Universidad Autónoma de Yucatán

Maestría en Ciencias de la Computación

Métodos Estadísticos de Machine Learning

Proyecto 2

Autor: Mario Herrera Almira

20 de marzo del 2023

En la página web https://datasciencedojo.com/blog/datasets-data-science-skills/?utm_campaign=DSD%20blogs%202022&utm_content=223545409&utm_medium=social&utm_source=twitter&hss_channel=tw-1318985240 se presenta un repositorio de datos, que se pueden utilizar con el objetivo de mejorar las habilidades en el área de Ciencia de Datos.

Utilizando la base de datos asignada al azar, escribe un reporte que incluya lo siguiente:

1. Una introducción, que incluya
    a. Descripción del problema, así como cada una de las variables involucradas en el estudio (8 puntos)
    b. Identificación de la variable dependiente y la independiente (2 puntos)
    c. Explicar si se trata de un problema de regresión o clasificación (2 puntos)
    d. Explicar la forma o criterios usados para eliminar registros con datos faltantes (si fue necesario) (3 puntos)
2. Sobre el modelo basado en árboles:
    a. Describir con mucho detalle el método Boosting utilizado para la construcción del modelo (15 puntos)
    b. Presentar los indicadores de desempeño del modelo (10 puntos)
    c. Describir con mucho detalle el método Bagging utilizado para la construcción del modelo (15 puntos)
    d. Presentar los indicadores de desempeño del modelo (10 puntos)
3. Ajuste un modelo de regresión Logística:
    a. Presente el modelo estimado y el método utilizado en la construcción (15 puntos)
    b. Presente los indicadores de desempeño del modelo (10 puntos)
4. Enviar el script r utilizado en el proyecto, con los comentarios necesarios para su entendimiento (10 puntos)

Este proyecto se debe entregar como un documento en formato pdf, a más tardar el martes 2 de mayo a las 23:50 horas.

1-)

a-) Descripción del problema, así como cada una de las variables involucradas en el estudio.

La base de datos que se utiliza en este proyecto se llama *Echocardiogram Data Set* donde la pregunta que se intenta contestar es si un paciente sobrevivirá o no después de un año de haber tenido un ataque al corazón.

Este conjunto de datos tiene 132 filas y 12 columnas. Todos los pacientes sufrieron ataques cardíacos en algún momento en el pasado. Algunos siguen vivos y otros no. Las variables *survival* y *still-alive*, cuando se toman en conjunto, indican si un paciente sobrevivió durante al menos un año después del ataque cardíaco. La parte más difícil de este problema es predecir correctamente que el paciente NO sobrevivirá. (Parte de la dificultad parece ser el tamaño del conjunto de datos).

