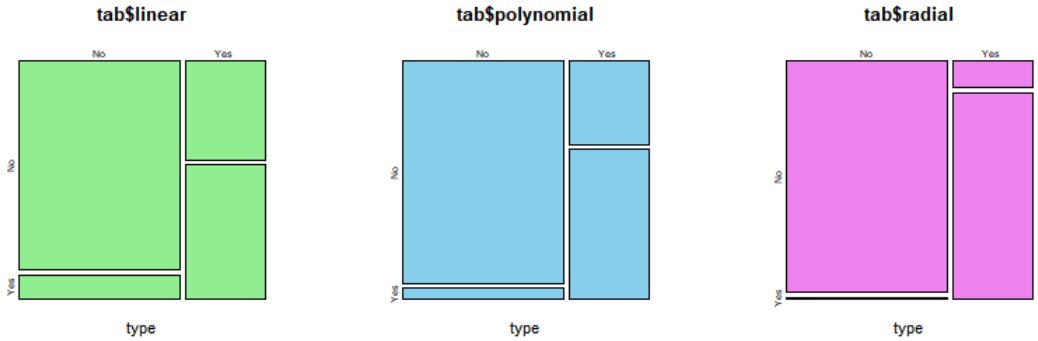


Figure 1: **Matriz de confusión de los 3 modelos svm.** *De manera práctica podemos observar como de forma descendiente disminuye las areas de confusión del modelo, siendo el modelo radial el de menor confusión*

Según lo observado podríamos elegir en primera instancia al modelo `svm.fit$radial`, por presentar menores errores de clasificasión.

Para la selección del modelo utilizamos la función `tune()` para encontrar los mejores parámetros y modelos.

Vemos los errores aparentes de los mejores modelos.

Modelo	Global	Grupo 0	Grupo 1
Mejor Lineal	21.05	10.42	42.37
Mejor Polinomial	17.86	3.66	46.33
Mejor Radial	9.59	3.94	20.90

Table 16: **Tasas de error aparentes de Mejores SVM.** *Las tasas están calculadas en porcentaje, teniendo que el mejor radial tiene errores más pequeños*

Seleccionamos el mejor modelo radial como nuestro modelo final. Tenemos que el mejor modelo es el radial con cost = 1 y gamma = 0.5 $\texttt{fit}_r ad\$best.parameters$

Utilizamos la técnica de remuestreo Training-test con 500 y 1000 simulaciones y particionamos la muestra y obtenemos los resultados de .

	Global	0	1
Aparentes Mejor Radial	21.05	17.86	9.586
T-t Mejor Radial B=500	22.00	45.77	10.061
T-t Mejor Radial B=1000	22.09	45.71	10.164

Table 17: **Errores de clasificación del mejor modelo radial.** *Los errores están expresados en porcentaje. Tenemos que el error global cambia ligeramente respecto a sus simulaciones.*

