Segundo Examen Parcial

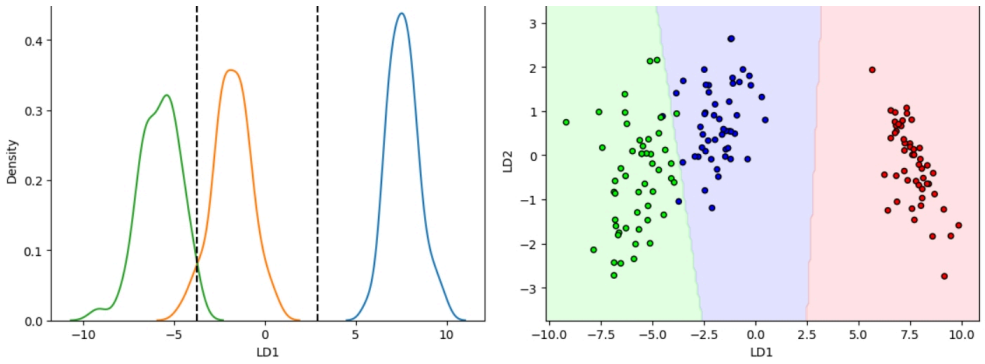
Luis Enrique Garcia Gallegos

Matricula: 649247

Pregunta 2

Las siguientes imágenes se generaron tras haber obtenido un modelo de linear discriminant analysis usando la función LinearDiscriminantAnalysis de sklearn, así como los métodos fit y transform, entre otros. La base de datos utilizada cuenta con 150 observaciones, 40 variables de entrada, y una variable de salida cualitativa de 3 clases.

- Basándote únicamente en las gráficas (no tomes en cuenta la descripción que te acabo de dar sobre la base de datos) ¿qué información sí se puede inferir sobre la base de datos y qué información no (cantidad de variables cualitativas, cantidad de variables cuantitativas, cantidad de observaciones, cantidad de clases, tipo de validación realizada)? Indica por qué sí o por qué no para cada caso.
- ¿Los colores de la gráfica de densidad corresponden con los de la gráfica de disperción? Es decir, ¿la curva azul de la primera gráfica tiene una relación directa con los puntos azules de la segunda gráfica?, ¿por qué?
- ¿Una de las dos gráficas representa de forma más precisa lo que el modelo hace?, si sí, ¿cuál y por qué?
- ¿Qué discriminante tiene mayor poder predictivo, LD1 o LD2?, ¿por qué?
- En las gráficas solo se hace referencia a LD1 y a LD2, ¿es posible que en este ejemplo existan LD3 y LD4 también, pero que simplemente no se haya deseado graficarlas?, ¿por qué?
- Explica, haciendo referencia a las características de la base de datos descrita y a las gráficas mostradas, ¿por qué también se considera al LDA como una técnica de reducción de dimensionalidad?



