.UBAfiuba
FACULTAD DE INGENIERÍA

Trabajo Práctico Final

Aprendizaje Estadístico
Segundo cuatrimestre 2020

Clasificación de mujeres con diabetes

Romero, Adrián

Índice

1. Introducción

El objetivo del presente trabajo es construir modelos capaces de clasificar a pacientes en diabéticos o no diabéticos. Para ello se utilizará un set de datos que tiene información acerca de diferentes factores médicos, como por ejemplo el índice de masa corporal, la cantidad de glucosa en sangre, la cantidad de insulina, entre otros. Además se buscarán cuáles son las variables mas relevantes para realizar la clasificación y qué métrica puede resultar conveniente para evaluar el desempeño de los modelos desarrollados. En particular a lo largo del trabajo, se desarrollaron los siguientes modelos: Árbol de decisión, Bagging, KNN y LDA/QDA.

El desarrollo del trabajo estará dividido en las siguientes secciones:

- **Análisis inicial**: se analizará el set de datos que se utilizará a lo largo del trabajo. En particular se evaluará si las variables predictoras están correlacionadas y cómo éstas se relacionan con la variable a predecir. Para ello se realizan distintos gráficos que permitan visualizar estas relaciones. Además en esta sección se hace énfasis en encontrar las variables que realizan una mejor división entre las clases.
- **Modelos**: se buscarán diferentes modelos de clasificación para resolver el problema y se escogerá entre estos modelos, el mejor según alguna métrica. Además cada uno de los modelos estará brevemente precedida por una explicación sobre el funcionamiento de ese modelo especifico. También se realizan variantes de los modelos que utilizan únicamente los features que se consideran mas relevantes según el análisis inicial. Para la evaluación de las métricas se utiliza siempre cross-validation.
- **Conclusiones**: Finalmente, en la conclusión, se resumirán los aspectos más destacados del trabajo y se propondrán potenciales mejoras a los modelos hallados.

2. Análisis inicial

El dataset con el que se estará trabajando a lo largo del trabajo proviene del "National Institute of Diabetes and Digestive and Kidney Diseases". Este dataset es un subconjunto de un dataset mucho mayor. En particular, se impuso como restricción que los pacientes a tener en cuenta son mujeres de 21 años de ascendencia Pima India.

A continuación se muestran las primeras entradas del dataset:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figura 1: Primeras entradas del dataset

Como se puede ver, las variables que forman parte del dataset son las siguientes:

- **Pregnancies**: cantidad de embarazos del paciente
- **Glucose**: cantidad de glucosa en sangre
- **BloodPressure**: presión sanguínea
- **SkinThickness**: grosor de la piel
- **Insuline**: cantidad de insulina en sangre
- **BMI**: indice de masa corporal, es una relacion entre el peso y la altura
- **DiabetesPedigreeFunction**: score que mide la probabilidad de tener diabetes basado en historia familiar
- **Age**: edad
- **Outcome**: 0 si la persona no es diabética, 1 si lo es

A priori se espera que la variable **glucosa** sea muy relevante a la hora de realizar la clasificación en personas diabéticas y no diabéticas De hecho, la diabetes surge cuando la cantidad de glucosa en sangre es eleveda.

El dataset contiene información acerca de 768 mujeres, 500 de las cuales no son diabéticas y 268 si lo son, por lo que las clases en el dataset no están balanceadas.

Se procede a realizar boxplots con los valores que toman las variables, clasificando previamente estos valores según si el paciente es diabético o no lo es:

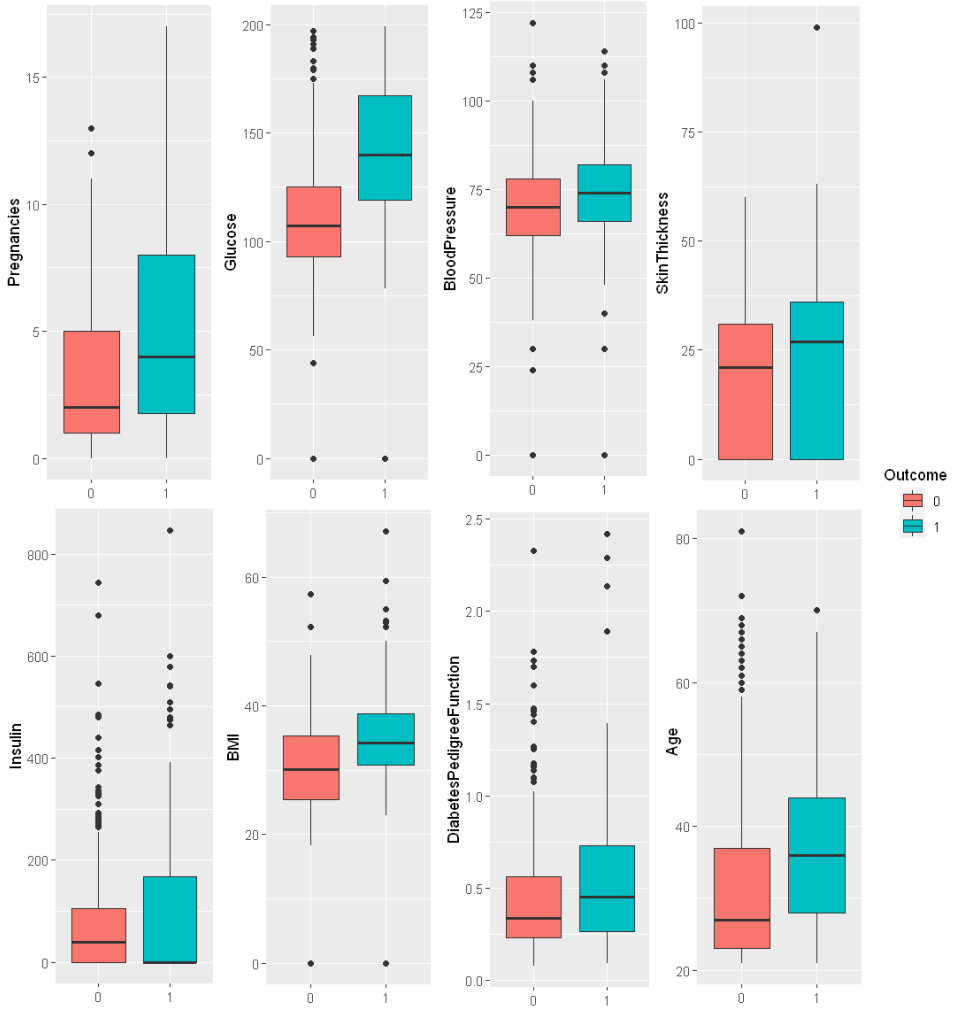


Figura 2: Boxplots

Se pueden realizar las siguientes observaciones a partir del boxplot anterior:

1. La cantidad de glucosa en sangre es significativamente mayor en las personas con diabetes que en las personas sin diabetes. El 75 % de las mujeres sin diabetes tiene un nivel de glucosa menor a 125 mientras que un 75 % de las mujeres con diabetes tiene un nivel de glucosa mayor a 120.

2. La mediana de la insulina en sangre para las personas con diabetes es aproximadamente 0. Sin embargo hay personas diabéticas que tienen una cantidad de insulina relativamente normal o incluso elevada. La insulina es una hormona que secreta el páncreas y regula la cantidad de glucosa en sangre, por lo que es posible que una persona diabética tenga baja insulina. Sin embargo, una persona diabética puede tener también altos niveles de insulina. Esto sucede cuando la persona en cuestión desarrolló resistencia a la insulina: su cuerpo detecta que tiene azúcar en sangre y genera insulina para regularla, sin embargo la insulina no logra su función y continua habiendo glucosa en sangre lo que dispara aun mas producción de insulina. Esto explicaría porque la mediana de la cantidad de insulina es menor en el caso de personas diabéticas pero el tercer cuartil es mayor.

3. La Función de Pedigree de Diabetes no presenta muchas diferencias para personas diabética y no diabéticas. Es levemente mayor en el caso de las personas con diabetes lo cual es esperado. Es posible que estas mujeres tengan un historial familiar similar, al ser todas mujeres de ascendencia Pima India y esto se vea reflejado con poca variabilidad entre los boxplots de mujeres diabéticas y no diabéticas.

4. Como es de esperarse, los valores de cantidad de embarazos, edad y BMI aumentan para personas diabéticas, estas variables también parecen tener diferencias notables entre las personas diabéticas y no diabéticas. La presión sanguínea y el grosor de la piel también aumentan para personas diabéticas, pero mas levemente.

En la siguiente tabla se muestran las correlaciones entre las variables:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54	0.22
Glucose	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26	0.47
BloodPressure	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24	0.07
SkinThickness	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11	0.07
Insulin	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04	0.13
BMI	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04	0.29
DiabetesPedigreeFunction	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03	0.17
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00	0.24
Outcome	0.22	0.47	0.07	0.07	0.13	0.29	0.17	0.24	1.00

Figura 3: Correlaciones

Se pueden realizar las siguientes observaciones de la tabla anterior:

1. Las variables no están muy correlacionadas entre si, en general las correlaciones son menores a 0.3.

2. Como es esperado, existe una correlación positiva entre edad y cantidad de embarazos.

3. Insulina y glucosa están correlacionados positivamente.

4. La variable mas correlacionada con el resultado o outcome es la glucosa en sangre y la siguiente el BMI

Finalmente, se realizan scatter plots de las variables para lograr encontrar algún par de variables que divida al menos visualmente a las personas diabéticas de las no diabéticas. En el siguiente gráfico se puede ver lo siguiente:

- Tanto el density plot como los scatter plots de para la variable glucosa muestran que este es el feature que mejor separa las clases

- BMI es la segunda variable que parece separar mejor a las clases. En los scatter plots de la fila correspondiente a BMI se ve que existe una tendencia a que las mujeres con diabetes tengan un BMI mayor que la gente sin diabetes.

Esta última observación motivó la realización de clasificaciones utilizando únicamente estas dos variables y viendo las regiones que quedan delimitadas al utilizar cada algoritmo.

