2. Clasificación supervisada

En el conjunto de datos `PimaIndiansDiabetes2` de la biblioteca `mlbench`, sin considerar NA´s, tenemos la variable `diabetes` con 2 grupos para hacer el estudio, estos son "pos" y "neg" que indican que la persona tiene diabetes y que no tiene, respectivamente. Tenemos 262 "neg" y 130 "pos", lo cual nos indica que es mayor el número de personas que no tiene diabetes. En el siguiente Cuadro, mostramos algunas estadísticas descriptivas de las variables numéricas del conjunto de datos, tales como pregnant, glucose, pressure, triceps, insulin, mass, pedigree y age.

Cuadro 1:

Statistic	N	Mean	St. Dev.	Min	Max
pregnant	392	3.301	3.211	0	17
glucose	392	122.628	30.861	56	198
pressure	392	70.663	12.496	24	110
triceps	392	29.145	10.516	7	63
insulin	392	156.056	118.842	14	846
mass	392	33.086	7.028	18.200	67.100
pedigree	392	0.523	0.345	0.085	2.420
age	392	30.865	10.201	21	81

Por otra parte, en la siguiente Figura podemos notar en los BoxPlot que las medianas de las variables numéricas para las personas que si tienen diabetes son mas altas, ademas este grupo presenta mas variabilidad o dispersión y sesgo a la derecha de acuerdo con las densidades. Es natural ver que el número de veces que se ha estado embarazada (pregmant) está positivamente correlacionado con la edad, y que a mayor índice de masa corporal (mass) mayor es el grosor de los triceps.

