Ejercicio 2: Clasificación supervisada

La base de datos PimaIndiansDiabetes2, proviene del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. El objetivo del conjunto de datos es predecir de forma diagnóstica si un paciente tiene diabetes o no, basándose en ciertas mediciones de diagnóstico incluidas en el conjunto de datos. Se impusieron restricciones, en particular, todas las pacientes de esta base de datos son mujeres de al menos 21 años de edad de ascendencia india Pima.

Analisis descriptivo de los datos

Se realizó un analisis descriptivo de los datos para visualizar medidas de tendencia central y de dispersión. Estas se muestran en el Cuadro 2

Cuadro 2: Análisis descriptivo de las variables predictoras

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age
Min.	0.0	56	24.0	7.0	14.0	18.2	0.085	21.0
1st Qu.	1.0	99	62.0	21.0	76.8	28.4	0.270	23.0
Median	2.0	119	70.0	29.0	125.5	33.2	0.450	27.0
Mean	3.3	123	70.7	29.1	156.1	33.1	0.523	30.9
3rd Qu.	5.0	143	78.0	37.0	190.0	37.1	0.687	36.0
Max.	17.0	198	110.0	63.0	846.0	67.1	2.420	81.0

Ademas se realizaron graficos para visualizar mejor los datos

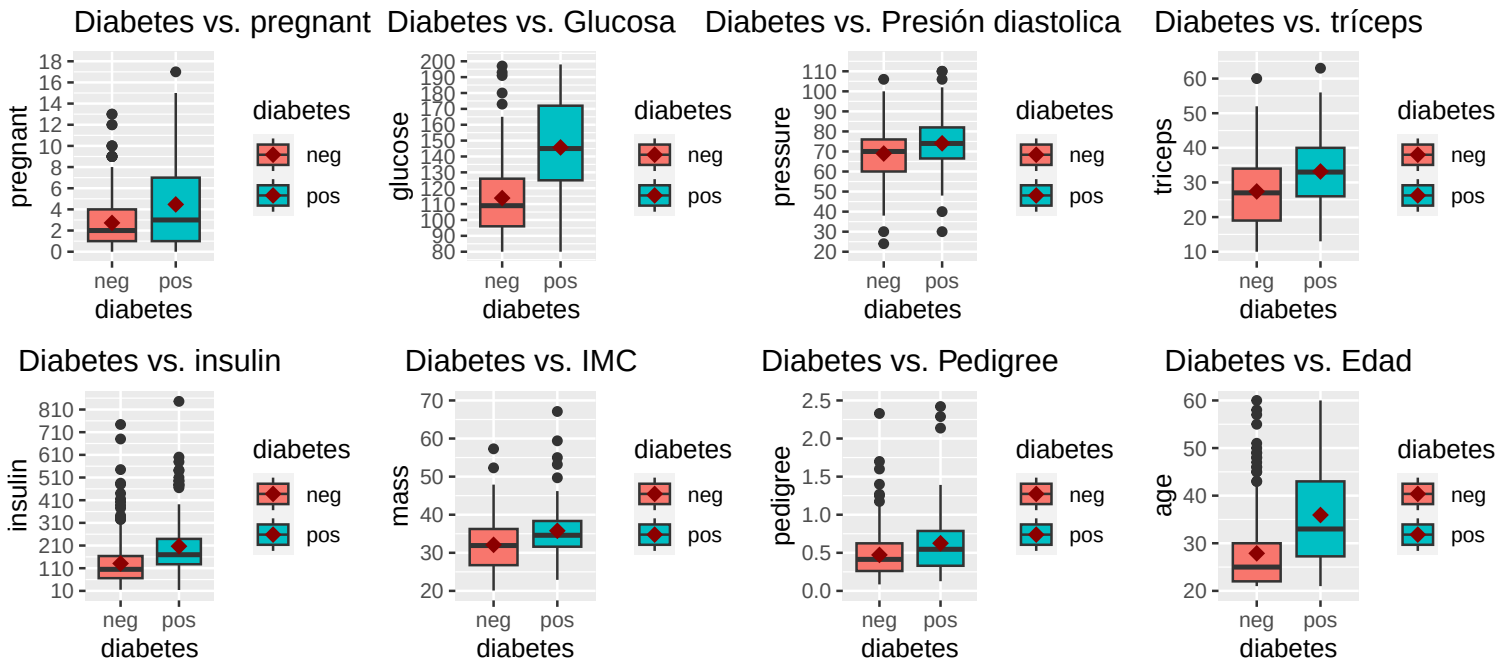

Figura 2: Boxplot de Diabetes v.s. cada variable