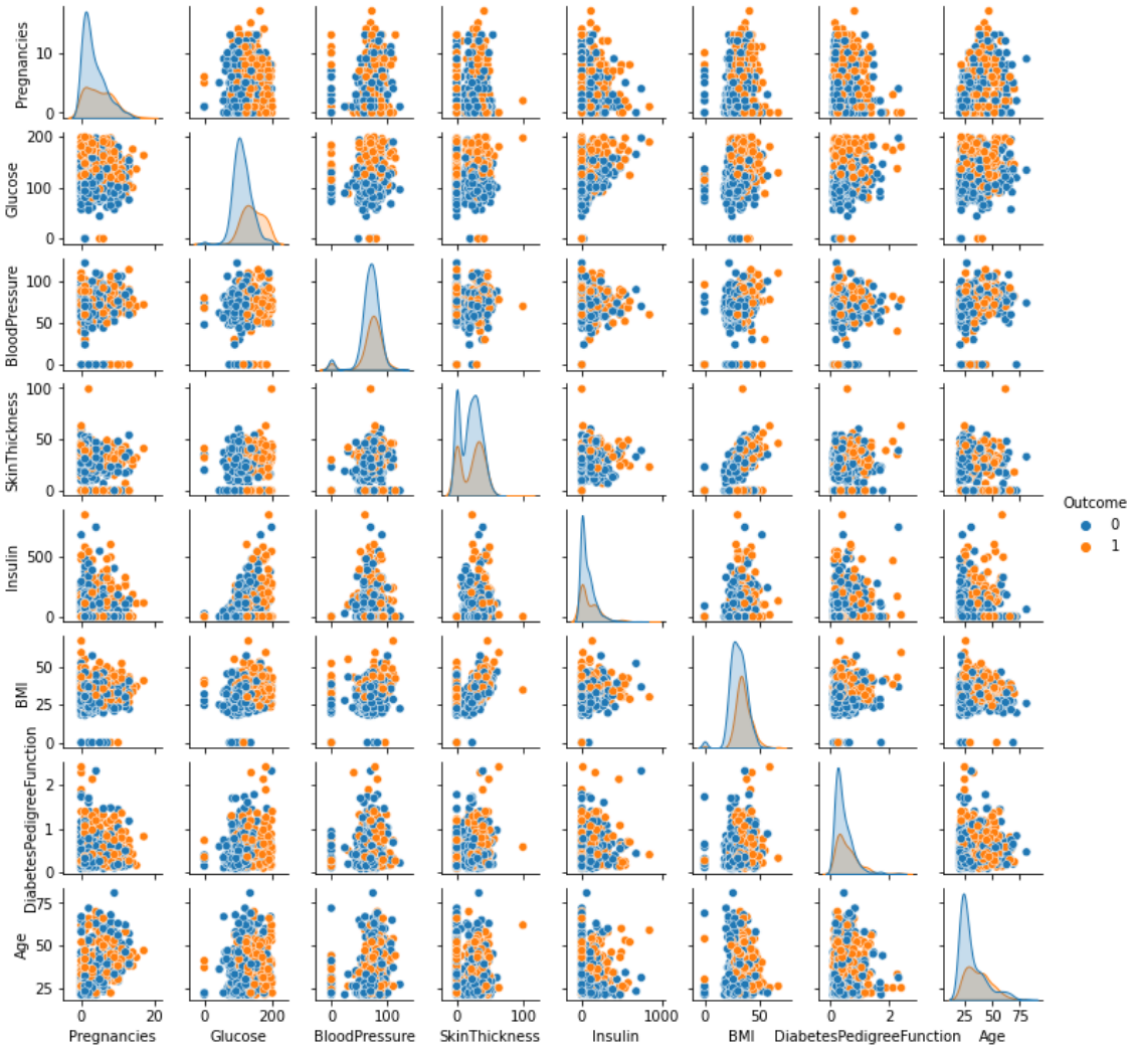


Figura 4: Scatters y density plots

3. Modelos

A continuación se realiza una presentación de los modelos utilizados para resolver el problema de clasificación. Estos fueron: árbol de decisión, bagging, KNN, LDA y QDA Cada modelo, se encuentra precedido de una breve explicación del modelo y de una justificación que motiva la elección del modelo.

Se muestra para cada modelo el valor de las estimaciones halladas por cross-validation para las métricas de:

1. Accuracy
2. Precsision
3. Recall

Además se muestra para los modelos de árbol de decisión, KNN, LDA y QDA como estos dividen el espacio de decisión formado por las variables Glucosa y BMI

3.1. Árbol de decisión

El primer modelo que se desarrolló es un modelo de árbol de decisión. Es un modelo simple y de muy alta interpretabilidad, es por esto que, a pesar de no ser un modelo con mucha precision de predicción, puede servir para proveer información relevante acerca del comportamiento de los datos.

El algoritmo para encontrar un árbol de decisión consiste en segmentar el espacio de predicción en regiones. Para realizar la clasificación de una nueva observación, se busca a que región del espacio de predicción pertenece esta nueva observación y se la clasifica según cual es la clase mayoritaria de las observaciones de la misma región.

Por ejemplo la siguiente es una partición del espacio de predicción en dos clases, una anaranjada y una celeste.

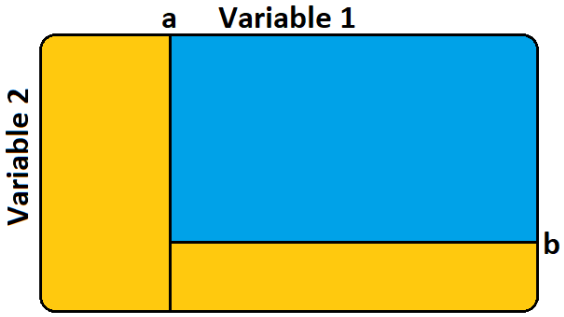


Figura 5: Partición de un espacio de predicción en 2 clases

Clasificar una nueva observación consiste en encontrar a cual de las 3 regiones de la partición anterior pertenece la nueva observación. Para ello se realizan las siguientes preguntas acerca de la nueva observación en el siguiente orden:

- ¿Toma la variable 1 de la nueva observación un valor menor que a? Si lo toma, entonces diremos que la nueva observación pertenece a la clase anaranjada, pues la mayor parte de las observaciones de entrenamiento que pertenecen a esa región eran anaranjadas. De lo contrario se procede a la siguiente pregunta.
- ¿Toma la variable 2 de la nueva observación un valor menor que b? Si lo toma, entonces la nueva observación se clasifica celeste y si no anaranjada.

Para hallar el árbol de decisión es necesario entonces determinar reiteradamente:

1. Cuál es la variable que se utilizará para partir el espacio de predicción
2. Cuál es el valor de la variable que se utilizará como umbral para realizar la división

Para hallar la variable y el valor que lograrán una mejor división del espacio se utilizará el criterio de pureza de Gini, cuya ecuación para el caso de una clasificación binaria puede escribirse como:

$$\mathrm{G} = p_1(1-p_1) + p_2(1-p_2) = 2\ p_1\ p_2 = 2p_1(1-p_1)$$

Donde p_1 es la proporción de observaciones de clase 1 en el set de entrenamiento en la región m. Para cada variable y para cada umbral posible para esa variable se calcula G a partir de como queden distribuidas las observaciones en las regiones correspondientes, luego se selecciona la variable y el umbral que logren un menor valor de G.