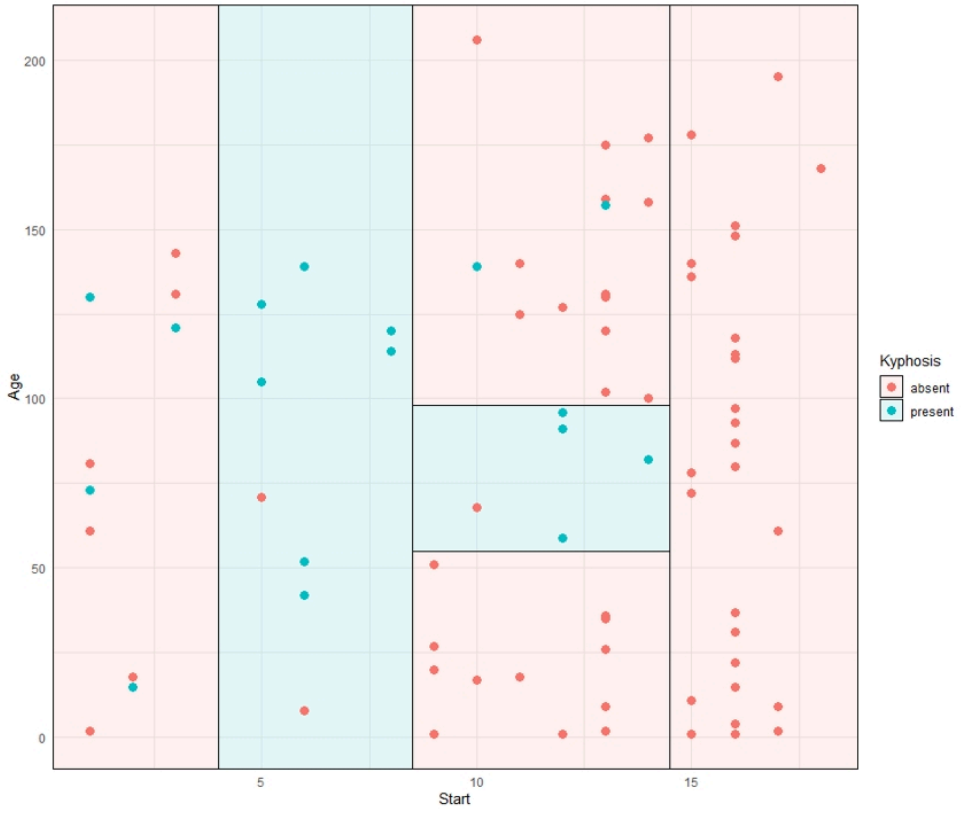
- Para el modelo de la base de datos tenemos una clase que es más fácil de diferenciar, ya que vemos en la grafica que densidad que esta esta mas alejada de las otras dos, mientras que las otras dos tienen ciertos datos que se encuentran en la partición y típicamente va a tener problemas para diferenciar.
- Los colores de las graficas no corresponden entre ellas, sin embargo podemos diferenciar entre ellas si nos fijamos en el eje LD1, por lo que podriamos interpretar el como se comporta nuestro modelo para la base de datos.
- En lo personal la grafica de densidad es más intepretable ya que si la comparamos con la otra veremos que la particion que hace de las clases es dificl de explicar, mientras que la grafica de densidad no, ya que podemos decir que si el LD1 es menor a -4 aproximadamente posiblemente pertenezca a una cle mientras que en la otra grafica tendriamos que fijarnos en la pendiente que tiene, volviendo al ejemplo pasado, si su LD1 es menor a -4 y LD2 es menor a 2 aproximadamente puede que sea de la clase tal, pero que pasaria si LD1 y LD2 fueran menor, entonces tambien serian de la misma clase, por ende es más facil de explicar la de densidad que la que compara ambas variables.
- LD1 tiene mayor poder predictivo, porque si nos fijamos en las particiones veremos que la pendiente respecto a LD1 es mucho mayor que en LD2 por lo que a la hora de predecir una clase LD2 no tiene un gran impacto ya que si nos fijamos en la gráfica veremos que en todas las clases pueden tener similar LD2 por ejemplo si nos fijamos entre LD2 de -1 y el de 1 veremos que costaria mas trabajo diferenciar las clases.
- Es posible que existan LD3 y LD4 sin embargo puede ser que estas variables tengan tanta varianza que provoquen que en el modelo no aporten información para predecir la clase, por lo que no se tomarían en cuenta y por ende no se gratifiquen.
- Porque LDA se basa mas en probabilidades de nuestros datos para poder trabajar con el teorema de Bayes, por lo que no importa que tan grande sea la base de datos esta sera reducida a valores entre 0 a 1.