Variable	Definición	Tipo de Dato	Ejemplo
Survival	El número de meses que sobrevivió el paciente (ha sobrevivido, si el paciente aún está vivo). Debido a que todos los pacientes tuvieron sus ataques cardíacos en diferentes momentos, es posible que algunos pacientes hayan sobrevivido menos de un año, pero aún estén vivos. Dichos pacientes no pueden utilizarse para la tarea de predicción mencionada anteriormente.	Cuantitativa	11, 57, 26
Still-alive	Una variable binaria que muestra si el paciente todavía está vivo (0: muerto al final del período de supervivencia, 1: todavía vivo).	Cuantitativa	0, 1
Age-at-heart-attack	Edad en años en que ocurrió el infarto.	Cuantitativa	71, 57, 62
Pericardial-effusion	El derrame pericárdico es líquido alrededor del corazón (0: sin líquido, 1: líquido).	Cuantitativa	0, 1
Fractional-shortening	Una medida de contracción alrededor del corazón, los números más bajos son cada vez más anormales.	Cuantitativa	0.23, 0.13, 0.45
Epss	Separación septal del punto E, otra medida de la contractilidad. Los números más grandes son cada vez más anormales.	Cuantitativa	6, 22, 12.062
Lvdd	Dimensión del ventrículo izquierdo al final de la diástole. Esta es una medida del tamaño del corazón al final de la diástole. Los corazones grandes tienden a ser corazones enfermos.	Cuantitativa	4.26, 4.23, 5.39
Wall-motion-score	Una medida de cómo se mueven los segmentos del ventrículo izquierdo.	Cuantitativa	14, 22.5, 27
Wall-motion-index	Es igual a la puntuación del movimiento de la pared dividida por el número de segmentos vistos. Por lo general, se ven 12-13 segmentos en un ecocardiograma. (Es preferible utilizar esta variable EN LUGAR de Wall-motion-score).	Cuantitativa	1, 1.625, 2
Mult	Una variable derivada (se sugiere ignorarla).	Cuantitativa	0.558, 1, 1.003
Name	El nombre del paciente.	Cualitativa	Nombre
Group	Grupo (Se ha considerado sin sentido y se sugirió ignorar).	Cuantitativa	1, 2
Alive-at-1	Es una variable de valor booleano derivada de los dos primeros atributos. (0: el paciente murió después de 1 año o había sido seguido por menos de 1 año, 1: el paciente estaba vivo al año).	Cuantitativa	1, 0

b-) Identificación de la variable dependiente y la independiente.

La variable dependiente que es la que se quiere predecir es *Alive-at-1,* hay otra variable que es dependiente porque se deriva de otras que es *Mult* pero esa no se desea predecir y se recomienda ignorarla por lo que no se va a tener en cuenta. El resto de las variables son independientes, pero hay algunas que van a ser ignoradas por las razones descritas a continuación:

- **Name:** Esta variable es solo el nombre del paciente, no aporta información relevante para el calculo estadístico por tanto no será tomada en cuenta.
- **Mult:** Es una variable derivada por tanto no será tomada en cuenta.
- **Group:** Se ha considerado sin sentido y por tanto se decidió excluirla del modelo.
- **Wall-motion-score:** Es preferible utilizar la variable *Wall-motion-index* en lugar de *Wall-motion-score* por lo que no se tomará en cuenta *Wall-motion-score* para construir el modelo.
- **Survival y Still-alive:** Estas dos variables no se pueden utilizar porque la variable que se desea predecir es el resultado de combinar estas dos por lo que habría información redundante causando que el modelo no funcionara correctamente.

c-) Explicar si se trata de un problema de regresión o clasificación.

En este caso se trata de un problema de Clasificación porque lo que se espera es predecir a que clase corresponde la variable predicha, es decir, si el resultado es 0 (el paciente murió después de 1 año o había sido seguido por menos de 1 año) o 1 (el paciente estaba vivo al año). No puede ser un problema de regresión porque no se intenta predecir un valor continuo como el precio de una casa, por ejemplo.

d-) Explicar la forma o criterios usados para eliminar registros con datos faltantes (si fue necesario).

En este caso sí fue necesario eliminar registros con datos faltantes, la base de datos fue procesada de manera manual para sustituir los registros faltantes que estaban representados con un signo de interrogación (“?”) por las letras “NA”. Luego mediante código R se eliminaron todos los registros que tuvieran datos NA utilizando la función “na.omit(datos)”.

2-)

a-) Describir con mucho detalle el método Boosting utilizado para la construcción del modelo.

El boosting es una técnica popular en el aprendizaje automático (machine learning) que se utiliza para mejorar el rendimiento de los algoritmos de clasificación y regresión. El método se basa en la combinación de varios modelos de aprendizaje débiles para crear un modelo final más preciso y robusto.

El boosting funciona mediante la creación de una secuencia de modelos, en la que cada modelo subsiguiente se entrena para corregir los errores del modelo anterior. El proceso se repite hasta que se alcanza un cierto nivel de precisión o hasta que se agota un número predefinido de modelos.