Dado que un árbol de decisión profundo tiende a sobreajustar, se tomó la decisión de entrenar un árbol de decisión de profundidad 4. En las siguientes imágenes se puede observar el árbol encontrado (fue dividido por ser muy grande):

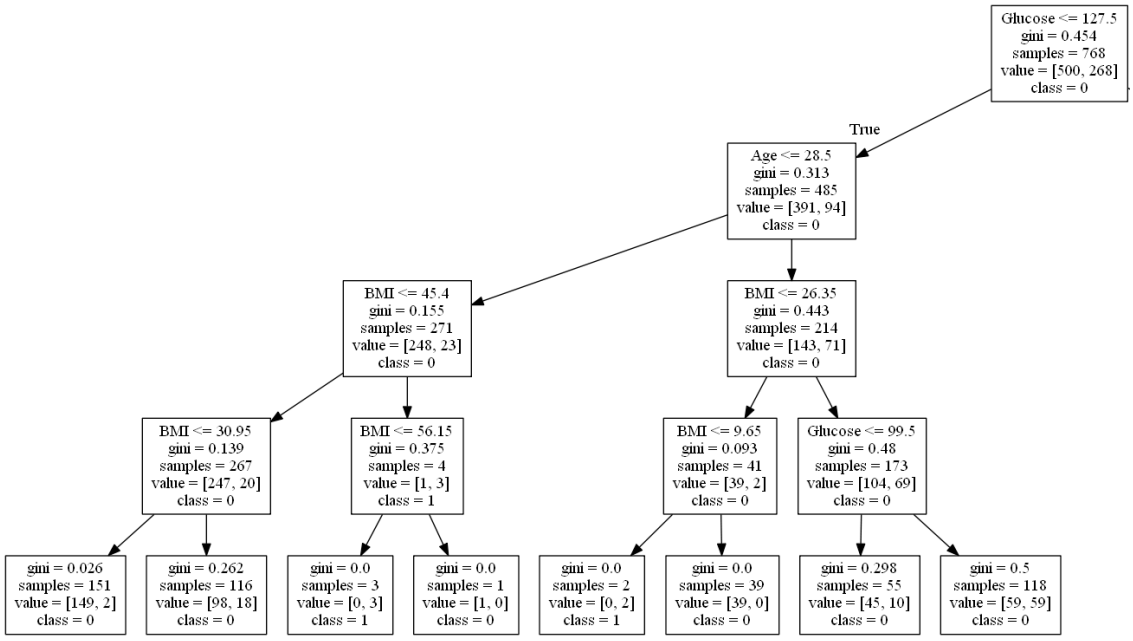


Figura 6: Subarbol izquierdo - Mujeres con glucosa menor a 127

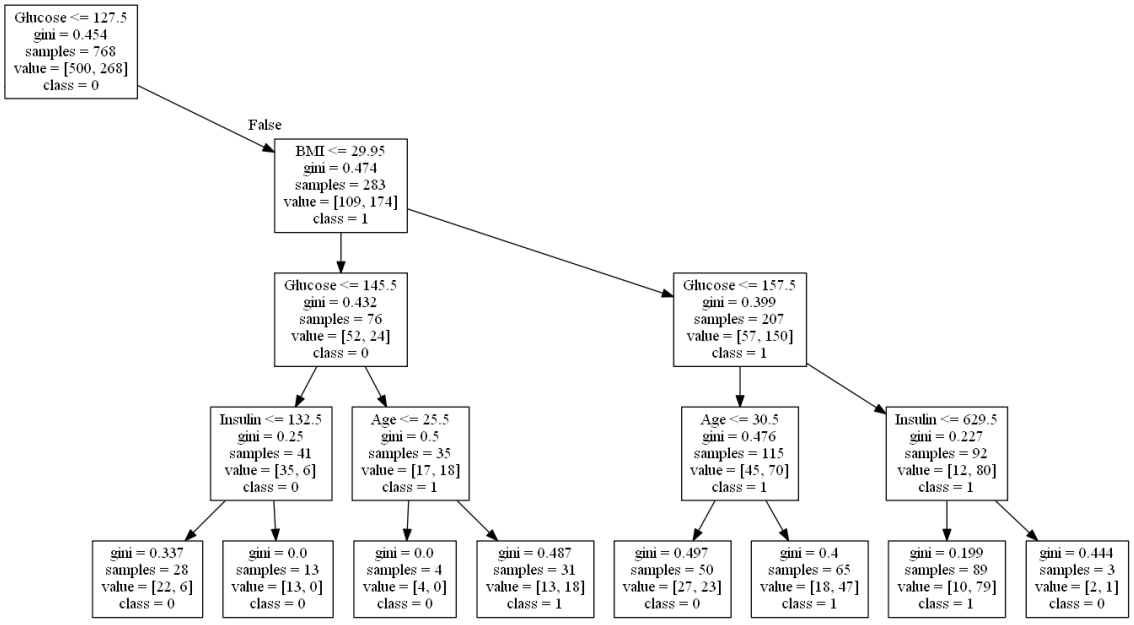


Figura 7: Subarbol derecho - Mujeres con glucosa mayor a 127

Se desprenden las siguientes observaciones del árbol encontrado:

- La primera variable que se utiliza para realizar una división del espacio de predicción es glucosa, y el valor que se utiliza para realizar la división es 127. En el análisis inicial se destacó que para valores similares a este se dividía aproximadamente un 75 % de las personas con diabetes de un 75 % de las personas sin diabetes.
- En el segundo nivel del árbol se utilizan las variables edad y BMI para realizar divisiones. De estas dos, solo BMI logra discriminar personas diabéticas de personas no diabéticas. Además BMI es una variable que aparece reiteradamente en el árbol por lo que es candidata a ser una variable relevante.
- En el árbol solo aparecen 4 variables: glucosa, BMI, edad e insulina.

En los modelos posteriores se encontrarán distintas particiones del espacio de predicción formado por las variables BMI y Glucosa. Por lo tanto, también se muestra aquí, la partición que realiza este árbol sobre el espacio formado por esas variables al considerar solo los 2 primeros niveles del árbol:

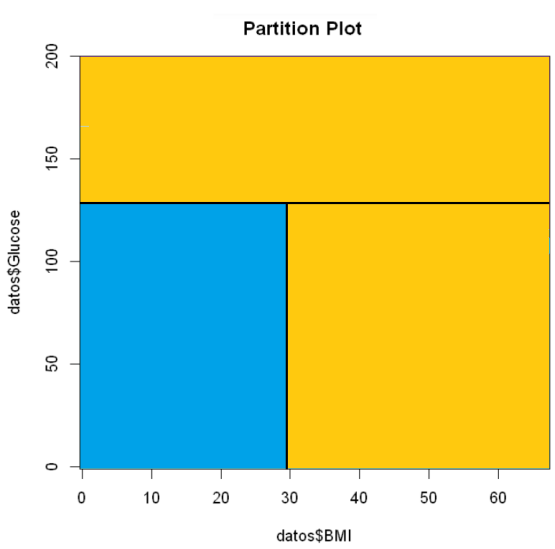


Figura 8: Partition plot - Árbol de decisión

Este árbol fue construido utilizando todo el set de datos como set de entrenamiento lo cual es útil para obtener insights sobre los datos pero no para evaluar la precision de la predicción del modelo.