Observamos la distribución de los errores calulados

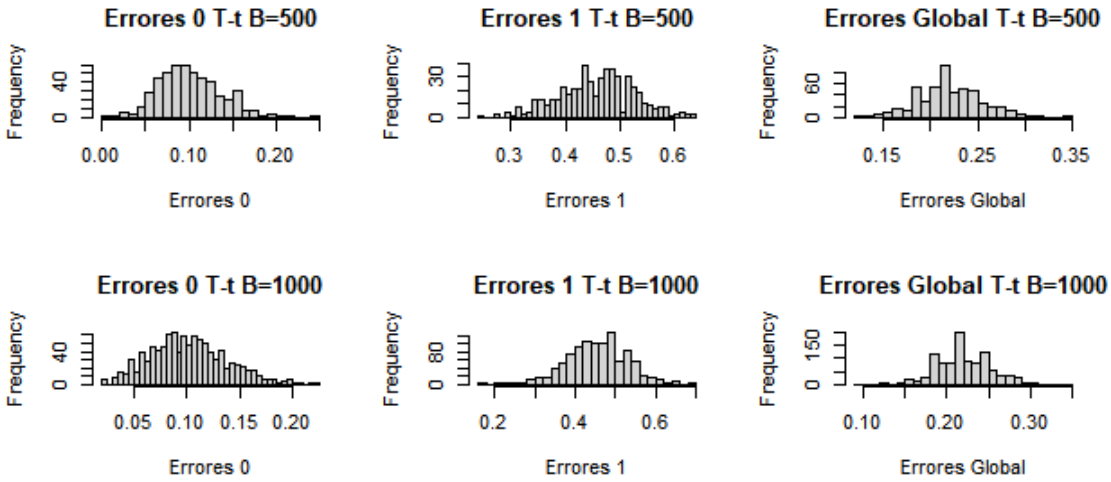


Figure 2: **Distribución de los errores calculados**

Tenemos que con este modelo acierta en su poder predictivo cerca del 80 por ciento. También notamos que solo difiere de la tasa global aparente un 1.05 por ciento cuando realizamos las 1000 simulaciones.

e) **Selección del modelo**

Modelo	Global	0	1
DLA	42.8	27.9	72.4
NaiveBayes	38.9	16.2	83.9
Regresión Log	21.4	10.9	42.0
SVM	22.1	45.7	10.2

Table 18: **Errores de clasificación Prueba finales.** *Los errores están expresados en porcentaje.*

Los mejores modelos son la Regresión logistica y SVM, el primero tiene una tasa de error global 0.7% menor, pero el modelo de SVM tiene tasa de error menor para el grupo 1 (falsos negativos) por lo que si consideramos que es más importante diagnosticar correctamente los casos positivos, elegiremos el modelo de SVM.

II. Este problema involucra predictores de tipo categórico. Es un problema de clasificación binaria. La variable CAD identifica a las dos clases.
Coronary artery disease Data.
A cross classified table with observational data from a Danish heart clinic. The response variable is CAD. A data frame with 236 observations on the following 14 variables.
La variable respuesta CAD es la presencia o ausencia de la Enfermedad de las Arterias Coronarias.

- `Sex` a factor with levels Female Male
- `AngPec` a factor with levels Atypical None Typical
- `AMI` a factor with levels Definite NotCertain
- `QWave` a factor with levels No Yes
- `QWavecode`a factor with levels Nonusable Usable
- `STcode` a factor with levels Nonusable Usable
- `STchange` a factor with levels No Yes
- `SuffHeartF` a factor with levels No Yes
- `Hypertrophi` a factor with levels No Yes
- `Hyperchol` a factor with levels No Yes
- `Smoker` a factor with levels No Yes
- `Inherit` a factor with levels No Yes
- `Heartfail` a factor with levels No Yes
- `CAD` a factor with levels No Yes

A lo largo de este ejercicio buscaremos encontrar el modelo con el mayor poder predictivo, tratando de minimizar las tasas de error de prueba. Especialmente cuando estas pertenezcan al grupo 1, esto debido a que se busca clasficar correctamente a aquellas personas que sí tengan la enfermedad.

	NB		**RL**		SVM					
					Lineal		Polinomial		Radial	
CAD	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
No	114	15	116	13	120	9	118	11	124	5
Yes	17	90	15	92	13	94	25	82	3	104

Table 19: **Matriz de confusión.** *De los modelos Naive Bayes (NB), Regresión Logística (RL) y Suport Vector Machine (SVM) con los distintos tipos de kernel (lineal, polinomial y radial).*

En la tabla 19 tenemos la matriz de confución de cada uno de los modelos, Suport Vector Machine utilizando el kerner radial, es el que asigna mejor a las observaciones, sólo falla en 8 observaciones que corresponden a un 3%.

Para los análisis subsecuentes nos apoyaremos de la tabla 20 , la cual muestra los errores obtenidos en cada modelo.

1. **Naive Bayes**: Este modelo pareciera tener un error aceptable siendo el aparente de un 13% el problema es claro cuando nos fijamos en los errores de pueba, son muy cercanos a un 50% inclusive para el grupo 1 es mayor a dicho porcentaje de tal modo que de haber elegido al azar entre ambas clases tendríamos mejores resultados.

