Universidad del Valle
Escuela de Estadística
Programa Académico de Pregrado en Estadística
Técnicas de Minería de Datos y Aprendizaje Automático
Cali, julio de 2023

Laboratorio 4

Aprendizaje Supervisado - Regresión

Yeimy Tatiana Marín[a], Miguel Enriquez[b]

1. Análisis exploratorio

1.1 Descripción de los datos

Los datos contienen 442 observaciones de pacientes con diabetes sobre 11 variables. La variable dependiente es la progresión de la enfermedad (Y)
Las variables independientes son: Edad, Sexo, Índice de masa corporal, Presión arterial promedio y seis mediciones de suero sanguíneo.
Al hacer una inspección preliminar a la hoja de datos **diabetes.txt** se corrobora que no hay datos faltantes (NA), y se modifica la estructura de la variable cualitativa **Sex** como factor.

1.2 Estadísticas descriptivas

De la siguiente tabla de resumen descriptivo obtenido para cada variable, se observa que en las cuantitativas el valor mínimo es menor al primer cuartil, el valor máximo es mayor al tercer cuartil y la media es levemente mayor que la mediana, infiriendo de lo anterior, posible presencia de datos atípicos (valores extremos) y una distribución en las observaciones sesgada a la derecha. La variable factor SEX muestra que el sexo correspondiente a "1" es levemente mayor al "2".

AGE	**SEX**	**BMI**	**BP**	**S1**	**S2**
Min: 19	**1:** 235	**Min:** 18	**Min:** 62	**Min:** 97	**Min:** 41.60
1st Qu: 38.25	**2:** 207	**1st Qu:** 23.20	**1st Qu:** 84	**1st Qu:** 164.2	**1st Qu:** 96.05
Mediana: 50		**Mediana:** 25.70	**Mediana:** 93	**Mediana:** 186	**Mediana:** 113
Media: 48.52		**Media:** 26.38	**Media:** 94.65	**Media:** 189.1	**Media:** 115.44
3rd Qu: 59		**3rd Qu:** 29.27	**3rd Qu:** 105	**3rd Qu:** 209.8	**3rd Qu:** 134.50
Max: 79		**Max:** 42.20	**Max:** 133	**Max:** 301	**Max:** 242.40
S3	**S4**	**S5**	**S6**	**Y**	
Min: 22	**Min:** 2	**Min:** 3.258	**Min:** 58	**Min:** 25	
1st Qu: 40.25	**1st Qu:** 3	**1st Qu:** 4.277	**1st Qu:** 83.25	**1st Qu:** 87	
Mediana: 48	**Mediana:** 4	**Mediana:** 4.620	**Mediana:** 91	**Mediana:** 140.5	
Media: 49.79	**Media:** 4.07	**Media:** 4.641	**Media:** 91.26	**Media:** 152.1	
3rd Qu: 57.75	**3rd Qu:** 5	**3rd Qu:** 4.997	**3rd Qu:** 98	**3rd Qu:** 211.5	
Max: 99	**Max:** 9.09	**Max:** 6.107	**Max:** 124	**Max:** 346	

Tabla 1: Resumén de las variables

[a]Código: 1524344. E-mail: yeimy.marin@correounivalle.edu.co
[b]Código: 2023796. E-mail: miguel.enriquez@correounivalle.edu.co

En la Figura 1, se observa la presencia de datos atípicos (valores extremos) para las variables BMI, S1, S2, S3 y S6, apoyando visualmente el resumen descriptivo de las variables. Además, de la variable AGE se tiene que entre los 40 a los 60 años se encuentran la mayoría de los pacientes diabéticos.