Para evaluar las métricas de predicción del árbol (y de cualquier modelo) se debe evitar clasificar observaciones que se utilizaron para entrenar, pues es poco representativo de lo que ocurrirá para observaciones que no se hayan visto durante el entrenamiento.

Con el propósito de evaluar las métricas de predicción se utilizará cross-validation y se entrenará un segundo árbol de decisión, también de profundidad 4 y utilizando el criterio de Gini para medir la pureza de las regiones. Se dividirá el dataset en un 90 % para entrenamiento y un 10 % para validación.

Las métricas resultaron:

Árbol de decisión	
Accuracy	0.7
Precision	0.6
Recall	0.375

Los resultados para el árbol así encontrado se pueden resumir en la siguiente matriz de confusión, donde también se observa que la cantidad de falsos negativos es alta, incluso mayor que la cantidad de positivos verdaderos, por eso se obtiene un recall tan bajo.

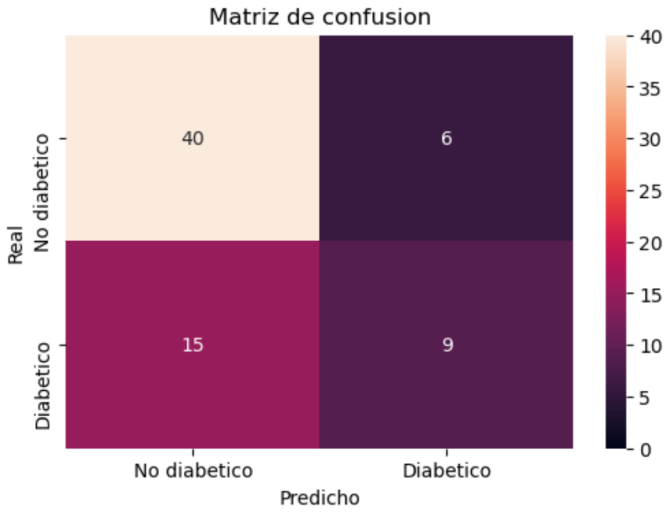


Figura 9: Matriz de confusión - Árbol de decisión

3.2. Bagging

Con el objetivo de obtener una mejor performance en la predicción, se busca entrenar un modelo de Bagging o Bootstrap Aggregation. Este consiste en entrenar varios modelos de árbol de decisión profundos con muestras que se obtienen mediante bootstrap o muestreo con reemplazo. Al predecir, se realiza una votación de las predicciones de cada uno de estos modelos, es decir se predecirá la clase que tenga mayor cantidad de votos.

La idea es que el muestreo con reposición logre disminuir la varianza de los modelos de árbol de decisión.

Los pasos que se utilizaron para construir y evaluar el modelo fueron los siguientes:

1. Dividir el set de datos en train y test.
2. Obtener muestras bootsrapeadas (muestreo con reposicion) del set de train
3. Para cada muestra entrenar un modelo de arbol de decisión
4. Una vez que se tienen todos los arboles de decisión, realizar las predicciones de estos con el set de test
5. Finalmente, realizar una votación de las predicciones de cada modelo para obtener la predicción final.

El siguiente esquema resume los pasos tomados para hallar y evaluar el modelo de bagging.

En particular para este trabajo se desarrollaron 42 modelos de bagging con diferentes cantidad de arboles de decisión y con muestras con reposición que representan distintas proporciones del set de entrenamiento. Los resultados obtenidos se resumen en los siguientes heatmaps:

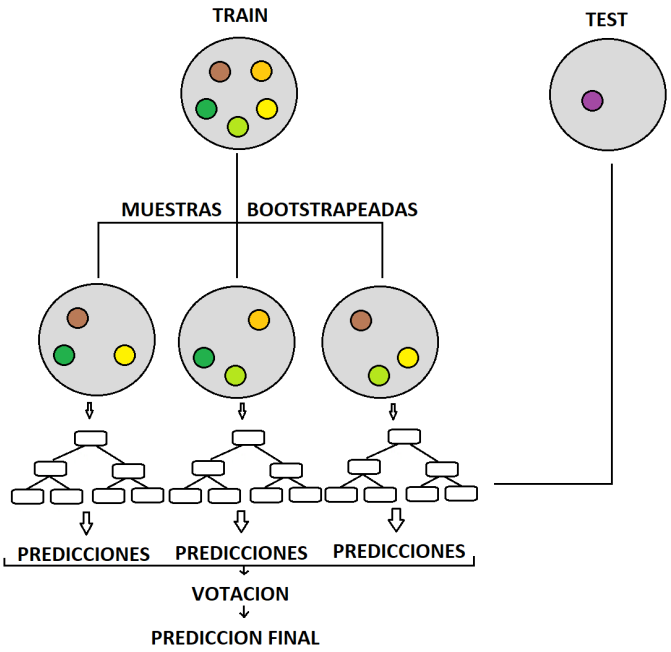
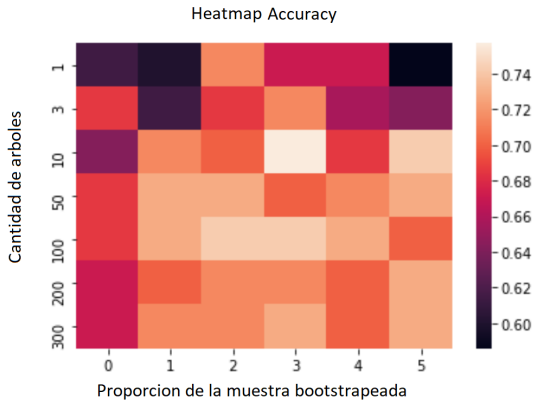
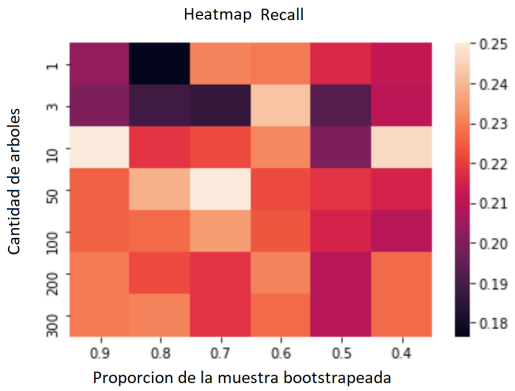
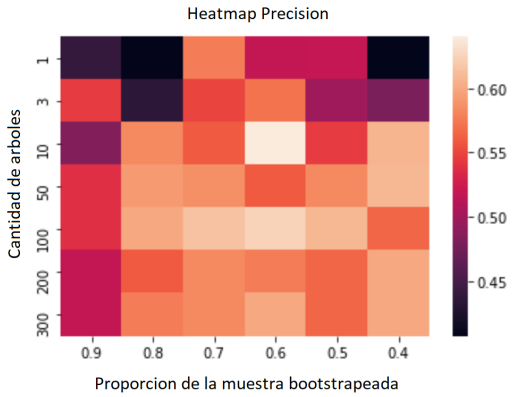


Figura 10: Esquema bagging







Se pueden observar las siguientes tendencias:

- A medida que la cantidad de árboles entrenados aumenta, también lo hacen los valores de las métricas, sin embargo la performance del modelo no parece mejorar significativamente a partir de 10 arboles entrenados.
- A medida que disminuye la proporción de datos tomados como muestra bootstrapeada, la performance del modelo mejora. Esto resulta muy interesante y es posible que se deba a que esto haga que los arboles sean más heterogéneos ya que se arman a partir de datos diferentes cuando la proporción de muestras bootsrapeadas es baja.

Dado que el modelo de bagging de 10 árboles y de proporción de muestra de 0.6 presenta los valores mas altos en accuracy y en precision, escogemos este modelo para compararlo con el obtenido en la seccion anterior de arbol de decision.

Bagging	
Accuracy	0.75
Precision	0.64
Recall	0.22

Comparando con el árbol de decisión hallado en la sección anterior, el modelo de Bagging sacrifica interpretabilidad a costa de mejorar la performance, se logra mejorar en un 5 % el accuracy y en un 4 % la precision. El recall sin embargo cae mas de un 10 %.

3.3. KNN

El modelo de K-nearest neighbors o vecinos mas cercanos es un modelo simple e interpretable. El modelo clasifica una nueva observación a partir de la clase de sus "vecinos", es decir, de las observaciones mas cercanas a esta. K es la cantidad de vecinos que se consideran para realizar la clasificación, se clasifica según cual es la clase mayoritaria en los vecinos de la nueva observación. Un enfoque tal vez mas cauto no le otorga la misma relevancia a todos los vecinos sino que pondera su importancia por la distancia: mientras mas cercano es un vecino, mayor es su importancia.

Para escoger el valor de K se buscaron varios clasificadores de KNN y se estimó el error de clasificación utilizando leave-one out cross-validation, luego se escogió el valor de K que minimiza este error.

Dado que este modelo es susceptible a diferencias en escala entre las variables, se consideró también un dataset normalizado que se obtiene restándole a cada variable su media y dividiéndola por el desvió estándar. Además, con el objetivo de visualizar las regiones en las que este algoritmo realiza la división, se tomo la decisión de considerar aparte sólo a las variables glucosa y BMI.

Los resultados fueron los siguientes:

Glucosa/BMI - Sin normalizar	
K	**Error**
1	0.31
5	0.28
10	0.29
15	0.26

Glucosa/BMI - Normalizado	
K	**Error**
1	0.30
5	0.28
10	0.26
15	0.26

Todas las variables - Sin normalizar	
K	**Error**
1	0.32
5	0.28
10	0.26
15	0.26

Todas las variables - Normalizado	
K	**Error**
1	0.29
5	0.26
10	0.25
15	0.26

Como era esperado, normalizar las variables del dataset arrojó mejores resultados, tanto al utilizar el dataset completo como al considerar solo 2 variables. Se puede ver que normalizar el set de datos completo otorga significativamente mejores resultados, esto se debe a que al tener mas variables hay mayor probabilidad de que las escalas de estas sean distintas.

Se obtuvieron buenos resultados para varios que modelos de KNN que logran un error de predicción de 26 %. El mejor hallado utiliza K = 10 y tiene un error de predicción estimado de 25 %. Sin embargo con el objetivo de visualizar la división del espacio de predicción, y compararla con aquella que realizan el árbol de decisión, LDA y QDA se muestra la partición que realiza el modelo de KNN que considera solo a glucosa y BMI sin normalizar:

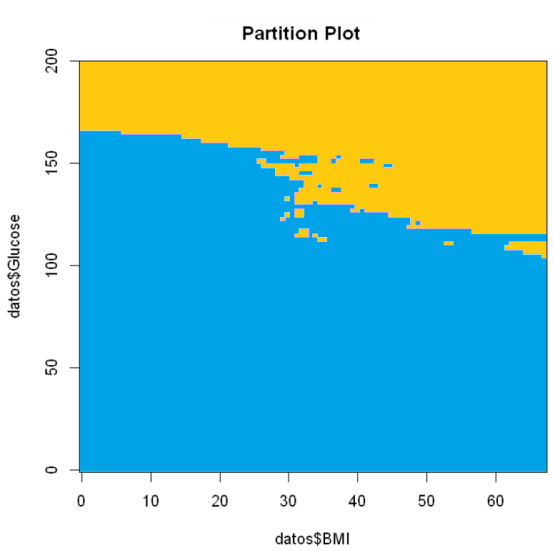


Figura 11: Partition plot - KNN

Podemos ver que esta partición es mucho mas flexible que la del árbol decisión. Por ultimo para realizar una comparación con las métricas de precision, recall y accuracy, buscamos los valores de estos para el modelo de KNN con todas las variables y el dataset normalizado, los resultados fueron:

Todas las variables - Normalizado	
	KNN
Accuracy	0.775
Precision	0.725
Recall	0.571

Podemos ver que para este modelo se obtuvieron muy buenos resultados respecto de accuracy, precision y recall, incluso mejores que los hallados con bagging. Se destaca que el recall mejoró significativamente.

3.4. LDA y QDA

En esta sección se busca resolver el problema de clasificación mediante LDA y QDA, tanto con todas las variables como también utilizando únicamente BMI y Glucosa para comparar la partición del espacio con las halladas anteriormente.

A continuación se procede a explicar brevemente en que consisten estos algoritmos:

Suponiendo que se cuenta con un conjunto de observaciones con sus etiquetas que indican a que clase de las k posibles pertenecen y una nueva observación x$\in \mathbb{R}^p$ que se busca clasificar.

Se define la variable aleatoria G = clase de x. De manera que P(G = j)=π_j con j = 1,2,...,k es la probabilidad a priori de que la observación x provenga de la clase j. Se puede estimar como: la proporción de observaciones cuya etiqueta es j. En este caso k = 2 y:

$$\hat{\pi}_1 = 500/768 \qquad \hat{\pi}_2 = 268/768$$

Suponiendo que la distribución de las variables aleatorias X|G=j es conocida, es decir que suponemos conocer la distribución de una observación sabiendo de que clase proviene, entonces es posible calcular P(G = j| X = x) la probabilidad a posteriori de que que la observación sea de la clase j con la regla de Bayes, pues:

$$\text{P(G = j| X= x)} = \frac{\pi_j \cdot f_{X|G=j}(x)}{\sum_{n=1}^{k} f_{X|G=n}(x)}$$

Y será posible asignar x a la clase para la cual la probabilidad a posteriori sea mayor.

Es decir, asignaremos x a la clase i si: P(G = i|X = x) >P(G = j|X = x) $\forall$ j = 1,2,...,k $\neq$ i (se observa que al plantear esta inecuación los denominadores se simplifican)

Si además X|G=j $\sim N_p(\mu_j, \Sigma_j)$, entonces la condición anterior puede escribirse como: asignaremos x a la clase i si:

$$Q_i(x) = -\tfrac{1}{2}\log|\Sigma_i| - \tfrac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) + \log\pi_i > Q_j(x) \ \forall \text{ j = 1,2,...,k} \neq \text{i}$$

En particular, si suponemos que las matrices de covarianza Σ_i son iguales para cada clase, entonces la condición se simplifica y se asignará x a la clase i si:

$$L_i(x) = \mu_i^T\Sigma^{-1}(x - \tfrac{1}{2}\mu_i) + \log\pi_i > L_j(x) \ \forall \text{ j = 1,2,...,k} \neq \text{i}$$

Las funciones dadas por $L_i(x) - L_j(x)$ y $Q_i(x) - Q_j(x)$ se llaman funciones discriminantes y el conjunto de puntos $Q_{ij} = \{x \in \mathbb{R}^p : Q_i(x) - Q_j(x) = 0\}$ y $L_{ij} = \{x \in \mathbb{R}^p : L_i(x) - L_j(x) = 0\}$ conforman una frontera de decisión. A un lado, la probabilidad de que la observación sea de la clase i es mayor a que sea de la clase j y del otro ocurre lo opuesto. Estas fronteras de decisión son funciones cuadráticas en x en el caso de QDA y funciones lineales en x en el caso de LDA.

Tomando en consideración solo las variables de BMI y Glucosa, las estimaciones halladas para las matrices de covarianza y los vectores de medias son las siguientes:

$$\hat{\Sigma}_1 = \begin{bmatrix} 59{,}133 & 26{,}484 \\ 26{,}484 & 683{,}362 \end{bmatrix} \qquad \hat{\Sigma}_2 = \begin{bmatrix} 52{,}750 & 11{,}695 \\ 11{,}695 & 1020{,}139 \end{bmatrix} \qquad \hat{\Sigma} = \begin{bmatrix} 56{,}908 & 21{,}329 \\ 21{,}329 & 800{,}750 \end{bmatrix}$$

$$\hat{\mu}_1 = \begin{bmatrix} 30{,}304 \\ 109{,}980 \end{bmatrix} \qquad \hat{\mu}_2 = \begin{bmatrix} 35{,}142 \\ 141{,}257 \end{bmatrix}$$

Donde $\hat{\Sigma}$ se obtiene realizando un promedio ponderado (por la cantidad de observaciones de la clase 1 - no diabéticos y 2 - diabéticos) de las matrices $\hat{\Sigma}_1$ y $\hat{\Sigma}_2$

Utilizando estas estimaciones, junto con las estimaciones de π_1 y π_2 se logra encontrar que la recta que es frontera de decisión en LDA es:

$$-0{,}0371y - 0{,}07x + 7{,}623 = 0$$

Mientras que la frontera de decision hallada con QDA es:

$$1{,}024x^2 + 0{,}408xy - 0{,}215y^2 - 200{,}5x - 1{,}07y = -7197{,}01$$

Y estas funciones, dividen al espacio de decisión en las siguientes regiones:

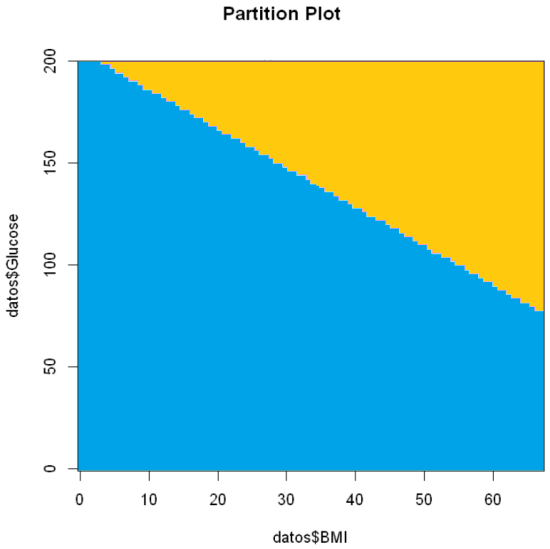


Figura 12: Particion LDA

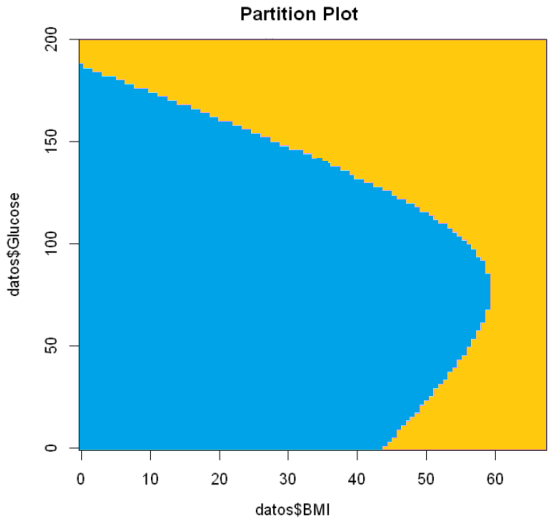


Figura 13: Particion QDA

Se pueden realizar las siguiente observaciones:

- Claramente QDA tiene mayor flexibilidad en la frontera de las regiones, esto sucede porque el algoritmo realiza menos supuestos acerca de los datos, en particular no presupone que la matriz de covarianza es la misma para todas las clases.
- Las regiones halladas con QDA son mas similares a las halladas con el árbol de decisión por lo tanto se espera que el desempeño de esta sea similar a la del árbol.
- Las regiones halladas con LDA son mas similares a las halladas con KNN por lo tanto se espera que el desempeño de esta sea similar a la de KNN

Los resultados obtenidos en las métricas para estos modelos, utilizando leave-one-out CV fueron los siguientes:

BMI - Glucosa		
	LDA	**QDA**
Accuracy	0.769	0.760
Precision	0.733	0.720
Recall	0.533	0.503

Todas las variables		
	LDA	**QDA**
Accuracy	0.775	0.74
Precision	0.725	0.649
Recall	0.571	0.552

4. Conclusiones

Se obtuvieron las siguientes conclusiones a partir de la realización del trabajo practico:

- Las variables mas relevantes para realizar la clasificación son Glucosa y BMI, pues:
  - Son las variables con mayor correlación con el resultado que indica si una persona es diabética o no.
  - Los boxplots muestran que los valores de estas variables son significativamente diferentes para las personas diabéticas que para las personas no diabéticas
  - El scatter plot de BMI vs. Glucosa muestra que estas variables separan bien a las observaciones de cada clase.
  - Al utilizar los distintos modelos con estas variables se obtuvieron resultados similares que al utilizar los modelos con todas las variables.
- En todos los modelos el recall fue la métrica con menor valor. Esta es una métrica muy importante sobre todo en este caso, el siguiente ejemplo lo demuestra. Si a las personas que se predice con diabetes reciben un medicamento y a las personas que se predicen no diabéticas no lo reciben, pueden darse las siguientes situaciones:
  - Una persona con diabetes recibe el medicamento
  - Una persona sin diabetes recibe el medicamento
  - Una persona con diabetes no recibe el medicamento
  - Una persona sin diabetes no recibe el medicamento

  Claramente el error que se comete cuando una persona con diabetes no recibe el medicamento que le es necesario es peor a el que se comete cuando una persona sin diabetes lo recibe. Es decir hay que buscar tener pocos falsos negativos (personas que se clasifican no diabéticas y que en verdad lo son). Al final de las conclusiones se proponen algunos cambios posibles a los algoritmos hallados para intentar mejorar esta métrica.
- El mejor modelo encontrado es:
- La división del espacio de decisión formado por las variables glucosa y BMI es muy diferente según el algoritmo que se utilice:
  - En árbol de decisión: las regiones están delimitadas por rectas horizontales y verticales que se obtienen como respuesta a las preguntas ¿Es Glucosa <a? ¿Es BMI <b? con a y b hallados de manera de minimizar el valor de Gini.

- En KNN las regiones no son tan rígidas y varían significativamente para distintos valores de K. En particular para K = 10 se puede ver que dentro de la región anaranjada hay pequeños grupos de instancias celestes, lo cual no se puede ver con ninguno de los otros métodos.
- LDA y QDA dividen el espacio decisión con una recta y con una hipérbola respectivamente. El espacio hallado con LDA es bastante parecido al que se halla con KNN y es por esto que los modelos presentan valores muy similares en cuanto a performance.

A continuación, se proponen las siguientes mejoras para el trabajo práctico:

- Una potencial mejora para este proyecto podría realizarse obteniendo un dataset menos sesgado. El dataset utilizado en el trabajo tiene información solo de mujeres y que además comparten ascendencia, lo cual potencialmente reduce la extrapolación de estos resultados.
- De todos los algoritmos utilizados en el trabajo puede extraerse una estimación de probabilidad de pertenecer a una clase. Por ejemplo:
  - En el árbol de decisión: la cantidad de observaciones de cada clase que hay en las hojas del árbol pueden utilizarse como estimación para la probabilidad de que una observación pertenezca a una clase.
  - En bagging: mientras mas arboles predigan que una observación es de cierta clase, mayor sera la probabilidad de que sea de esa clase.
  - En KNN: contando la cantidad de vecinos de una clase se puede obtener la probabilidad de que sea de esa clase

  Seria interesante ver que ocurre con el recall y el precision al variar el umbral de probabilidad para el cual uno considera que una persona es diabática.

5. Bibliografía

1 Flury B., A First Course in Multivariate Statistics

2 Seber G., Multivariate Observations

3 Rencher, A. (n.d.). Methods of Multivariate Analysis (2nd ed.). Brigham Young University: John Wiley Sons, Inc.

4 Larry A. Wasserman, All of Statistics: A Concise Course in Statistical Inference