2. **Regresión logística**: Si nos fijamos en los errores globales, el aparente y el que obtuvimos del método Cross Validation son muy cercanos sólo variando un .4% mientras que el de prueba se dispara creciendo un 4% .

3. **Suport Vector Machine:** En este modelo podemos comparar entre los distintos kernels, es fácil notar que el mejor es el radial, siendo que las tasas de error de este son más de 2 veces mejores a las demás, esto permanece así independientemente de si calibramos o no los parámetros.

4. **Selección del modelo**: Para elegir el mejor de los 3 modelos se usará nuevamente la tabla 20, si comparamos los errores de prueba de todos los modelos veremos que el mínimo se encuentra en el modelo de Suport Vector Machine utilizando el kerner radial, el cual tiene un error muy bajo únicamente un 3%

		Global	Grupo 0	Grupo 1
NB	Aparentes	13.56	11.63	15.89
	Entrenamiento	13.60	11.56	16.12
	Prueba	**48.01**	**38.66**	**57.82**
	10-fold CV, B=500	13.53	11.53	15.92
RL	Aparentes	11.86	10.08	14.02
	CV δ_1	11.43		
	Prueba	15.47	18.19	12.80
Parámetros sin calibrar				
SVM	Lineal	9.32	6.98	12.15
	Polinomial	14.83	13.18	16.82
	Radial	3.39	3.88	2.80
Parámetros calibrados				
SVM	Lineal	9.75	5.43	14.95
	Polinomial	9.32	7.75	11.21
	Radial	3.39	3.88	2.80
	CV K=10, B=200	15.58	11.62	20.04
	Repeated Holdout (B=50, 4/5)	16.29	12.15	20.93

Table 20: **Tasas de error de clasificación** Expresada en porcentaje de los modelos Naive Bayes (NB), Regresión Logística (RL) y Suport Vector Machine (SVM).

III.
En este ejercicio trabajamos con la base de datos `Glucose1`, la cual contiene 145 observaciones de 7 variables: la primera `Patient` corresponde al número de paciente, la cual no es de interés. Las demás variables son:

- `Weight`: Peso relativo.
- `Fglucose`: Glucosa plasmática en ayunas.
- `GlucoseInt`: Intolerancia a la glucosa.

- `InsulinResp`: Respuesta en la insulina a la administración oral de glucosa.
- `InsulineResist`: Resistencia a la insulina.
- `Class`: Etiqueta de la clase. Corresponde al tipo de diabetes que presenta cada paciente: 1 (*normal*), 2 (`chemical diabetes`), 3 (`overt diabetes`).

La variable `Class` fue transformada a factor con los 3 niveles descritos. Las otras cinco variables fueron guardadas como numéricas.

Usando las primeras cinco variables numéricas como predictores y la variable `Class` como variable respuesta, ajustamos un modelo de regresión logística multinomial con la finalidad de evaluar su poder predictivo.

Como primer modelo usamos todas las variables predictoras (**Modelo 1**) usando la función `multinom()` del paquete `nnet`. En la Tabla 21 se muestra el AIC y BIC de este modelo.
Como segunda opción ajustamos un modelo que además de las cinco variables predictoras incluye la interacción entre las variables `Fglucose` y `GlucoseInt` (**Modelo 2**) basándonos en la correlación observada entre estas variables (ver el Anexo 1.2).
Por último, ajustamos también un modelo con todas las variables predictoras, la interacción entre `Fglucose` y `GlucoseInt` y la interacción entre `GlucoseInt` e `InsulineResist` (**Modelo 3**).