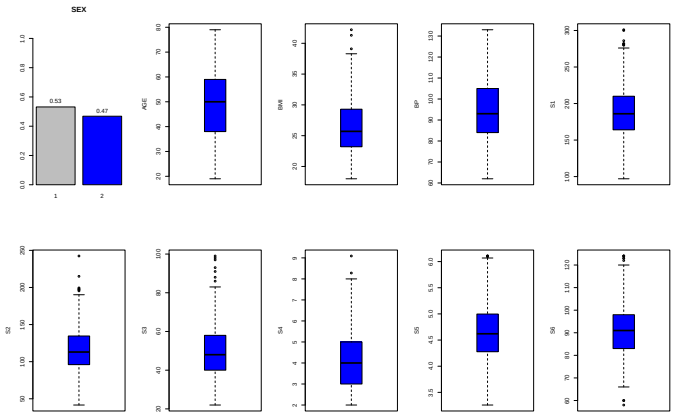

FIGURA 1: Boxplot. visualización gráfica de las variables Cuantitativas

Al observar la relación bivariada con la variable de respuesta Y en la Figura 2, se infiere una fuerte relación relativa entre las variables BMI, BP, S4, S5, y en la figura 3, se observan las variables S1, S2, S3, S4 y S5 altamente correlacionadas entre dos o más variables predictoras de estas, indicando posible presencia de multicolinealidad para modelos de regresión.

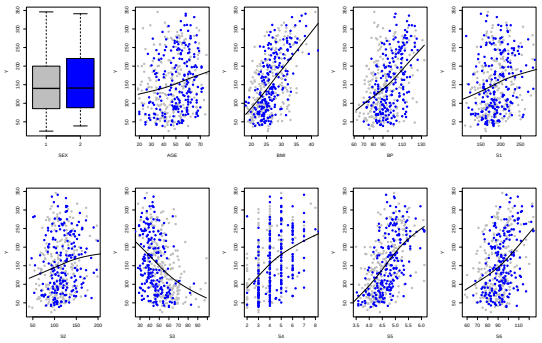

FIGURA 2: Visualización bivariada. X's vs Y

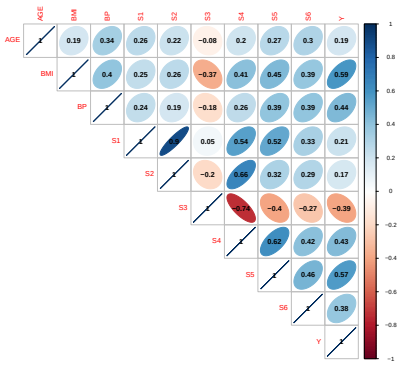

FIGURA 3: Matriz de correlación entre las variables

1.3. Visualización Multivariada apoyada en Componentes Principales

En las figuras 4:7 de representación simultánea, se analiza que la variable de respuesta esta relacionada con las variables BMI y BP, al igual que el grupo de variables S1 y S2, y el grupo de variables S6 y S5. También que aproximadamente la mitad de los pacientes diabéticos independiente de su sexo están siendo representados por las variables y la otra mitad no presenta estás características, siendo mejor representadas en conjunto en el plano 2-3. Por último, se observa que el sexo no es un factor clave en la determinación de la diabetes.

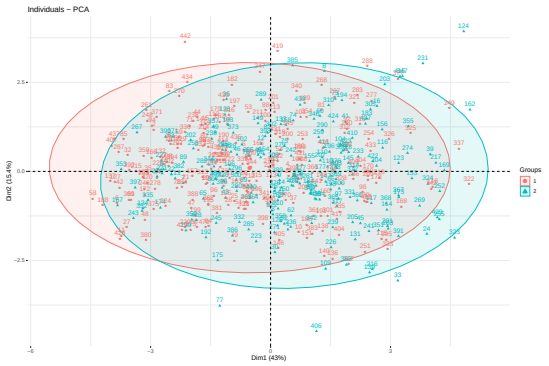

FIGURA 4: Agrupación de diabéticos por Sexo 1 o 2

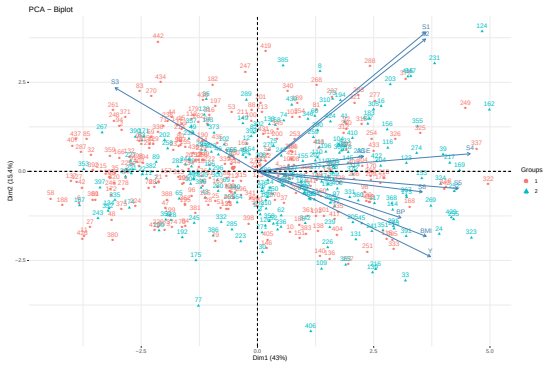

FIGURA 5: Representación simultánea en el plano 1-2

FIGURA 6: Representación simultánea en el plano 1-3

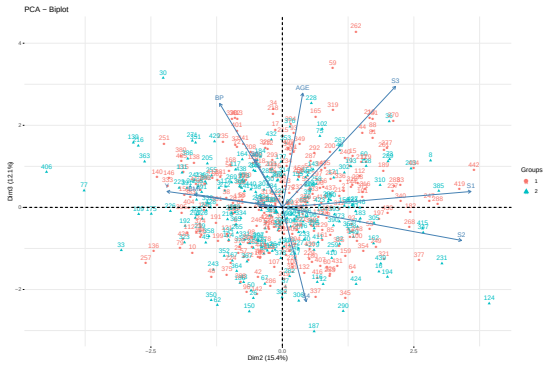

FIGURA 7: Representación simultánea en el plano 2-3

2. Selección de las variables - Modelo Regresión Lineal Múltiple (MRL).

De acuerdo al análisis exploratorio, se detecta presencia de multicolinealidad debido a la alta correlación entre dos o más variables predictoras.
Por tanto, con el fin de conocer las variables necesarias para explicar el Modelo de Regresión Lineal Múltiple (MRL) y abordar la multicolinealidad en el problema estudiado, se utilizó el método AIC con el algoritmo Backward que permite eliminar una a una las variables altamente correlacionadas en el modelo de regresión.
Se inicia con el ajuste de un modelo saturado, es decir, modelo con todas las variables, en el cuál se observa que AGE, S1, S2, S3, S4 y S6 no son significativas al modelo, y que las variables S1, S2, S3 son las que más presentan multicolinealidad. Luego, con el método de eliminación hacia atrás (Backward), se obtiene el modelo reducido que mejor explica los datos con el mínimo número de parámetros y con el mínimo AIC.

Modelo ajustado de MRL.

$$Y = \beta_1 SEX + \beta_2 BMI + \beta_3 BP + \beta_4 S1 + \beta_5 S4 + \beta_6 S5 \tag{1}$$

3. Entrenamiento y comparación por validación cruzada - Modelos de Regresión.

Con el fin de emplear los modelos de regresión, se realiza la selección del porcentaje para dividir los datos en dos conjuntos más pequeños: 80 % datos de entrenamiento (trainig) y 20 % datos de prueba (test). El subconjunto de datos de entrenamiento será utilizado para estimar los parámetros del modelo y el subconjunto de datos de test se empleará para comprobar el comportamiento del modelo estimado.

3.1 Modelo Estadístico. Regresión parcial por mínimos cuadrados (PLS1).

Se crea el modelo con un número máximo de componentes 8 y con validación cruzada simple (k=10 pliegues por defecto). Luego, por la prueba Raíz del error (RMSEP) se determina que para este problema la cantidad de componentes cp=2 identifican las variables más relevantes en el modelo PLS1 y evitan presencia de multicolinealidad.
Observando que las 2 Componentes explican el 50.12 % de la variabilidad de Y, y las variables predictoras con mayor importancia en el modelo PLS1 son: BMI, BP, S3, S4, S5 y S6.

3.2 Modelo computacional. Árbol de regresión (CART).

Se ajuste un modelo de árbol de decisión para evaluar el número de nodos adecuados, estableciendo un punto de corte o parámetro cp en 0.0001 para controlar la complejidad del árbol y evitar un sobreajuste. Ahora, por medio de validación cruzada se selecciona el punto óptimo de equilibrio que controla la complejidad del árbol y su capacidad para ajustarse a los datos, en este caso, de acuerdo al error de validación cruzada obtenido se tiene un cp=0.02135193. Luego, se procede a podar el árbol con el cp óptimo escogido, obteniendo por ser mayor el valor de CP árbol resultante con n=5, es decir, más simple. Se observa que las variables con su respectivo punto de corte (Atributo) que mejor divide los datos son: BMI las variables predictoras con mayor importancia en en son: BMI, S5 y S6.

3.3 Modelo computacional. K vecinos cercanos (KNN).

Se utiliza la librería caret para realizar validación cruzada y tuneo de hiperparámetros, estableciendo una estructura de validación cruzada con k-fold repetido 10 veces y ejecutando una validación con tuning para k=1:15. Obteniendo por R^2 que el mejor k=10.

3.4 Modelo computacional. Maquina de soporte vectorial (SVM).

Se ajusta un modelo SVM definiendo los rangos de valores que se probarán para los hiperparámetros **epsilon** y **cost** del modelo SVM, donde **epsilon** tomará valores desde 0.1 hasta 0.5 en incrementos de 0.05 y **cost** tomará valores de 2^0 hasta 2^5. Luego, por Validación cruzada (10 fold repetido 10 veces) se selecciona el mejor modelo encontrado por la combinación óptima de los hiperparámetros **epsilon= 0.4** y **cost= 1** que producen el mejor rendimiento de este modelo.

3.5 Modelo de ensamble. Random Forest (RF).

Se ajusta un modelo de Random Forest por validación cruzada (10 fold repetido 10 veces) con el fin de tunear el hiperparámetro mtry=2:6 para definir el número de variables predictoras del proceso de aleatorización. De lo anterior, se obtiene el modelo RF ajustado con 500 árboles de decisión y mtry=3 variables predictoras según su importancia (BMI, S5 y BP).

3.6 Comparación bondad ajuste.

De acuerdo a las medidas de desempeño ICC (Índice de Correlación Intraclase) y R^2 se realizó una tabla (ver tabla 2) entre modelos para evaluar su bondad de ajuste. Obteniendo que el mejor predictor (Algoritmo en este caso) que al registro de observaciones X (con datos trainig) asigna un mejor pronóstico a la variable respuesta Y (progresión de la enfermedad de diabetes) es el de Maquina de soporte vectorial (SVM).

Modelos	**Rsquared**	**ICC**
MRL	0.507	0.673
PLS1	0.501	0.668
CART	0.473	0.642
KNN	0.425	0.555
SVM	0.633	0.761
RF	0.445	0.608

Tabla 2: Bondad de ajuste. Datos de entrenamiento del modelo

Ahora, con las medidas de desempeño ICC (Índice de Correlación Intraclase) y R^2 se realiza una tabla (ver tabla 3) entre los mismos modelos para evaluar su bondad de ajuste con datos de prueba (test) obteniendo que el mejor modelo es el de Regresión Lineal Múltiple (MRL)

Modelos	**Rsquared**	**ICC**
MRL	0.528	0.683
PLS1	0.515	0.666
CART	0.489	0.635
KNN	0.267	0.402
SVM	0.479	0.638
RF	0.484	0.625

Tabla 3: Bondad de ajuste. Datos de prueba del modelo

4. Ventajas y Desventajas de los Modelos

Partiendo de que cada modelo tiene una construcción teórica distinta, una mejor capacidad interpretativa, algunos predictiva y otros sirven para implementar en conjunto de datos extremadamente grandes. Se describe a continuación, los usos, fundamentos teóricos y las ventajas de cada uno de los modelos utilizados anteriormente.

Modelo	Enfoque	Fundamento teórico	Ventajas
Regresión Lineal Múltiple	Se utiliza para predecir una variable continua a partir de múltiples variables predictoras.	Se basa en supuestos estadísticos, como linealidad, independencia, homocedasticidad y normalidad de los errores.	Adecuado cuando existe una relación lineal entre las variables y se cumplen los supuestos de regresión lineal.
PLS1 (Partial Least Squares)	Utiliza una técnica de regresión que busca maximizar la covarianza entre las variables predictoras y la variable de respuesta.	Se basa en la descomposición de matrices y la reducción de dimensionalidad.	Útil cuando hay multicolinealidad entre las variables predictoras y se desea una reducción de dimensionalidad.
Árbol de Regresión	Utiliza un enfoque de partición recursiva del espacio de características para predecir una variable continua.	Se basa en la estructura jerárquica de árbol y la división del espacio de características en subregiones.	Adecuado para problemas con relaciones no lineales y cuando la interpretación en términos de reglas de decisión es importante.
KNN (K-Nearest Neighbors)	Utiliza la similitud entre observaciones para hacer predicciones.	Se basa en la idea de que observaciones similares tienen etiquetas similares.	Útil cuando la estructura local de los datos es relevante y no se requiere una interpretación detallada del modelo.
SVM (Support Vector Machines)	Utiliza vectores de soporte para separar las clases en un espacio de alta dimensionalidad.	Se basa en la idea de encontrar un hiperplano óptimo que maximice el margen entre las clases.	Útil cuando hay una separación clara entre las clases y se busca un buen rendimiento de clasificación.

Tabla 4: Resumén teórico de los modelos