El método de boosting se puede aplicar a cualquier algoritmo de aprendizaje automático, pero se utiliza con mayor frecuencia con árboles de decisión y redes neuronales. Los árboles de decisión son modelos de aprendizaje débiles que se utilizan como base para los modelos subsiguientes. En cada iteración, el

árbol de decisión se entrena para clasificar correctamente los datos mal clasificados por el modelo anterior.

El boosting también utiliza una técnica llamada "peso de ejemplo", que se utiliza para dar más importancia a los ejemplos de entrenamiento que son difíciles de clasificar. En cada iteración, los pesos de los ejemplos que se clasificaron incorrectamente se incrementan, lo que los hace más propensos a ser clasificados correctamente en el siguiente modelo. De esta manera, el modelo final puede centrarse en las áreas de los datos que son más difíciles de clasificar.

En resumen, el método de boosting es una técnica de aprendizaje automático que combina varios modelos de aprendizaje débiles para crear un modelo final más preciso y robusto. El método utiliza árboles de decisión y redes neuronales como base y utiliza pesos de ejemplo para centrarse en las áreas de los datos más difíciles de clasificar.

b-) Presentar los indicadores de desempeño del modelo.

El modelo obtenido mediante Boosting presenta un AUC (Area Under the Curve) de 0.875.

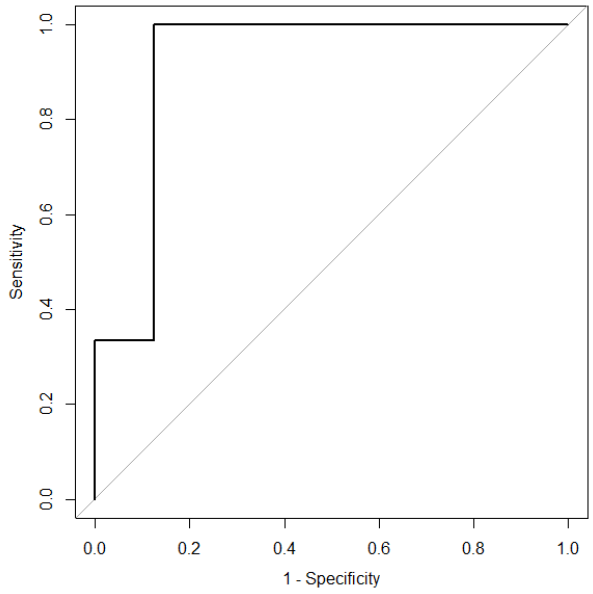
La matriz de confusión es la siguiente:

Predicción	x0	x1
x0	7	2
x1	1	1

Otras métricas de desempeño son las siguientes:

Accuracy	0.7273
Sensitivity:	0.33333
Specificity	0.87500
Pos Pred Value	0.50000
Neg Pred Value	0.77778
Prevalence	0.27273
Detection Rate	0.09091
Detection Prevalence	0.18182
Balanced Accuracy	0.60417

El gráfico de la curva ROC es el siguiente:



c-) Describir con mucho detalle el método Bagging utilizado para la construcción del modelo.

El método Bagging (Bootstrap Aggregating) es una técnica de ensamblaje o "ensemble" que se utiliza en el aprendizaje automático para mejorar la precisión de los modelos predictivos. En R, el paquete "randomForest" es una de las implementaciones más populares del método Bagging.

La idea detrás del método Bagging es entrenar múltiples modelos de aprendizaje automático independientes utilizando diferentes conjuntos de datos de entrenamiento que se generan mediante el muestreo con reemplazo (bootstrap) de los datos de entrenamiento originales. Cada modelo se entrena con un conjunto de datos de entrenamiento diferente, lo que significa que cada modelo puede aprender de una perspectiva diferente y puede tener fortalezas y debilidades distintas.