Modelo	Fórmula	AIC	BIC
1	`Weight` + `Fglucose` + `GlucoseInt` + `InsulineResp` + `InsulineResist`	24	30
2	$\sim$. + `Fglucose:GlucoseInt`	28	70
3	$\sim$. + `Fglucose:GlucoseInt` + `GlucoseInt:InsulineResist`	32	80

Table 21: **Modelos propuestos.** *El modelo 1 tiene un AIC 14% menor que el modelo 2.*

Los modelos antes descritos fueron ajustados usando todos los datos de la base, con lo que calculamos las tasas de error aparentes para cada uno.
Para evaluar el poder predictivo de cada modelo usamos el método de *training-test* con un conjunto de entrenamiento de 116 observaciones (80% de los datos) seleccionadas al azar. Repetimos el proceso 1000 veces para cada modelo, con lo que obtuvimos las tasas de error de clasificación del conjunto de prueba que se muestran en la Tabla 22

Tasas de error	3	1	2	Global
Aparente Modelo 1	0	0	0	0
Aparente Modelo 2	0	0	0	0
Aparente Modelo 3	0	0	0	0
Prueba Modelo 1	8.1	7.2	3.9	5.8
Prueba Modelo 2	7.5	5.0	3.1	4.7
Prueba Modelo 3	10.3	7.2	3.1	5.9

Table 22: **Tasas de error de clasificación** aparentes y de prueba con B=1000 para cada modelo.

Las columnas 3, 1 y 2 de la Tabla 22 hacen referencia a las tasas de error de clasificación de cada categoría, tomando en cuenta las observaciones que pertenecían a dicha categoría pero fueron clasificadas como otra.

Las tasas de error aparente para los tres modelos son 0, por lo que en principio no podríamos decir cual es mejor para predecir, por lo que las tasas en el conjunto de prueba nos dan mejor información al respecto. Notamos que en los tres modelos las tasas de error de clasificación global no superan el 10%, por lo que son modelos relativamente buenos, sin embargo, el Modelo 2 tiene una tasa 1.1% menor que el Modelo 1 y 1.2% menor que el Modelo 3, lo que es una

diferencia relativa de un 20%, por lo que elegimos el Modelo 2 como el mejor modelo predictivo.

1 Anexo

1.1 `Pima.tr, Pima.te`: Análisis exploratorio.

Como se mencionó al inicio del Ejercicio I, esta base de datos cuenta con 532 observaciones de 8 variables.
En la Tabla 23 se muestran las correlaciones de las siete variables numéricas de la base. Podemos notar que estas correlaciones son en su mayoría menores a 0.4 con excepción de una de 0.65 correspondiente a las variables `bmi` y `skin`, lo cual era de esperarse debido a que un índice de masa corporal alto puede estar asociado con un mayor espesor en la piel debido a la grasa que se encuentra en sus capas. Otra correlación importante se presenta entre las variables `age` y `npreg`, la cual es de 0.64 y puede explicarse debido a que en general las mujeres que han tenido un mayor número de embarazos suelen tener mayor edad que las que han tenido menos embarazos.
Exceptuando estos dos casos, podríamos decir que en general no tenemos problemas de colinealidad en las variables predictoras.

	`npreg`	`glu`	`bp`	skin	`bmi`	`ped`	`age`
`npreg`	1.00	0.13	0.20	0.10	0.01	0.01	0.64
`glu`	0.13	1.00	0.22	0.23	0.25	0.17	0.28
`bp`	0.20	0.22	1.00	0.23	0.31	0.01	0.35
skin	0.10	0.23	0.23	1.00	0.65	0.12	0.16
`bmi`	0.01	0.25	0.31	0.65	1.00	0.15	0.07
`ped`	0.01	0.17	0.01	0.12	0.15	1.00	0.07
`age`	0.64	0.28	0.35	0.16	0.07	0.07	1.00

Table 23: **Matriz de correlaciones.** *Las correlaciones más altas se presentan entre las variables* `bmi` *y* `skin` *y las variables* `age` *y* `npreg`*.*

Para ver si las observaciones de cada variable predictora según el nivel de la variable `type` siguen una distribución normal se aplicó la prueba Shapiro para normalidad. Los p valores correspondientes resultaron menores a 0.001 para la mayoría de las variables a excepción de la variable `bp` con el nivel 0 de `type` que tuvo un valor de 0.01 y la variable `bp` con el nivel 1 que tuvo un p valor de 0.05. Como en la prueba de Shapiro, la hipótisis nula es normalidad, entonces solo estas dos variables cumplen con la normalidad a un nivel de 0.01 de significancia.

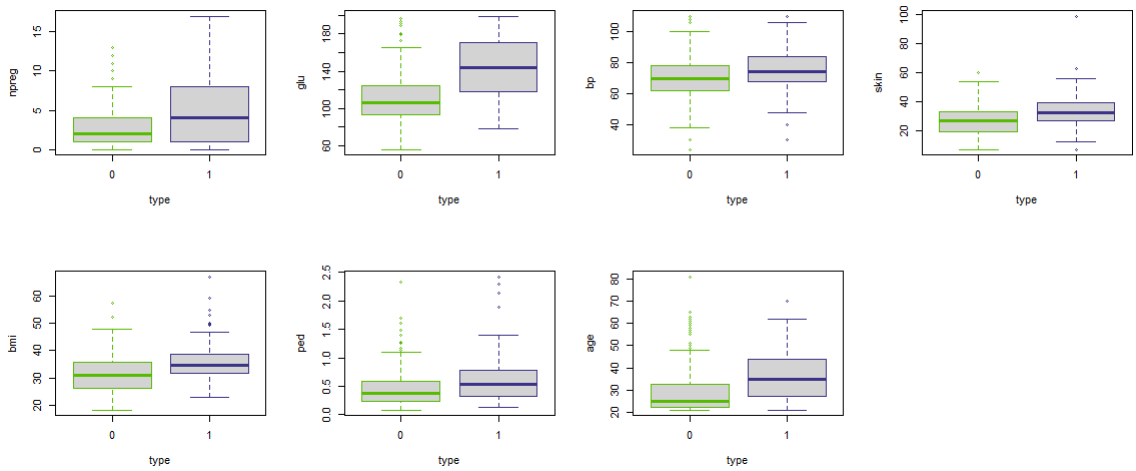


Figure 3: **Boxplots de las variables predictoras según la variable `type`.**

En la Figura 3 se puede notar que para algunas variables como `glu` y `age` la diferencia entre los niveles de `type` es más notoria que en otras, lo que es un indicio de que tal vez estas variables sean más adecuadas que otras para hacer la clasificación de la variable `type`.

Nos gustaría ver si es posible identificar gráficamente una separación entre las categorías de la varible `type` (dado que es la variable respuesta), pero como tenemos siete variables predictoras, no podemos realizar una observación directa que incluya estas siete variables.
Para intentar visualizar algún tipo de separación entre las dos categorías de la variable `type` graficamos los primeros dos componentes principales de las variables predictoras. Esta gráfica su muestra en la Figura 4.

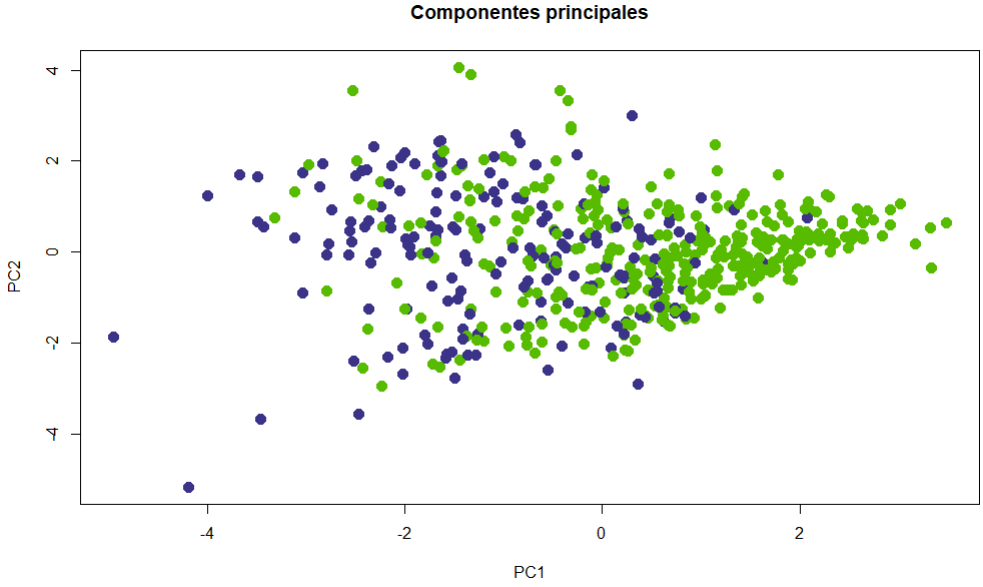


Figure 4: **Primeros dos componentes principales.** *No se muestra una separación muy clara de las dos categorías, sin embargo es posible identificar una concentración de las obsvaciones de una categoría en el lado derecho de la gráfica.*

Por otro lado, podemos tratar de identificar si existe una combinación de dos de las siete variables predictoras que separe las dos categorías de la respuesta usando las gráficas de dispersión por pares.
En la Figura 5 podemos observar que en general no es evidente una separación de las categorías de `type`, pero algunos pares de variables muestran una separación más clara que otras. Por ejemplo, tomando en cuenta solo las variables `glu` y `skin` notamos que las observaciones de la categoría 1 se encuentran más concentradas en valores bajos de `skin` y altos de `glu`, mientras que para la combinación de `npreg` y `ped` no es posible hacer una separación análoga.

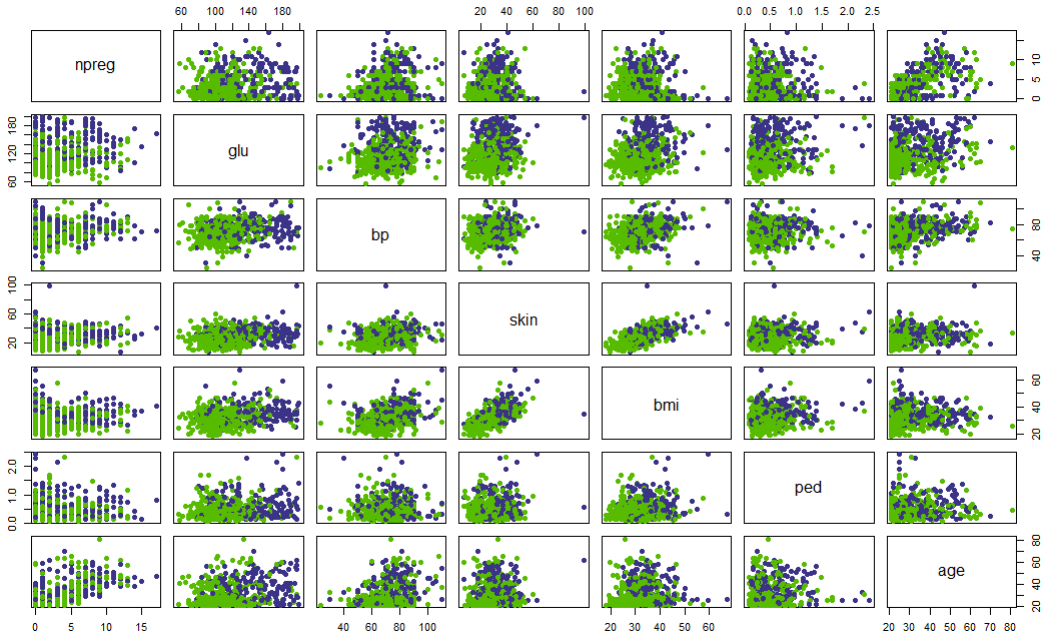


Figure 5: **Gráfica de dispersión por pares.** *Las observaciones que pertenecen a la categoría 0 están coloreadas en verde. Las combinaciones que mejor parecen separar las categorías son las que incluyen a la variable* ***`glu`****.*

1.2 `Glucose1`: Análisis exploratorio.

La base de datos `Glucose1` contiene 145 observaciones de laa 6 variables descritas al inicio del Ejercicio III.
El número de observaciones por nivel de `Class` es:

- 1 (*normal*): 76 observaciones,
- 2 (`chemical diabetes`): 36 observaciones,
- 3 (`overt diabetes`): 33 observaciones.

En la Tabla 24 observamos que la mayoría de las variables tiene una correlación menor a 0.5. Las variables que superan este nivel de correlación pueden indicar un problema de colinealidad.

	Weight	Fglucose	GlucoseInt	InsulinResp	InsulineResist
Weight	1.00	-0.01	0.05	0.22	0.38
Fglucose	-0.01	1.00	0.94	-0.40	0.72
GlucoseInt	0.05	**0.94**	1.00	-0.32	0.73
InsulinResp	0.22	-0.40	-0.32	1.00	0.00
InsulineResist	0.38	**0.72**	**0.73**	0.00	1.00

Table 24: **Matriz de correlaciones.** *Se presentan 3 correlaciones mayores a 0.5.*

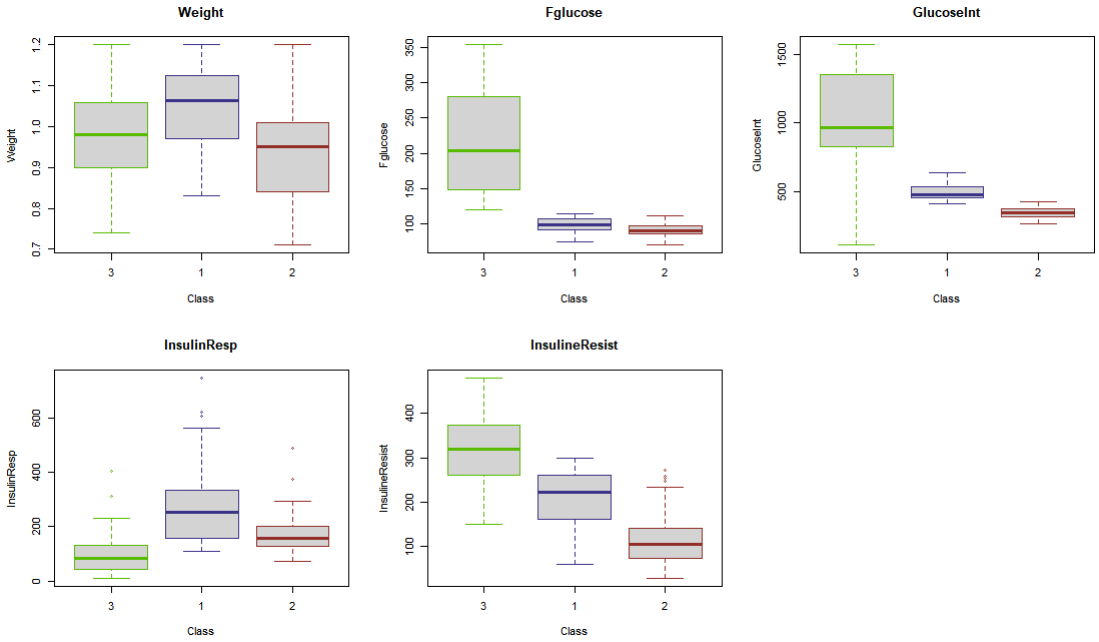


Figure 6: **Boxplots de las variables predictoras de acuerdo a la variable** `Class`.

En la Figura 6 notamos que en la categoría 3 de `Class` se observa mayor diferencia en las observaciones con respecto a las otras dos categorías, por lo que se eligió fijar dicha categoría como categoría de referencia.