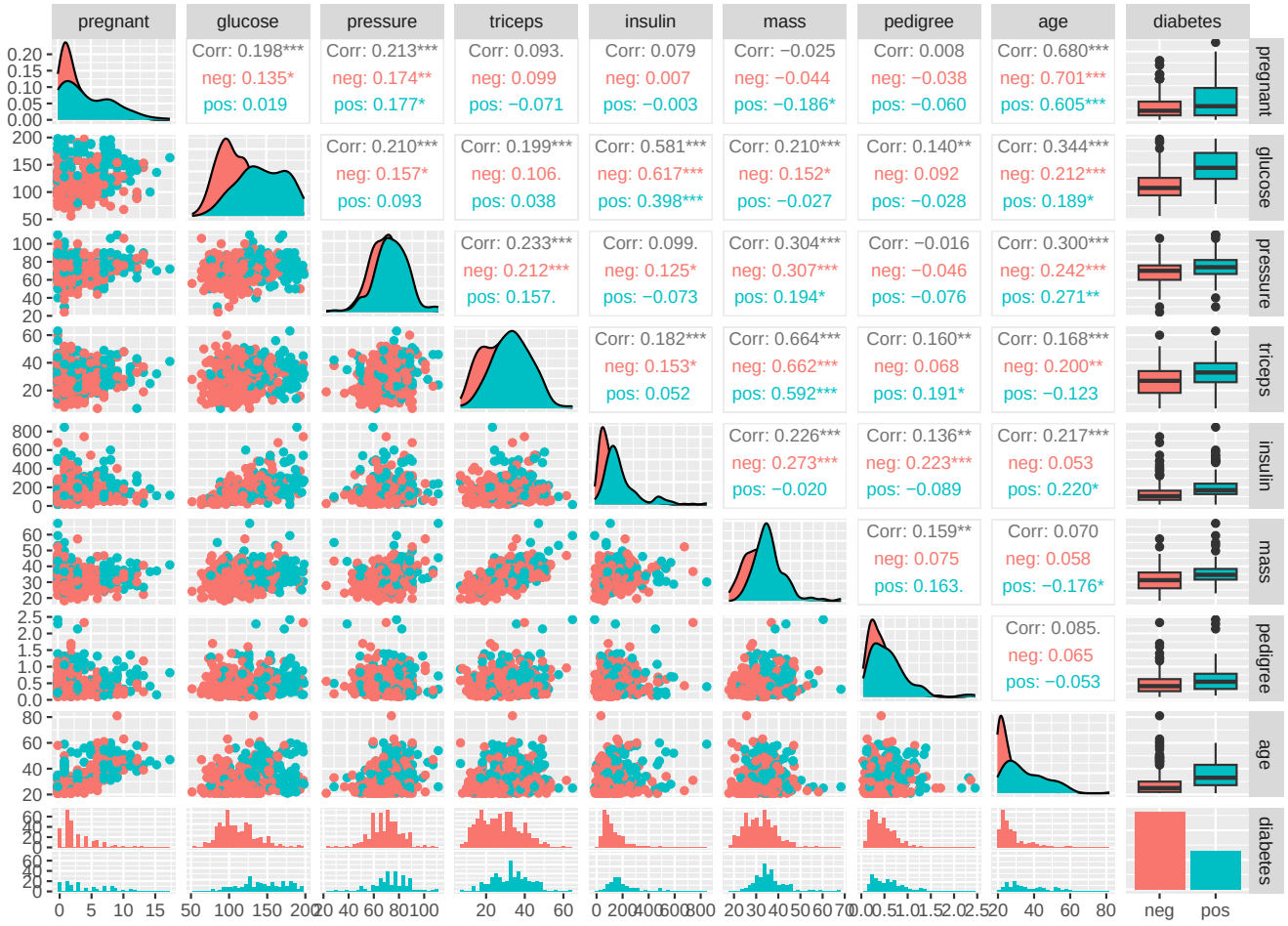

Figura 1: Variables predictoras por grupos de respuesta

Lo que sigue es obtener las componentes principales con el fin de obtener un mejor modelo y que ademas nos servirá para desechar variables que no aporten tanta información al mismo. Ya que se obtuvieron las componentes principales obtuvimos lo siguiente:

Hasta la componente 4 conserva un 78 % de variabilidad de los datos originales por lo que usaremos estas para el análisis, ademas veremos la correlación que tiene cada componente con las variables originales.

Para el caso de la componente 1 obtuvimos que las variables con las que esta mas correlacionada son glucose, triceps y age, arriba de 0.6 para los 3 casos.

Luego para la segunda componente se obtuvo que las variables con las que esta mas correlacionada son pregnant y age, aunque en este caso la correlación es negativa, de ahí se tiene que con mayor correlación positiva se encuentran mass y pedigree.

Con la tercer componente se tiene que las variables con las que esta mas correlacionada fueron glucose e insuline.

Por ultimo la cuarta componente esta mas correlacionada con pedigree.