Con base a la Figura 2 se observó que para el grupo de mujeres que fueron diagnosticadas con diabetes, tienen mayor concentración de glucosa en plasma. Este mismo grupo tiene mayor grosor del pliegue cutáneo del tríceps, mayores valores en 2-Hour serum insulin, además mayores valores en el indice de masa corporal (IMC) y mayor edad.

Además con esta gráfica se puede observar que las variables tienen distinta escala.

Componentes principales

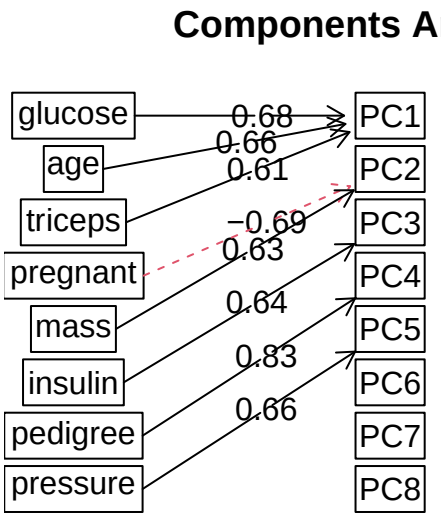

Figura 3: Analisis de componentes principales.

A partir de la Figura 3, se observó que las variables que se encuentran relacionadas con el primer componente principal, son las variables glucose, age y triceps

Además las variables que se encuentran relacionadas con el segundo componente son las variables mass y pregnant.

La variable que se encuentra relacionada con el tercer componente es insulin, la variable que se encuentra relacionada con el cuarto componente es pedigree y variable que se encuentra relacionada con el quinto componente es pressure.

A continuación se presentan los diagramas de dispersión entre los primeros tres componentes principales coloreando las observaciones de acuerdo con los dos grupos a clasificar(diabetes=neg o diabetes=pos)