Pregunta 3

Con respecto a los árboles de decisión, reponde las siguientes preguntas:

- Se sabe que los árboles de decisión, si no se les pone algún tipo de límite, tienden a generar resultados con un sobreajuste excesivo, ¿por qué? Apoya tu explicación con el ejemplo de un árbol que crece sin restricciones. Para lograrlo, indica una situación real en que utilizar un árbol de decisión haga sentido. Asume que, dentro de dicho escenario, tienes una base de datos con 20 observaciones y 3 variables. Crece un árbol sin ningún tipo de restricción para dicha base de datos. indicando en cada nodo (intermedios y finales) la cantidad de observaciones de cada clase y la decisión utilizada (obviamente, las decisiones serán completamente arbitrarias, pues no contamos con datos reales).
- Para la imagen que se muestra, indica al menos dos posibles restricciones específicas que se le hayan definido al modelo para restringir su crecimiento (por ejemplo, si indicas que la cantidad de algo se restringió, especifica el número).

- Genera pseudocódigo que muestre el proceso de poda de un árbol de decisión con la metodología de cost complexity pruning. Asume que ya cuentas con funciones que pueden calcular y/u optimizar métricas (por ejemplo, puedes agregar una línea que indique que se usa una función que entrega los coeficientes que optimizan el RSS), pero no puedes asumir que se cuentan con funciones que realizan operaciones complejas (por ejemplo, no puedes agregar una línea que indique que se realiza el proceso de poda con la función prune).



- Si hicieramos un arbol de decisiones de una base de datos con 20 observaciones y lo dejaramos creecer puede generarse sobreajuste porque diriamos por ejemplo si para la variable 1 tenemos valore menor a 3 entonces puede petenecer a tal clase y lo mismo al caso contrario y digamos que esto provoco que la mitad de mis datos se fueran a una rama y las demas a la otra, y repetimos lo mimso pero si su variable 2 es si entonces puede pertenecer a tal clase y lo mismo al caso contrario y asi sucesivamente hasta tener un arbol donde tenemos muchas hojas donde cada hoja pertenezca a una observación, y en cuanto llegen otros datos para probar nuestro modelo veremos que no logra predecir correctamente, justamente porque creamos un sobreajuste.
- No genere mas particiones si en estas tienen menos de 5 observaciones y otra restriccion seria no haga más particiones si tenemos un porcentaje menor del 10% aproximadamente observaciones de una clase.
- Si en nuestro árbol tenemos hojas donde tengamos menor de n cantidad de observaciones entonces quita esa rama, si en la hojas no tienes al menor cierto

porcentaje de observaciones de una clase entonces corta esa rama; repetir este proceso en cada hoja hasta que se cumplan ambas condiciones.

Pregunta 4

A continuación se muestra pseudocódigo que describe un algoritmo no revisado en clase.

- ¿A qué algoritmo(s) de los vistos en clase se parece? Indica claramente por qué y en qué se parecen.
- Para dicho algoritmo, ¿se tendría que trabajar forzosamente con árboles de decisión como modelo base?, ¿por qué?
- Este algoritmo, tal y como se describe, ¿para qué tipo de problemas puede ser utilizado (clasificación, regresión, ambos, ninguno)?, ¿por qué?
- Existe una diferencia muy marcada entre modelos de bagging y de random forests. Modifica el pseudocódigo para incluir dicha característica al algoritmo (que el modelo se parezca a random forests en ese sentido).

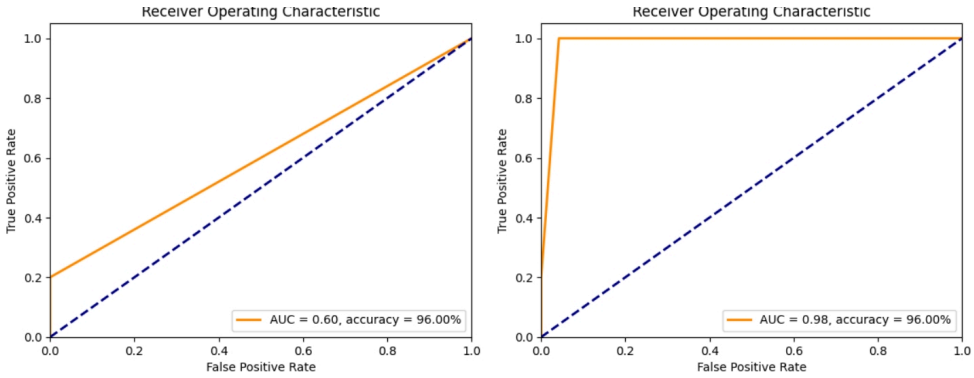
```
    - Datos de entrenamiento D
    - Datos de prueba P
    - Número de modelos M
    - Tipo de modelo base
  Salida:
    - Predicción en P

  Método:
    1. Inicializa un peso igual para cada punto de datos en D.
    2. Para m = 1 hasta M hacer:
        a. Si m es impar:
            i.   Crea una muestra seleccionando con reemplazo de D usando los pesos actuales.
            ii.  Entrena un modelo en la muestra.
            iii. Actualiza los pesos de los puntos de datos en D para enfatizar los puntos mal clasificados
        b. Si m es par:
            i.   Crea una muestra seleccionando con reemplazo de D ignorando los pesos actuales.
            ii.  Entrena un modelo en la muestra.
    3. Combina todos los modelos M en un conjunto:
        a. Para una entrada dada p, la predicción del conjunto es la mediana de los modelos M.
```

- El metodo lo siento como Random Forests, ya que queremos generar varios modelos que sean disntintos entre ellos y al mismo tiempo quemos genere un promedio de los outputs de los modelos, sin embargo se van a crear hasta M modelos como si fueran Bossting y Bagging fusionados ya que partiremos nuestros datos para generar modelo al estilo bootstrap pero estos modelos alterarn nuestros datos datos como en el Bossting.
- Creo que no forzosamente se trabaje con arboles de decisiones porque usualmente los arboles deciones se usan para problemas de clasifiacion por lo que al intenar usar varios modelos y usar la mediana para predecir una variable cualitativa no tendria mucho sentido por el orden de estas por ejemplo si trataramos de predecir colores como asignarias el orden.
- Para este algoritmo siento que seria más que nada para problemas de regresión porque como obtendremos un numero podemos sacar la mediana facilmente, mientra que para problemas de clasifiacion sacar una mediana seria más conflictivo.
- La decidir si un modelo va para el caso a o b y diectamente ir al caso a seria más acorde a random forest

Pregunta 5

Para una base de datos con 100 observaciones se tiene un modelo de clasificación binaria con una exactitud de 96%:

- Indica un escenario específico en que, a pesar de la exactitud tan alta, dicho modelo se considere que tiene un desempeño pobre. Explica qué métrica o métricas generadas a partir de los valores que se encuentran en una matriz de confusión podrían dejar en claro que el desempeño del modelo es pobre, especifica los valores de dichas métricas en el escenario que generaste.
- Para este escenario, ¿habría algún beneficio potencial de utilizar una estrategia de validación cruzada con k-folds estratificada contra una estrategia de validación cruzada con k-folds tradicional?, ¿por qué?
- Se genera un nuevo modelo que trata de corregir los problemas del primero; el nuevo modelo tiene la misma exactitud (96%), pero sus curvas ROC y el área bajo las mismas son extremadamente distintas. ¿Cómo es esto posible? Explíca esta situación al interpretar las curvas que se muestran. Enfócate en explicar por qué dos modelos con la misma exactitud pueden tener diferentes valores de AUC.



- Cuando queremos dectectar un falso ya que puede ser que exista un falso positivo cuando no era haci esto se puede deber a un desbalance de datos, lo cual no puedo saber si no tengo una matriz de confusion
- No debido a lo que mencione ante
- Puede deberse que a un desbalance de datos donde casualmente la mayoria de datos de un tipo cayeron en el entrenamiento.

Firma de Honor: Doy mi palabra que he realizado esta actividad con integridad académica