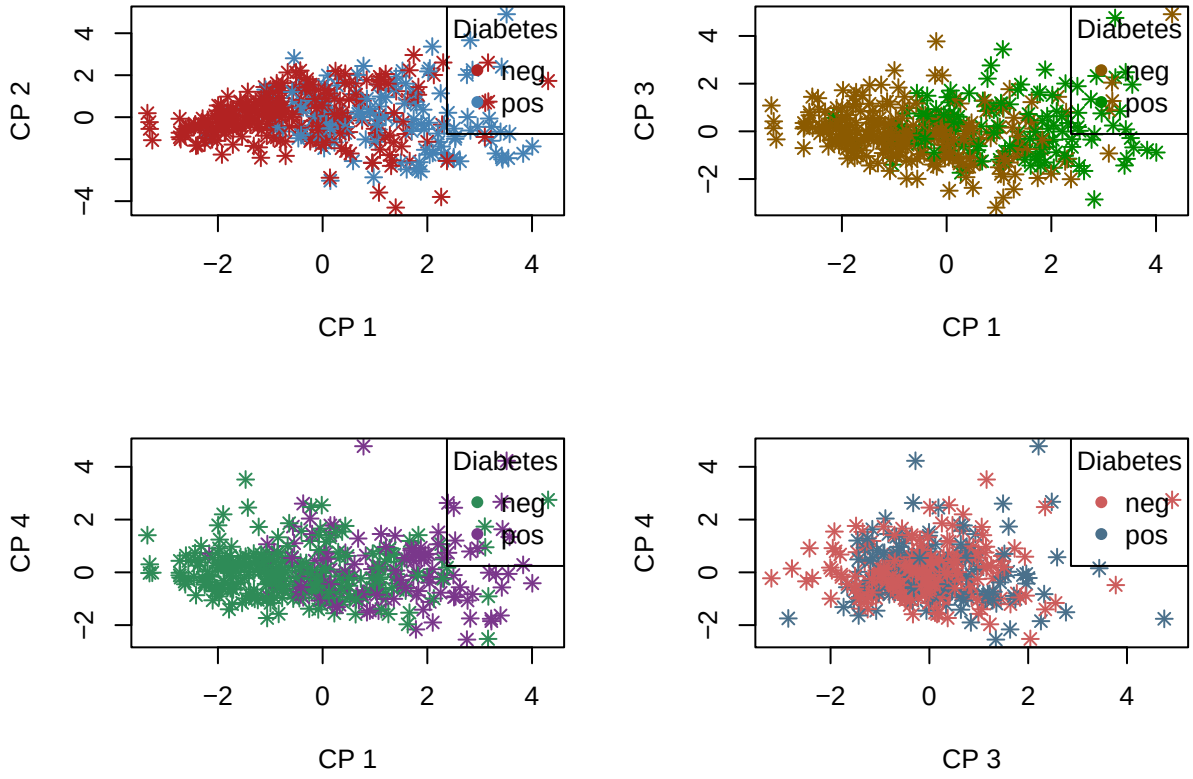

Figura 2: Componentes principales por grupos

Pasando a los modelos se decidió que se ajustaran los siguientes:

1.Regresión logit con efectos principales

2.Regresión logit con interacciones, variables al cuadrado y selección de variables con el método del mejor subconjunto.

3.Regresión logit con selección de variables y el método por pasos both.

4.Regresión logit con interacciones, variables al cuadrado, selección lasso con lambda tuneado.

5.Naive classifier, LDA, QDA, KNN.

6.Random Forest con 200 arboles tuneado.

Cuadro 2: Entrenamientos

Modelo	accuracy	recall	specificity
Regresión logit con efectos principales.	0.768	0.546	0.875
Regresión logit con interacciones variables al cuadrado, y selección de variables con el método del mejor subconjunto.	0.77	0.554	0.881
Regresión logit con selección de variables y el método por pasos both.	0.773	0.556	0.877
Regresión logit con interacciones variables al cuadrado selección lasso con lambda tuneado.	0.775	0.535	0.892
Naive classifier	0.77	0.634	0.835
LDA	0.77	0.54	0.881
QDA	0.758	0.58	0.844
KNN	0.745	0.487	0.871
Random Forest con 200 arboles tuneado.	0.745	0.549	0.848

De los resultados obtenidos por cada modelo registrado en la tabla con sus métricas correspondientes podemos decir que en términos de "accuracy" o precisión el modelo con mayor puntuación fue el modelo de regresión logit con interacciones, variables al cuadrado y selección lasso con lambda tuneado con un valor de 0.775, sin embargo si observamos la tabla también existen otros 2 modelos que tuvieron una muy buena puntuación y que no se quedan tan atrás del modelo lasso, por ejemplo, el modelo en el que se uso selección de variables y método por pasos both obtuvo un valor en la métrica "accuracy" de 0.773 el cual esta muy cercano al valor registrado para el método lasso.No obstante dado que en la base de datos y gracias al análisis descriptivo que se hizo al comienzo notamos que existen mas valores clasificados en la categoría de negativos por lo que usar la métrica "acurracy" traería problemas al momento de clasificar, por otra parte el modelo lasso siguió siendo el mejor puntuado en la métrica "specificity" con un valor de 0.89, en este aspecto si lo que se busca es reducir la cantidad de falsos negativos este modelo nos sera de mucha ayuda ya que en dado caso de tener nuevas observaciones que sean clasificadas en la categoría negativo tendremos una buena clasificación.

Para el caso de la métrica "recall" tenemos como mejor modelo el "Naive classifier" con un valor de 0.634.

De todos los modelos aquí presentados algo que notamos es que los coeficientes que mayor efecto tuvieron en el diagnostico de la diabetes fueron pedigree, glucosa, insulina y edad, lo cual nos hace bastante sentido ya que es bien sabido en el ámbito medico que la genética siempre va a ser un factor determinante en el desarrollo de ciertas enfermedades de distinta índole no solamente hablando de desarrollar diabetes, adicionalmente a la genética el nivel de insulina en cada persona y su edad son otros factores que influyen directamente en el desarrollo de diabetes incluso seria interesante ver como influiría el genero de una persona para el desarrollo o no de esta enfermedad.