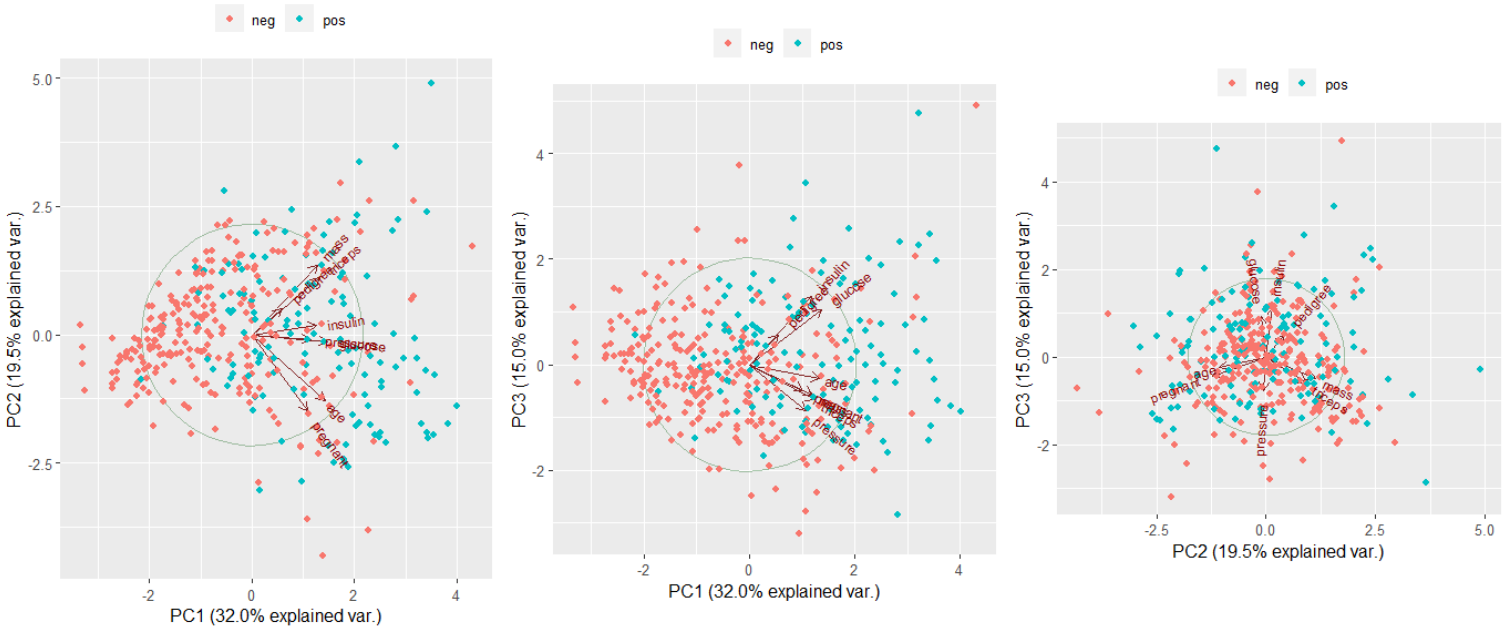

Figura 4: Diagramas de dispersión entre los primeros tres componentes principales

En la Figura 4 notamos que a mayor valor en el componente principal 1 (PC1), tenemos mayores valores para todas las variables.

Notamos que a mayor valor en el componente principal 2 (PC2), se tienen mayores valores para las variables mass, pedigree, triceps e insulin, sin embargo se tienen valores menores en las variables pressure, glucose, age y pregnant

Se observa a mayor valor en el componente principal 3 (PC3), se tienen mayores valores para las variables pedigree, insulin y glucose, sin embargo se tienen valores menores en las variables age, pregnant, pressure, triceps y mass.

Cuadro 3: Análisis de las métricas de poder predictivo mediante la aplicación de diversos métodos y modelos.

Esquema_Entrenamiento	Tuneo	accuracy	recall	specificity
Reg. Log. Efectos principales	No	0.7782051	0.5630769	0.8857692
Reg. Log. efectos principales, iteraciones y cuadrados	No	0.8220513	0.6553846	0.9053846
Reg. Log. con selección usando lasso (K-CV)	Valor Lambda	0.7800000	0.5584615	0.8907692
G.L.M. selección con liga probit	No	0.7661538	0.5107692	0.8938462
naive classifier	No	0.7661538	0.6384615	0.8300000
LDA continuo, considerando variables binarias	No	0.7661538	0.6384615	0.8300000
QDA continuo, considerando variables binarias	No	0.7733333	0.6161538	0.8519231
Modelo K-NN (Categorico)	Valor K	0.7592308	0.5307692	0.8734615
Random Forest (200 árboles)	Valor mtry	0.7735897	0.5961538	0.8623077

De acuerdo con las métricas presentadas en el Cuadro 3 , se observa que los valores más elevados corresponden a la columna “specificity”. Esto indica una mejora en la capacidad de clasificación predictiva para ese grupo, siendo el modelo “Reg. Log. efectos principales, iteraciones y cuadrados” el que exhibe el valor más alto.

No obstante, es relevante señalar que este mismo modelo también ostenta la mejor métrica predictiva global, con una tasa del 82 %. Sin embargo, es pertinente destacar que la métrica “recall” muestra un valor relativamente bajo. En consecuencia, sería prudente considerar la exploración de otras métricas o la incorporación de parámetros adicionales en este esquema.

Por este motivo, se ha optado por el esquema denominado “Reg. Log. efectos principales, iteraciones y cuadrados”, dado que este exhibe las tasas más elevadas de poder predictivo en las métricas calculadas. En este sentido, se observa una métrica predictiva global con una tasa del 82 %, una métrica predictiva para el primer grupo con una tasa del 65 %, y una tercera métrica predictiva para el segundo grupo con una tasa del 90.5 %.