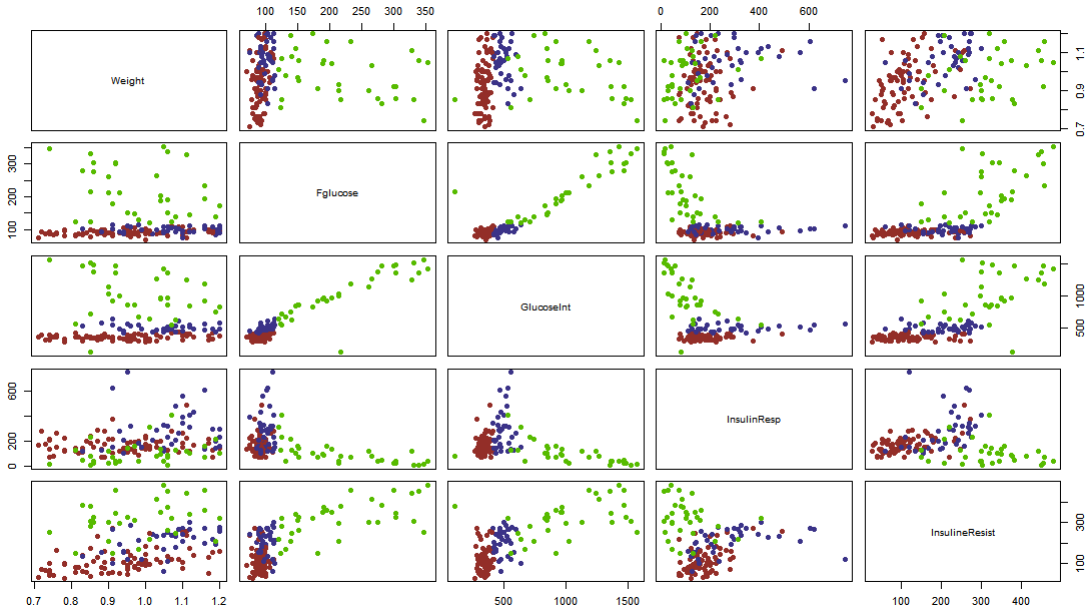


Figure 7: **Gráficas de dispersión por pares.** *Las observaciones están coloreadas de acuerdo al nivel de* `Class` *correspondiente: 1 (morado), 2 (rojo), 3 (verde).*

En las gráficas de dispersión por pares notamos que las observaciones de la categoría 3 se encuentran más dispersas y separadas de las categorías 1 y 2, las cuales están muy juntas y concentradas en una zona de cada gráfica. Esto es otro indicio de que la categoría 3 puede influenciar más al ajustar un modelo para clasificar las observaciones.
Como en las gráficas anteriores (Figura 7) solo podemos observar la relación entre las variables predictoras por pares, realizamos un análisis de componentes principales y graficamos los primeros dos componentes (Figura 8).

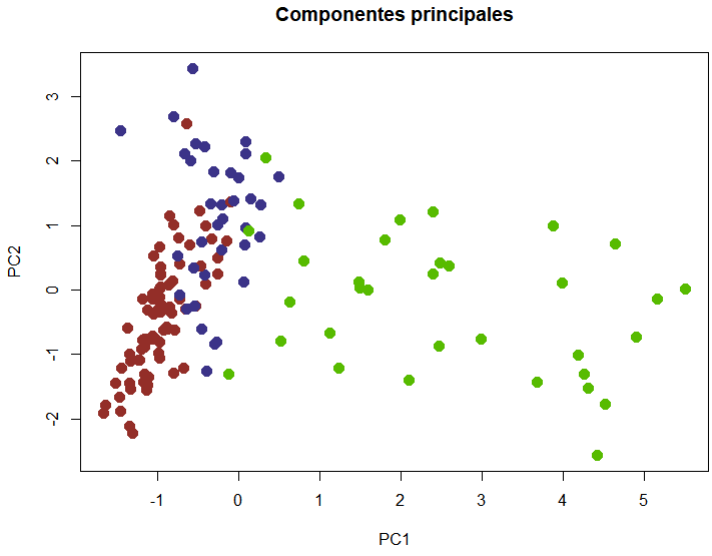


Figure 8: **Gráfica de los dos primeros componentes principales.** *La separación de la categoría 3 es más marcada, pero también es posible identificar una separación entre las categorías 1 y 2.*