Una vez que se han entrenado los modelos, se utiliza un método de agregación para combinar las predicciones de los diferentes modelos en una única predicción final. En el caso de Bagging, la técnica de agregación utilizada es el promedio de las predicciones de los modelos individuales.

El resultado final es un modelo de conjunto (ensemble) que puede ser más preciso y generalizar mejor que un único modelo entrenado con todo el conjunto de datos de entrenamiento original.

Es importante tener en cuenta que el método Bagging funciona mejor con modelos que son propensos a sobreajustarse (overfitting) a los datos de entrenamiento. Al entrenar múltiples modelos en diferentes conjuntos de datos de entrenamiento, el Bagging reduce la varianza y, por lo tanto, ayuda a evitar el sobreajuste y mejora la precisión del modelo predictivo.

d-) Presentar los indicadores de desempeño del modelo.

El modelo obtenido mediante Bagging presenta un AUC de 1.0.

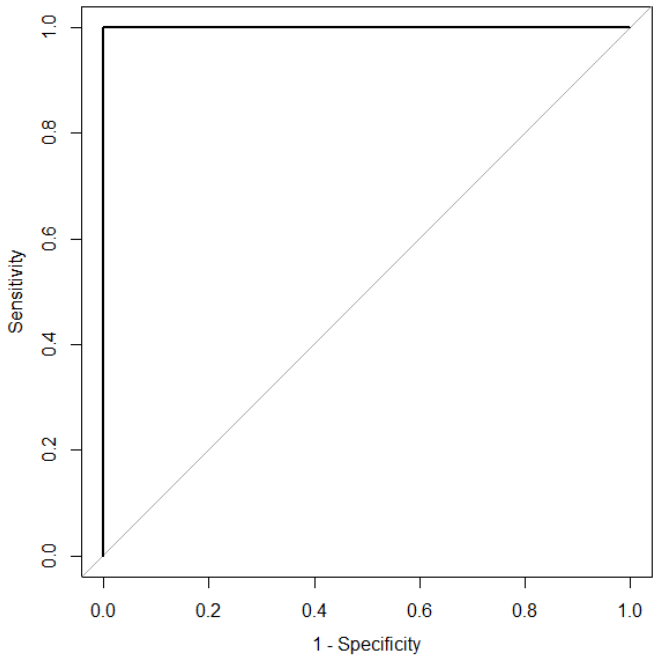
La matriz de confusión es la siguiente:

Predicción	x0	x1
x0	8	1
x1	0	2

Otras métricas de desempeño son las siguientes:

Accuracy	0.9091
Sensitivity:	0.6667
Specificity	1.0000
Pos Pred Value	1.0000
Neg Pred Value	0.8889
Prevalence	0.2727
Detection Rate	0.1818
Detection Prevalence	0.1818
Balanced Accuracy	0.8333

El gráfico de la curva ROC es el siguiente:



3-) Ajuste un modelo de regresión Logística.

a-) Presente el modelo estimado y el método utilizado en la construcción.

Para la construcción del modelo se utilizaron las variables Age-at-heart-attack, Pericardial-effusion, Fractional-shortening, Epss, Lvdd, Wall-motion-index y Alive-at-1, esta última es la variable que se desea predecir. El resto de las variables son ignoradas por los motivos expuestos en la sección 1.b de este trabajo.

El método utilizado para la construcción del modelo logístico fue la función de R “glm”.

b-) Presente los indicadores de desempeño del modelo.

El modelo obtenido mediante Regresión Logística presenta un AUC de 1.0.

La matriz de confusión es la siguiente:

Predicción	x0	x1
x0	8	1
x1	0	2

Otras métricas de desempeño son las siguientes:

Accuracy	0.9091
Sensitivity:	0.6667
Specificity	1.0000
Pos Pred Value	1.0000
Neg Pred Value	0.8889
Prevalence	0.2727
Detection Rate	0.1818
Detection Prevalence	0.1818
Balanced Accuracy	0.8333

El gráfico de la curva ROC es el siguiente: