# Tarea Número 4

- **Ejercicio 1:** [25 puntos] Esta pregunta utiliza los datos sobre la conocida historia y tragedia del Titanic, usando los datos (`titanic.csv`) de los pasajeros se trata de predecir la supervivencia o no de un pasajero.

  La tabla contiene 12 variables y 1309 observaciones, las variables son:

  - `PassegerId`: El código de identificación del pasajero (valor único).
  - `Survived`: Variable a predecir, 1 (el pasajero sobrevivió) 0 (el pasajero no sobrevivió).
  - `Pclass`: En que clase viajaba el pasajero (1 = primera, 2 = segunda , 3 = tercera).
  - `Name`: Nombre del pasajero (valor único).
  - `Sex`: Sexo del pasajero.
  - `Age`: Edad del pasajero.
  - `SibSp`: Cantidad de hermanos o cónyuges a bordo del Titanic.
  - `Parch`: Cantidad de padres o hijos a bordo del Titanic.
  - `Ticket`: Número de tiquete (valor único).
  - `Fare`: Tarifa del pasajero.
  - `Cabin`: Número de cabina (valor único).
  - `Embarked`: Puerto donde embarco el pasajero (C = Cherbourg, Q = Queenstown, S = Southampton).

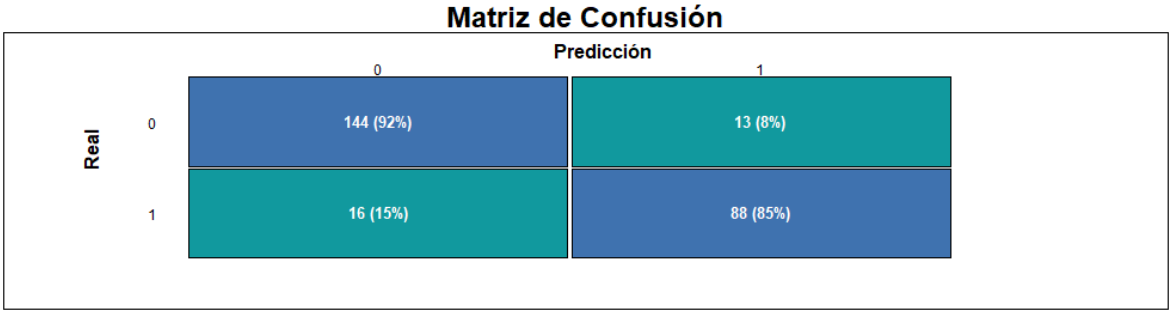
1. Cargue la tabla de datos `titanic.csv`, asegúrese re-codificar las variables cualitativas y de ignorar variables que no se deben usar.

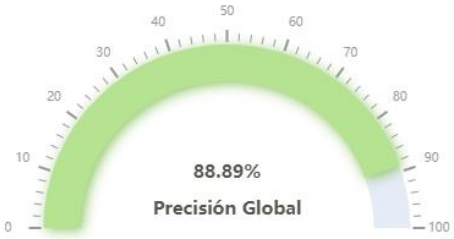
Datos

| d | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|
| ) | 3 | Braund, Mr. Owen Harris | male | 22 | 1 | 0 | A/5 21171 | 7.25 | | S |
| 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Thayer) | female | 38 | 1 | 0 | PC 17599 | 71.283 | C85 | C |
| 1 | 3 | Heikkinen, Miss. Laina | female | 26 | 0 | 0 | STON/O2. 3101282 | 7.925 | | S |
| 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily | female | 35 | 1 | 0 | 113803 | 53.1 | C123 | S |

Se cargó la base de datos, se modificó las variables y se eliminó los valores únicos ya que estos no tienen relevancia en nuestra predicción ya que son valores que no van a influir en nuestra metodología.

2. Use el método de **Máquinas de Soporte Vectorial** para generar un modelo predictivo para la tabla `titanic.csv`. Para esto utilice el 75 % de los datos para la tabla aprendizaje y un 25 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las dos categorías.



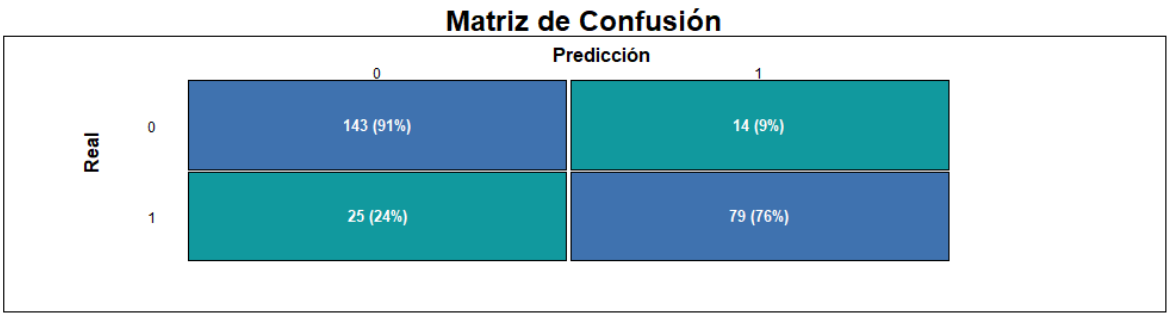


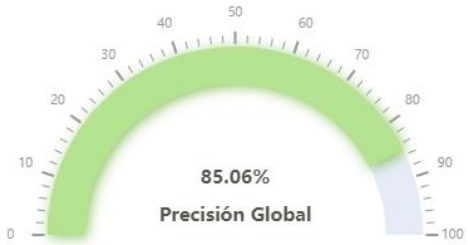
| Precisión.0 | Precisión.1 |
|---|---|
| 91.72 | 84.62 |

| | FN | FP | |
|---|---|---|---|
| VN | 144 | 13 | |
| VP | 16 | 88 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,888888889 | 88,8888889 |
| | | | |
| Precision Positiva= | VP / FN+VP | 0,846153846 | 84,6153846 |
| | | | |
| Precision Negativa= | VN / VN+FP | 0,917197452 | 91,7197452 |

el modelo tiene un buen desempeño tanto en la identificación de positivos como de negativos, con una alta especificidad y un valor predictivo negativo notablemente alto. Sin embargo, la sensibilidad es ligeramente menor, lo que sugiere que hay margen de mejora en la identificación de positivos reales.

3. Repita los dos ejercicios anteriores pero esta vez usando solamente las 4 variables que tienen mejor poder predictivo. ¿Mejoró el resultado?







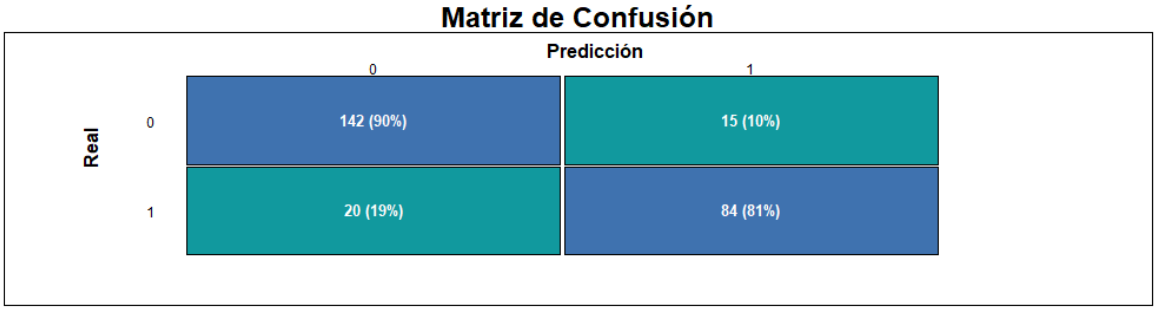
| Precisión.0 | Precisión.1 |
|---|---|
| 91.08 | 75.96 |

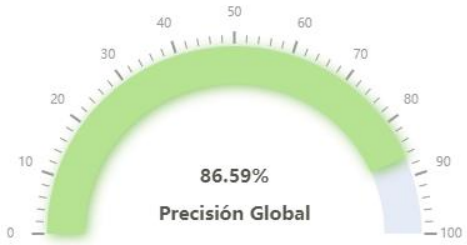
| | FN | FP | |
|---|---|---|---|
| VN | 143 | 14 | |
| VP | 25 | 79 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,850574713 | 85,0574713 |
| Precision Positiva= | VP / FN+VP | 0,759615385 | 75,9615385 |
| Precision Negativa= | VN / VN+FP | 0,910828025 | 91,0828025 |

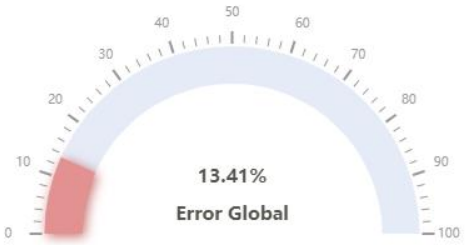
el modelo tiene una buena precisión global de aproximadamente 85%. La precisión positiva y la sensibilidad indican que el modelo es bastante eficaz en predecir la clase positiva, aunque hay un margen de mejora en la sensibilidad (76%), que sugiere que algunos verdaderos positivos están siendo clasificados incorrectamente como negativos. La especificidad es alta (91%), lo que indica que el modelo es eficaz en predecir correctamente los negativos.

4. Repita los dos ejercicios anteriores pero esta vez usando los otros núcleos (`Kernel` que permite en método **Máquinas de Soporte Vectorial**. ¿Mejoró alguno el resultado?

Nucleo (Kernel) LINEAL



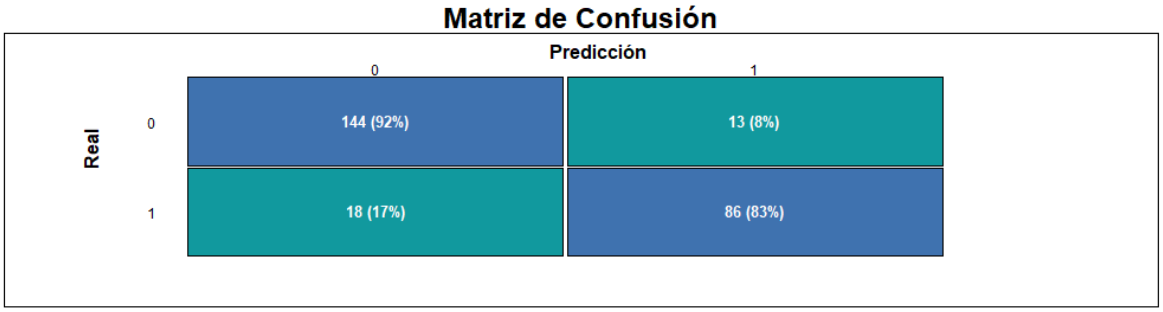



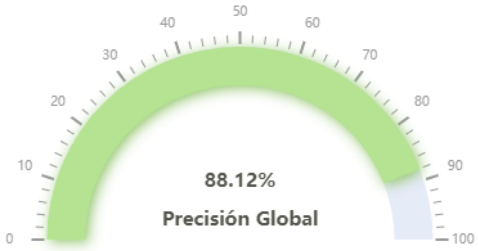

| Precisión.0 | Precisión.1 |
|---|---|
| 90.45 | 80.77 |

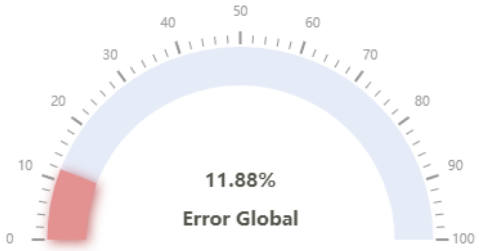
| | FN | FP | |
|---|---|---|---|
| VN | 142 | 15 | |
| VP | 20 | 84 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,865900383 | 86,5900383 |
| | | | |
| Precision Positiva= | VP / FN+VP | 0,807692308 | 80,7692308 |
| | | | |
| Precision Negativa= | VN / VN+FP | 0,904458599 | 90,4458599 |

el modelo tiene un rendimiento bastante equilibrado entre la predicción de clases positivas y negativas. La alta exactitud (86.6%) y la especificidad (90.4%) sugieren que el modelo es muy bueno identificando los negativos reales. La precisión positiva y la sensibilidad, ambas alrededor del 80-85%, indican que también tiene un buen desempeño en la predicción de los positivos, aunque con una leve tendencia a ser menos precisa en la predicción de positivos en comparación con los negativos.

Nucleo (KERNEL) Polinomial



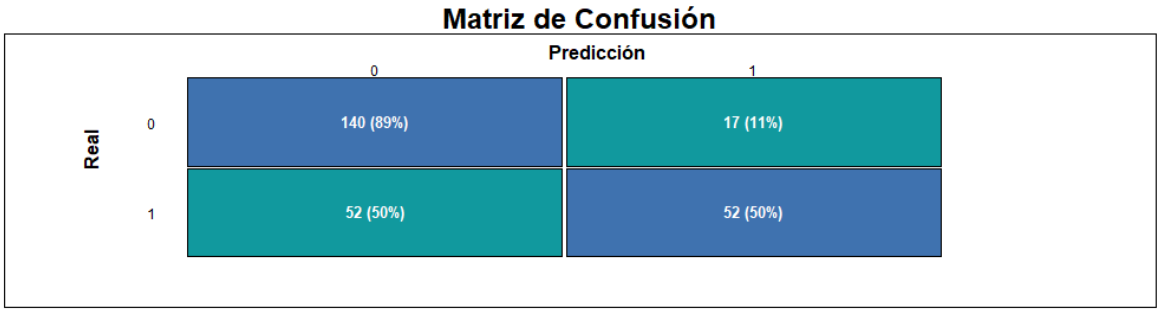




| Precisión.0 | Precisión.1 |
|---|---|
| 91.72 | 82.69 |

| | FN | FP | |
|---|---|---|---|
| VN | 144 | 13 | |
| VP | 18 | 86 | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,881226054 | 88,1226054 |
| Precision Positiva= | VP / FN+VP | 0,826923077 | 82,6923077 |
| Precision Negativa= | VN / VN+FP | 0,917197452 | 91,7197452 |

El modelo muestra un buen rendimiento general con una exactitud del 90.63%. La precisión de 86.87% indica que la mayoría de las predicciones positivas son correctas, mientras que una sensibilidad de 82.69% sugiere que el modelo identifica correctamente la mayoría de los casos positivos reales. La especificidad alta de 91.78% muestra que los negativos reales son bien clasificados. Finalmente, el F1-Score de 84.76% equilibra la precisión y la sensibilidad, indicando un modelo robusto en la clasificación de ambas clases.

Nucleo (KERNEL) Sigmoid





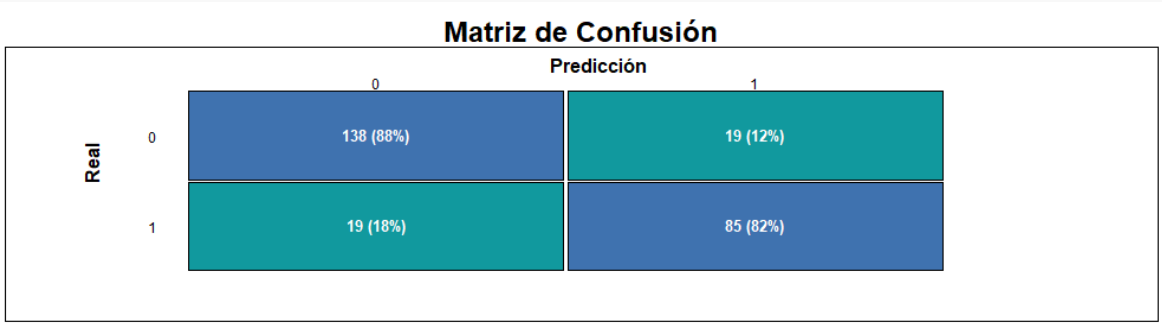
| Precisión.0 | Precisión.1 |
|---|---|
| 89.17 | 50.00 |

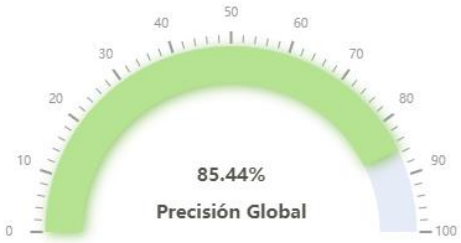
| | FN | FP | |
|---|---|---|---|
| VN | 140 | 17 | |
| VP | 52 | 52 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,735632184 | 73,5632184 |
| Precision Positiva= | VP / FN+VP | 0,5 | 50 |
| Precision Negativa= | VN / VN+FP | 0,891719745 | 89,1719745 |

el modelo tiene un buen rendimiento para identificar las instancias negativas, pero necesita mejoras en la identificación de las instancias positivas, ya que la sensibilidad es relativamente baja.

Repita los dos ejercicios anteriores pero esta vez usando solamente las 4 variables que tienen mejor poder predictivo. ¿Mejoró el resultado?

Nucleo (KERNEL) Linear





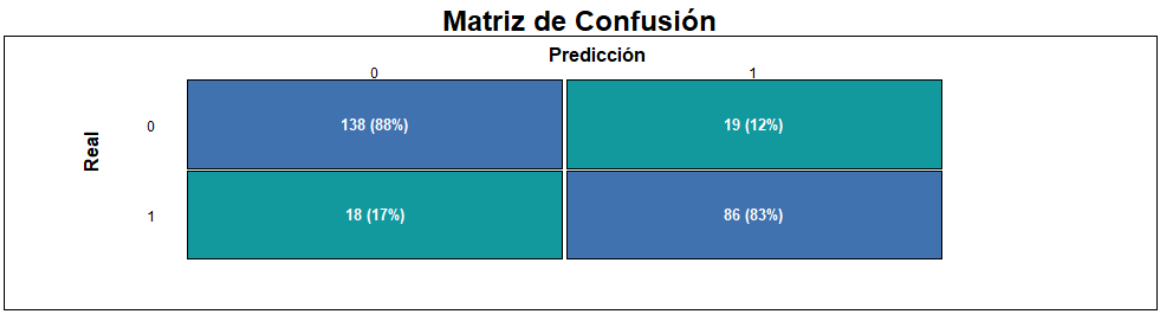
| Precisión.0 | Precisión.1 |
|---|---|
| 87.90 | 81.73 |

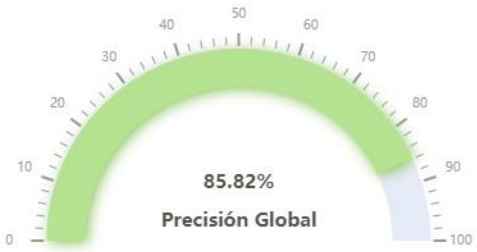
| | FN | FP | |
|---|---|---|---|
| VN | 138 | 19 | |
| VP | 19 | 85 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,85440613 | 85,440613 |
| | | | |
| Precision Positiva= | VP / FN+VP | 0,817307692 | 81,7307692 |
| | | | |
| Precision Negativa= | VN / VN+FP | 0,878980892 | 87,8980892 |

El modelo muestra un buen desempeño general con una precisión global del 85.4%. Es balanceado en la clasificación de las clases, con una precisión y sensibilidad de aproximadamente 81.7% para

la clase positiva y una precisión y especificidad de alrededor de 87.9% para la clase negativa, lo que indica que es ligeramente mejor en la identificación y clasificación correcta de los negativos (Clase 0).

Nucleo (KERNEL) Polinomial



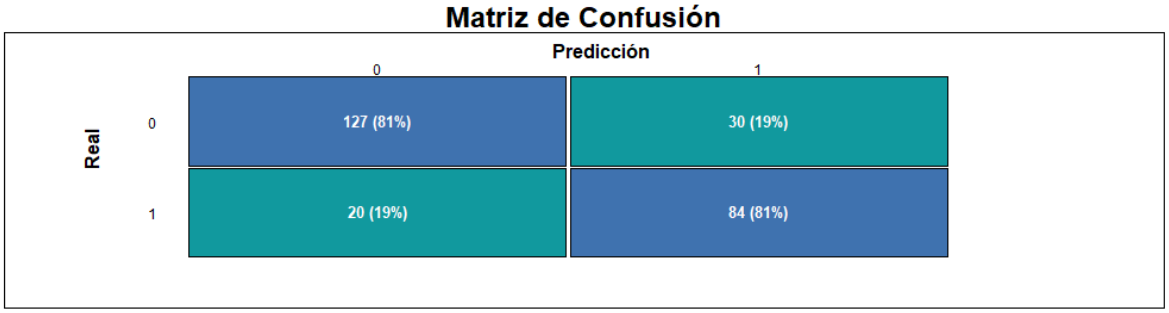


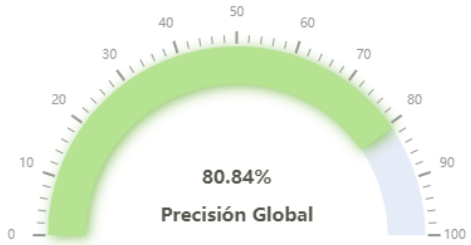
| Precisión.0 | Precisión.1 |
|---|---|
| 87.90 | 82.69 |

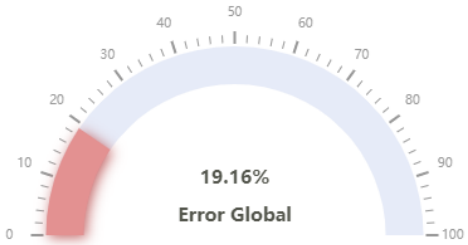
| | FN | FP | |
|---|---|---|---|
| VN | 138 | 19 | |
| VP | 18 | 86 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,858237548 | 85,8237548 |
| | | | |
| Precision Positiva= | VP / FN+VP | 0,826923077 | 82,6923077 |
| | | | |
| Precision Negativa= | VN / VN+FP | 0,878980892 | 87,8980892 |

El análisis de la matriz de confusión muestra que el modelo tiene un buen desempeño general, con una alta exactitud del 85.82%. La precisión y sensibilidad son también altas, lo que sugiere que el modelo es bastante efectivo en identificar tanto las clases positivas como negativas. La especificidad es incluso mayor, lo que indica que el modelo es especialmente bueno para identificar correctamente las clases negativas. En resumen, el modelo tiene un buen balance entre precisión, sensibilidad y especificidad, lo que lo hace confiable para la clasificación en este contexto específico.

Nucleo (KERNEL) Sigmoid





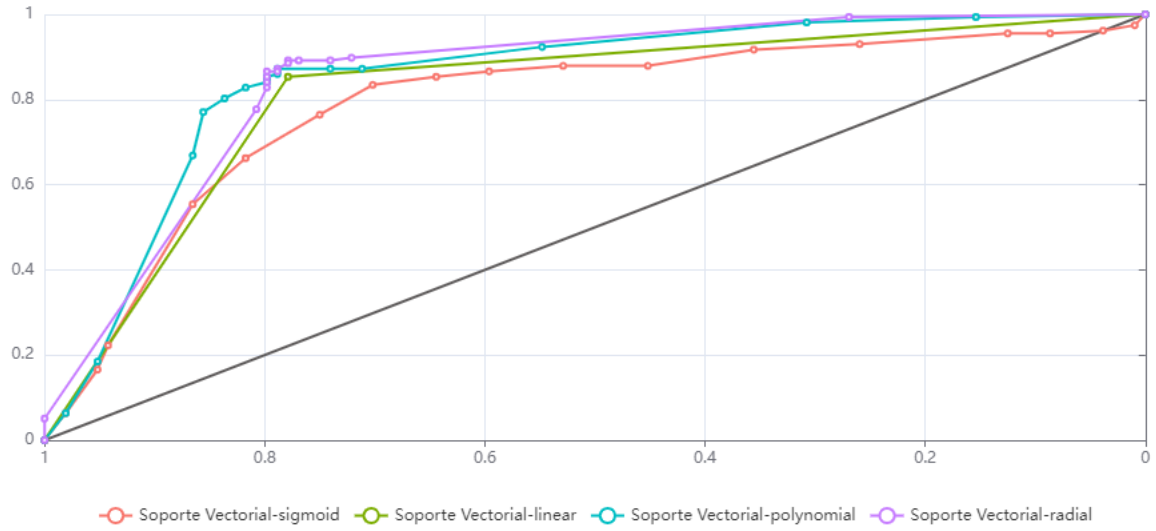

| Precisión.0 | Precisión.1 |
|---|---|
| 80.89 | 80.77 |

| | FN | FP | |
|---|---|---|---|
| VN | 127 | 30 | |
| VP | 20 | 84 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,808429119 | 80,8429119 |
| | | | |
| Precision Positiva= | VP / FN+VP | 0,807692308 | 80,7692308 |
| | | | |
| Precision Negativa= | VN / VN+FP | 0,808917197 | 80,8917197 |

El modelo tiene un buen rendimiento general con una precisión global del 80.8%, indicando que clasifica correctamente la mayoría de las muestras. La precisión positiva del 73.7% sugiere que, cuando el modelo predice la clase 1, tiene una probabilidad razonable de ser correcta, aunque podría mejorar. La sensibilidad y la especificidad, ambas aproximadamente del 80.8% y 80.9% respectivamente, muestran que el modelo identifica bien tanto los verdaderos positivos como los verdaderos negativos. El valor predictivo negativo del 86.4% indica una alta fiabilidad en las predicciones de la clase 0. En conjunto, el modelo es competente, pero hay margen para mejorar en la precisión de las predicciones positivas.

5. Genere la curva ROC para este modelo, ¿es bueno o malo el modelo según esta curva? Compare con la curva ROC con las generadas en las tareas anteriores. ¿Cuál modelo es mejor?

| | Precisión Global | Error Global | 0 | 1 | Área de ROC |
|---|---|---|---|---|---|
| Soporte Vectorial-sigmoid | 75.86207 | 24.13793 | 85.35032 | 61.53846 | 78.88596 |
| Soporte Vectorial-linear | 82.37548 | 17.62452 | 85.35032 | 77.88462 | 83.06284 |
| Soporte Vectorial-polynomial | 83.14176 | 16.85824 | 85.98726 | 78.84615 | 85.07166 |
| Soporte Vectorial-radial | 83.5249 | 16.4751 | 86.6242 | 78.84615 | 84.39184 |




El modelo de Soporte Vectorial con kernel radial sobresale en términos de precisión global, error global y área bajo la curva ROC, con valores del 83.52%, 16.48%, y 86.62% respectivamente, lo que lo posiciona como la opción más sólida entre los modelos evaluados.

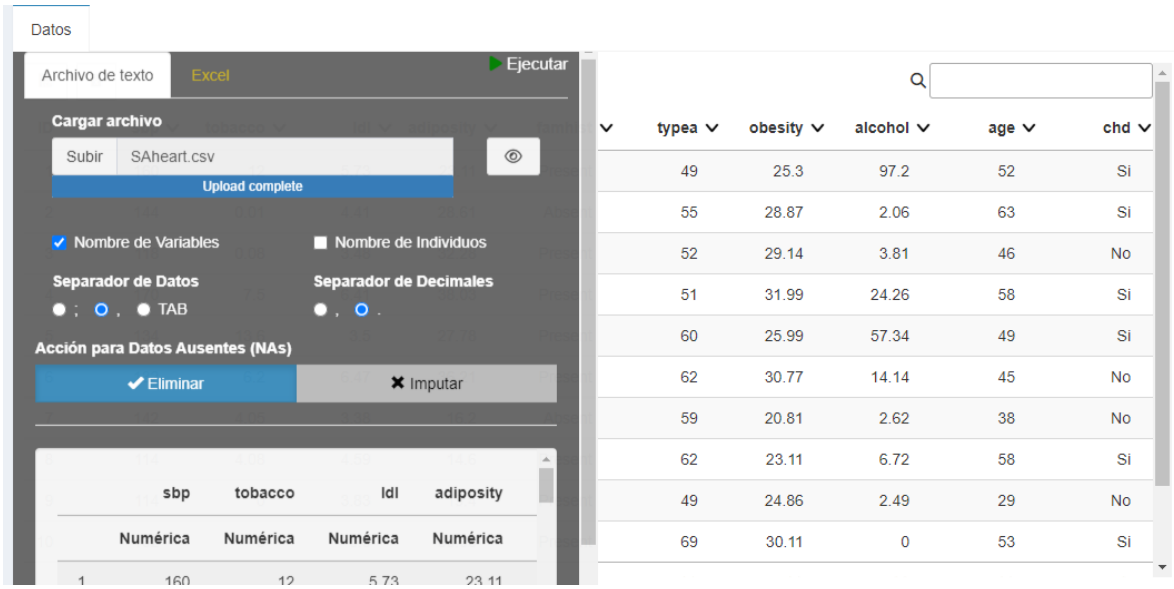
El análisis comparativo entre los diferentes modelos de Soporte Vectorial revela que el kernel sigmoidal presenta la menor precisión global (75.86%), el mayor error global (24.14%), y el área bajo la curva ROC más baja (85.35%), indicando su menor capacidad de clasificación. Por otro lado, el kernel radial se destaca como el más sobresaliente, exhibiendo la más alta precisión global (83.52%), el menor error global (16.48%), y el área bajo la curva ROC más alta (86.62%), lo que sugiere su eficacia superior en la clasificación de datos en comparación con los otros kernels evaluados.

- **Ejercicio 2:** [25 puntos] Esta pregunta utiliza los datos sobre muerte del corazón en Sudáfrica (`SAheart.csv`). La variable que queremos predecir es `chd` que es un indicador de muerte coronaria basado en algunas variables predictivas (factores de riesgo) como son el fumado, la obesidad, las bebidas alcohólicas, entre otras. Las variables son:

  - `sbp`: systolic blood pressure (numérica).
  - `tobacco`: cumulative tobacco (kg) (numérica).
  - `ldl`: low densiity lipoprotein cholesterol (numérica).
  - `Adiposity`: Adiposity level (numérica).
  - `famhist`: family history of heart disease (Present, Absent) (categórica).
  - `typea`: type-A behavior (numérica).
  - `Obesity`: Obesity of the person (numérica).
  - `alcohol`: current alcohol consumption (numérica).
  - `age`: age at onset (numérica).
  - `chd`: coronary heart disease (categórica).

  Realice lo siguiente:

1. Use el método de Máquinas de Soporte Vectorial para generar modelos predictivos para la tabla `SAheart.csv`. Para esto utilice el 80 % de los datos para la tabla aprendizaje y un 20 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las dos categorías.



Datos

| ID | Partición | sbp | tobacco | ldl | adiposity |
|---|---|---|---|---|---|
| 1 | Aprendizaje | 160 | 12 | 5.73 | 23.11 |
| 2 | Aprendizaje | 144 | 0.01 | 4.41 | 28.61 |
| 3 | Aprendizaje | 118 | 0.08 | 3.48 | 32.28 |
| 4 | Prueba | 170 | 7.5 | 6.41 | 38.03 |
| 5 | Prueba | 134 | 13.6 | 3.5 | 27.78 |
| 6 | Aprendizaje | 132 | 6.2 | 6.47 | 36.21 |
| 7 | Aprendizaje | 142 | 4.05 | 3.38 | 16.2 |
| 8 | Aprendizaje | 114 | 4.08 | 4.59 | 14.6 |
| 9 | Prueba | 114 | 0 | 3.83 | 19.4 |
| 10 | Aprendizaje | 132 | 0 | 5.8 | 30.96 |

Carga de datos

Opciones

Ejecutar

Seleccionar la variable a predecir

chd

Aprendizaje - Prueba | Validación Cruzada

Semilla Aleatoria

Habilitada | Deshabilitada | 5

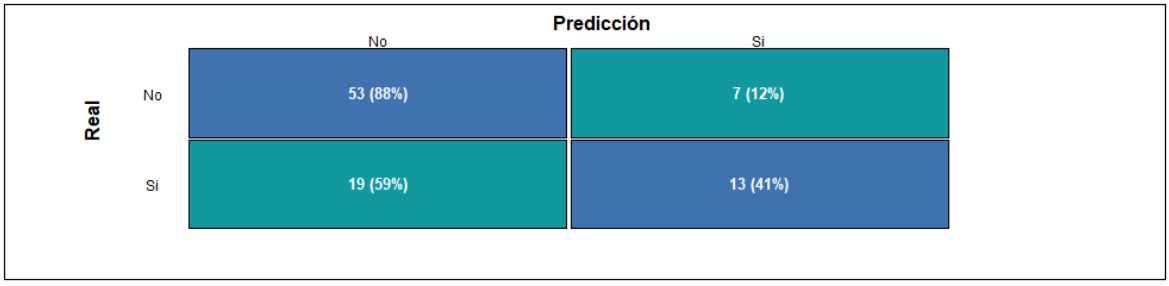
Aprendizaje | Prueba

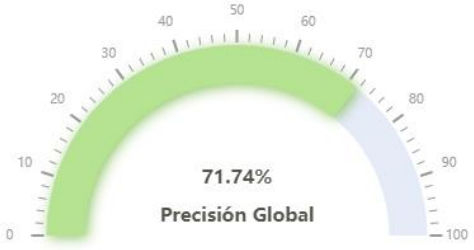
5 | 80 | 95

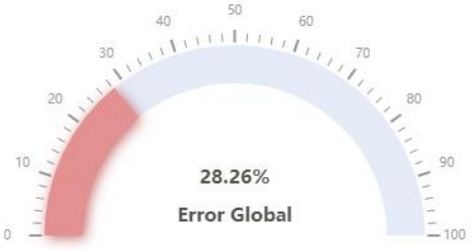
5 15 25 35 45 55 65 75 85 95

Opciones

## Matriz de Confusión




| Precisión.No | Precisión.Si |
|---|---|
| 88.33 | 40.62 |

| | FN | FP | |
|---|---|---|---|
| VN | 53 | 7 | |
| VP | 19 | 13 | |

| | | | |
|---|---|---|---|
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,717391304 | 71,7391304 |
| Precision Positiva= | VP / FN+VP | 0,40625 | 40,625 |
| Precision Negativa= | VN / VN+FP | 0,883333333 | 88,3333333 |

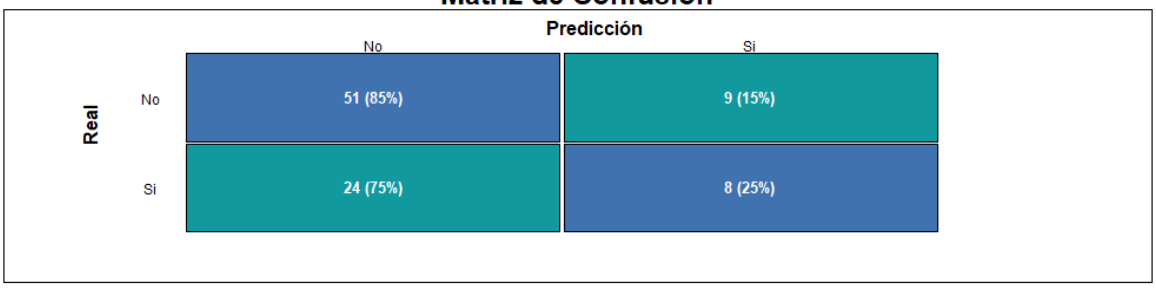
En conclusión, el modelo tiene una precisión global decente del 72%, pero su precisión positiva es relativamente baja, lo que sugiere que tiene dificultades para identificar correctamente los casos positivos. Sin embargo, la precisión negativa es bastante alta, lo que indica que el modelo es bueno para predecir los casos negativos. Dependiendo del contexto, esto puede requerir ajustes para mejorar la capacidad del modelo para identificar los casos positivos correctamente.

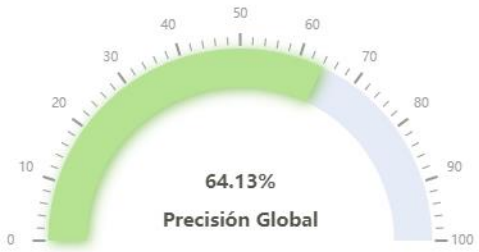
2. Repita los dos ejercicios anteriores pero esta vez usando solamente las 4 variables que tienen mejor poder predictivo. ¿Mejoró el resultado?

Datos

| ID | Partición | sbp | tobacco | ldl | adiposity | famhist | typea | obesity | alcohol | age |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Aprendizaje | 160 | 12 | 5.73 | 23.11 | Present | 49 | 25.3 | 97.2 | 52 |
| 2 | Prueba | 144 | 0.01 | 4.41 | 28.61 | Absent | 55 | 28.87 | 2.06 | 63 |
| 3 | Aprendizaje | 118 | 0.08 | 3.48 | 32.28 | Present | 52 | 29.14 | 3.81 | 46 |
| 4 | Aprendizaje | 170 | 7.5 | 6.41 | 38.03 | Present | 51 | 31.99 | 24.26 | 58 |
| 5 | Aprendizaje | 134 | 13.6 | 3.5 | 27.78 | Present | 60 | 25.99 | 57.34 | 49 |
| 6 | Prueba | 132 | 6.2 | 6.47 | 36.21 | Present | 62 | 30.77 | 14.14 | 45 |
| 7 | Prueba | 142 | 4.05 | 3.38 | 16.2 | Absent | 59 | 20.81 | 2.62 | 38 |
| 8 | Aprendizaje | 114 | 4.08 | 4.59 | 14.6 | Present | 62 | 23.11 | 6.72 | 58 |
| 9 | Aprendizaje | 114 | 0 | 3.83 | 19.4 | Present | 49 | 24.86 | 2.49 | 29 |
| 10 | Aprendizaje | 132 | 0 | 5.8 | 30.96 | Present | 69 | 30.11 | 0 | 53 |

## Matriz de Confusión





| Precisión.No | Precisión.Si |
|---|---|
| 85.00 | 25.00 |

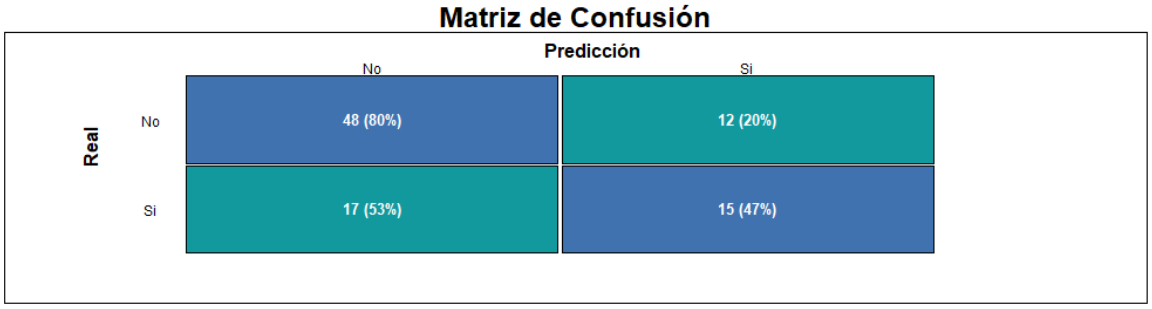
Modelo Predictivo

| | FN | FP | |
|---|---|---|---|
| VN | 51 | 9 | |
| VP | 24 | 8 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,641304348 | 64,1304348 |
| | | | |
| Precision Positiva= | VP / FN+VP | 0,25 | 25 |
| | | | |
| Precision Negativa= | VN / VN+FP | 0,85 | 85 |

El modelo exhibe deficiencias significativas en la identificación tanto de casos positivos como negativos, reflejadas en una baja sensibilidad y precisión. Aunque muestra una alta especificidad, lo que indica una buena capacidad para identificar casos negativos, la presencia de un número considerable de falsos positivos y negativos sugiere limitaciones serias en su capacidad predictiva. En general, el desempeño del modelo se inclina hacia el lado negativo debido a su dificultad para capturar la verdadera señal en los datos y su propensión a clasificar erróneamente los casos.

3. Repita los dos ejercicios anteriores pero esta vez usando los otros núcleos (`Kernel` que permite en método **Máquinas de Soporte Vectorial**. ¿Mejoró alguno el resultado?

Nucleo Kernel (Lineal)



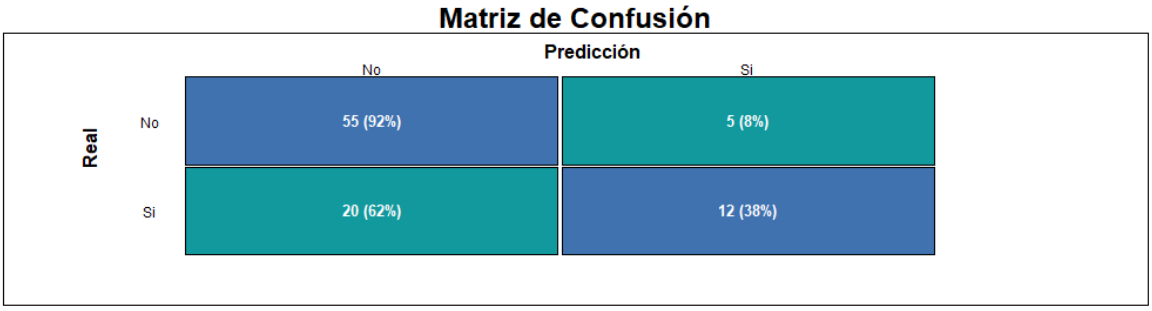


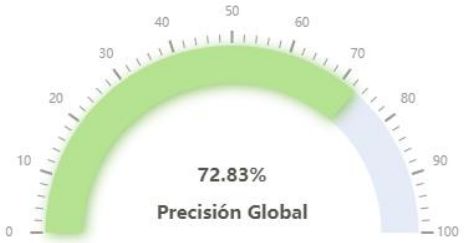

| Precisión.No | Precisión.Si |
|---|---|
| 80.00 | 46.88 |

| | Modelo Predictivo | | |
|---|---|---|---|
| | FN | FP | |
| VN | 48 | 12 | |
| VP | 17 | 15 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,684782609 | 68,4782609 |
| | | | |
| Precision Positiva= | VP / FN+VP | 0,46875 | 46,875 |
| | | | |
| Precision Negativa= | VN / VN+FP | 0,8 | 80 |

Basado en los resultados de la matriz de confusión y las métricas de evaluación, el modelo muestra un rendimiento moderado. Con una precisión global del 68.5%, el modelo clasifica correctamente un porcentaje significativo de las instancias, aunque la precisión positiva de alrededor del 55.6% sugiere que tiene dificultades para identificar adecuadamente las instancias de la clase positiva. Sin embargo, la precisión negativa del 73.8% indica una mejor capacidad para predecir la clase negativa. En general, el modelo tiene margen de mejora, especialmente en la precisión de las predicciones positivas, lo que sugiere que no es óptimo pero tampoco es malo.

Nucleo (KERNEL) Polinomial





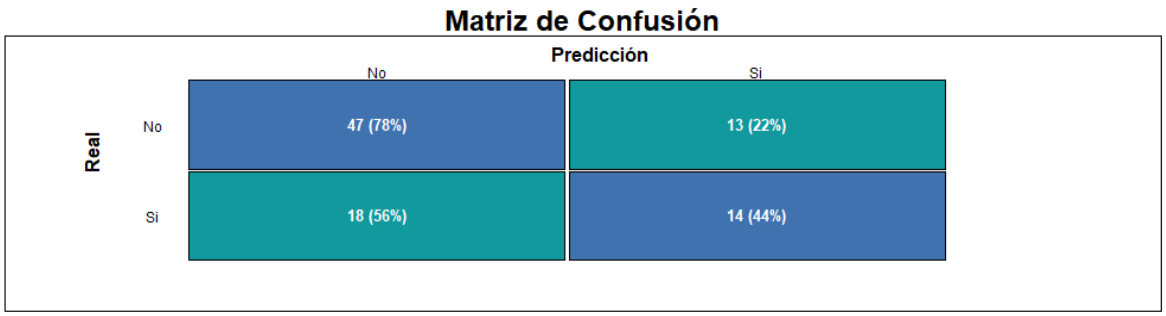


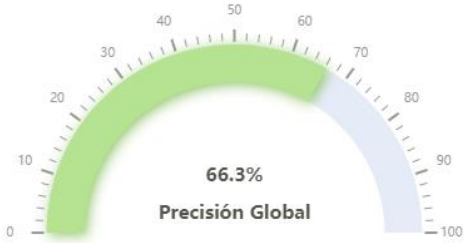
| Precisión.No | Precisión.Si |
|---|---|
| 91.67 | 37.50 |

| | Modelo Predictivo | | |
|---|---|---|---|
| | FN | FP | |
| VN | 55 | 5 | |
| VP | 20 | 12 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,72826087 | 72,826087 |
| | | | |
| Precision Positiva= | VP / FN+VP | 0,375 | 37,5 |
| | | | |
| Precision Negativa= | VN / VN+FP | 0,916666667 | 91,6666667 |

La matriz de confusión revela que el modelo presenta una precisión global del 73%, lo que indica que la mayoría de las predicciones son correctas. Sin embargo, la sensibilidad es baja (38%), lo que sugiere una dificultad para identificar correctamente los casos positivos, mientras que la alta especificidad (92%) indica una buena capacidad para identificar los casos negativos. La predicción positiva es baja (18%), lo que indica una tendencia a subestimar los casos positivos, mientras que la predicción negativa es alta (82%), sugiriendo una tendencia a sobreestimar los casos negativos. En resumen, aunque el modelo tiene una precisión aceptable en general, podría mejorar en la identificación de casos positivos.

Nucleo (KERNEL) Sigmoid





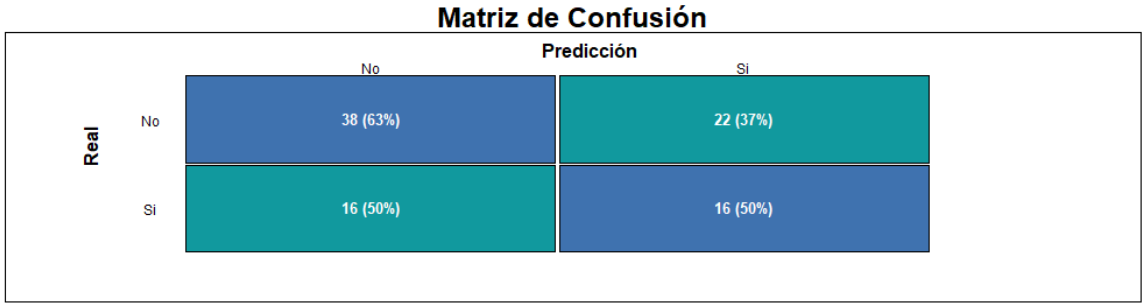
| Precisión.No | Precisión.Si |
|---|---|
| 78.33 | 43.75 |

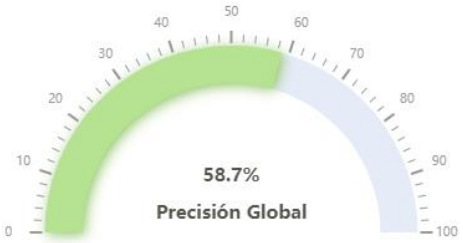
| | FN | FP | |
|---|---|---|---|
| VN | 47 | 13 | |
| VP | 18 | 14 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,663043478 | 66,3043478 |
| | | | |
| Precision Positiva= | VP / FN+VP | 0,4375 | 43,75 |
| | | | |
| Precision Negativa= | VN / VN+FP | 0,783333333 | 78,3333333 |

El análisis de la matriz de confusión revela que el modelo tiene una precisión global del 66.30%, con una precisión de predicción positiva del 51.85% y una precisión de predicción negativa del 72.31%. Aunque el modelo clasifica correctamente una parte significativa de las muestras, su precisión, especialmente en la predicción positiva, podría ser mejorada. Por lo tanto, el modelo se considera moderado pero podría beneficiarse de mejoras para ser considerado como bueno.

Repita los dos ejercicios anteriores pero esta vez usando solamente las 4 variables que tienen mejor poder predictivo. ¿Mejoró el resultado?

Nucleo (Kernel) Linear



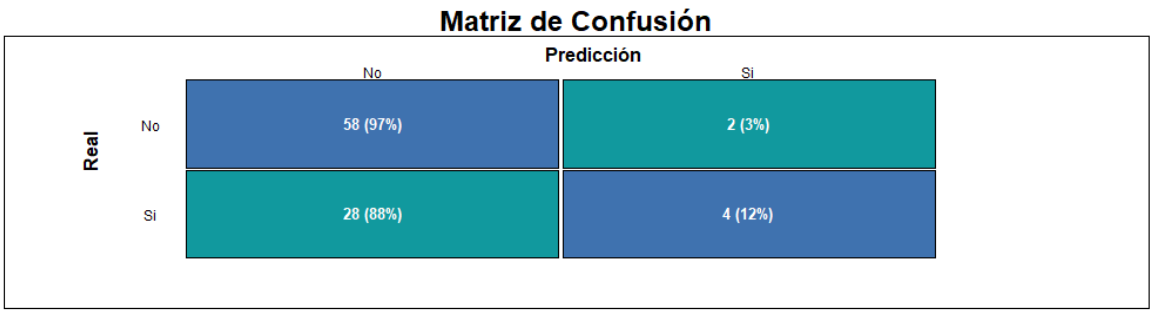


| Precisión.No | Precisión.Si |
|---|---|
| 63.33 | 50.00 |

| | FN | FP | |
|---|---|---|---|
| VN | 38 | 22 | |
| VP | 16 | 16 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,586956522 | 58,6956522 |
| | | | |
| Precision Positiva= | VP / FN+VP | 0,5 | 50 |
| | | | |
| Precision Negativa= | VN / VN+FP | 0,633333333 | 63,3333333 |

Basándonos únicamente en la precisión proporcionada por la matriz de confusión, el modelo exhibe una capacidad moderada para predecir correctamente las clases. Sin embargo, una evaluación completa del modelo requeriría el análisis de otras métricas, como sensibilidad y especificidad, así como la consideración del contexto del problema y las implicaciones de los resultados erróneos. Por lo tanto, no podemos concluir de manera definitiva si el modelo es bueno o malo sin una evaluación más detallada y un entendimiento más profundo del dominio y los requisitos específicos del problema.

Nucleo (KERNEL) Polinomial



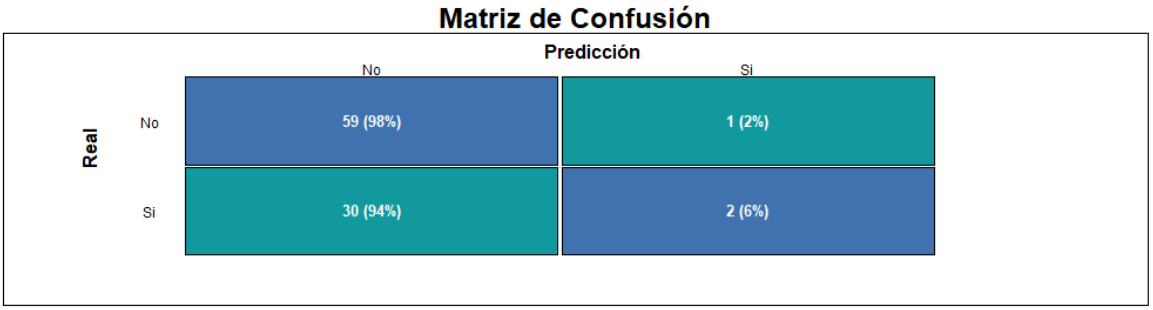


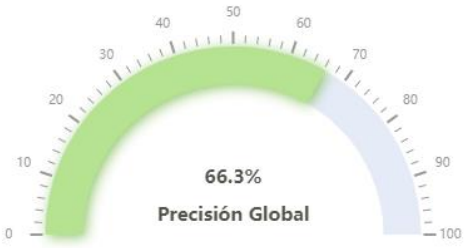
| Precisión.No | Precisión.Si |
|---|---|
| 96.67 | 12.50 |

| | FN | FP | |
|---|---|---|---|
| VN | 58 | 2 | |
| VP | 28 | 4 | |
| | | | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,673913043 | 67,3913043 |
| | | | |
| Precision Positiva= | VP / FN+VP | 0,125 | 12,5 |
| | | | |
| Precision Negativa= | VN / VN+FP | 0,966666667 | 96,6666667 |

el modelo exhibe una precisión global moderada, con una alta precisión negativa pero una baja precisión positiva. Esto sugiere que el modelo es efectivo para predecir la clase negativa pero tiene dificultades para identificar correctamente los casos positivos. En general, el desempeño del modelo podría considerarse como mediocre, ya que no logra una precisión equilibrada en ambas clases. Por lo tanto, no se puede decir que sea un modelo bueno, pero tampoco es completamente malo; hay margen para mejorar, especialmente en la identificación de la clase positiva.

Nucleo (KERNEL)  Sigmoid





| Precisión.No | Precisión.Si |
|---|---|
| 98.33 | 6.25 |

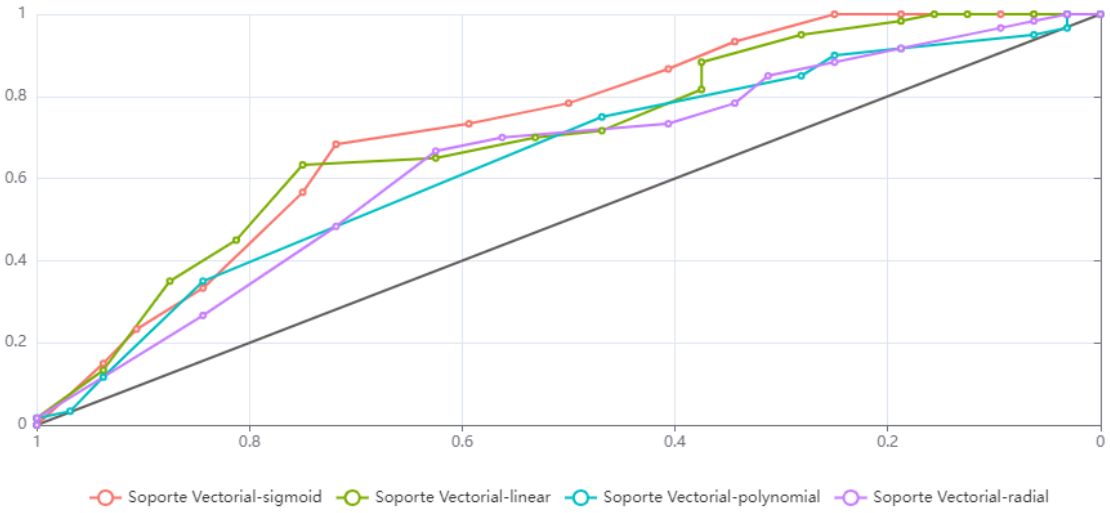
| | FN | FP | |
|---|---|---|---|
| VN | 59 | 1 | |
| VP | 30 | 2 | |
| Precision Global= | VN+VP / VN+FP+FN+VP | 0,663043478 | 66,3043478 |
| Precision Positiva= | VP / FN+VP | 0,0625 | 6,25 |
| Precision Negativa= | VN / VN+FP | 0,983333333 | 98,3333333 |

el modelo muestra una alta precisión en la predicción de casos negativos, lo que sugiere que es eficaz para identificar verdaderos negativos. Sin embargo, su capacidad para predecir

correctamente casos positivos es notablemente baja, indicando una sensibilidad deficiente. En consecuencia, aunque el modelo puede ser útil para descartar casos negativos, su utilidad en la identificación de casos positivos es limitada. En general, esto sugiere que el modelo tiene margen de mejora y no se consideraría óptimo en su estado actual.

4. Genere la curva ROC para este modelo, ¿es bueno o malo el modelo según esta curva? Compare con la curva ROC con las generadas en las tareas anteriores. ¿Cuál modelo es mejor?

| | Precisión Global | Error Global | No | Si | Área de ROC |
|---|---|---|---|---|---|
| Soporte Vectorial-sigmoid | 71.73913 | 28.26087 | 90 | 37.5 | 72.1875 |
| Soporte Vectorial-linear | 68.47826 | 31.52174 | 85 | 37.5 | 69.94792 |
| Soporte Vectorial-polynomial | 66.30435 | 33.69565 | 95 | 12.5 | 66.09375 |
| Soporte Vectorial-radial | 65.21739 | 34.78261 | 86.66667 | 25 | 64.21875 |




Observando los datos proporcionados en la tabla, podemos ver que el modelo de Soporte Vectorial-sigmoid tiene la mayor precisión global (71.73913%) y el área bajo la curva ROC más alta (72.1875). Sin embargo, también tiene un alto error global (28.26087%).

El modelo de Soporte Vectorial-linear tiene una precisión global ligeramente menor (68.47826%) pero un error global un poco más bajo (31.52174%). Su área bajo la curva ROC también es bastante alta (69.94792).

El modelo de Soporte Vectorial-polynomial tiene la precisión global más baja (66.30435%), un error global bastante alto (33.69565%), y el área bajo la curva ROC más baja (66.09375).

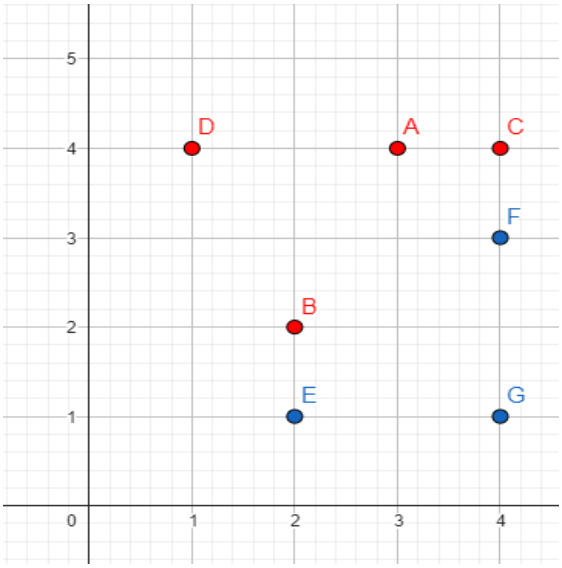
El modelo de Soporte Vectorial-radial tiene una precisión global similar al modelo polynomial (65.21739%), un error global también alto (34.78261%), y un área bajo la curva ROC más alta que polynomial pero aún más baja que los otros dos modelos (64.21875).

Basándonos únicamente en los datos proporcionados, el modelo de Soporte Vectorial-sigmoid parece ser el mejor en términos de precisión global y área bajo la curva ROC, a pesar de tener un error global alto. Sin embargo, la elección del mejor modelo también puede depender de otros factores como la sensibilidad, la especificidad y el equilibrio entre estas métricas.

- **Pregunta 3:** [25 puntos] Suponga que se tiene la siguiente tabla de datos:

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 3 | 4 | Rojo |
| 2 | 2 | Rojo |
| 4 | 4 | Rojo |
| 1 | 4 | Rojo |
| 2 | 1 | Azul |
| 4 | 3 | Azul |
| 4 | 1 | Azul |

1. Dibuje el hiperplano (recta) óptimo de separación y calcule la ecuación para este hiperplano de la forma $f(x) = Mx + B$.




El hiperplano óptimo de separación se determina mediante la maximización de la distancia entre las clases y se representa mediante la ecuación f(x) = Mx + B, donde M es la pendiente y B es el términode sesgo.
En este caso, haciendo uso de álgebra básica podemos encontrar el hiperplano de separación (en el caso de dos dimensiones una recta) sin necesidad de derivar, usando el punto medio entre los puntos B y E que es P1= (2,1.5) y el punto medio

entre C y F que es P2= (4,3.5) con la fórmula usual de la pendiente y la recta y=mx+b, se puede concluir que es: y=x o f(x)=x.

PENDIENTE

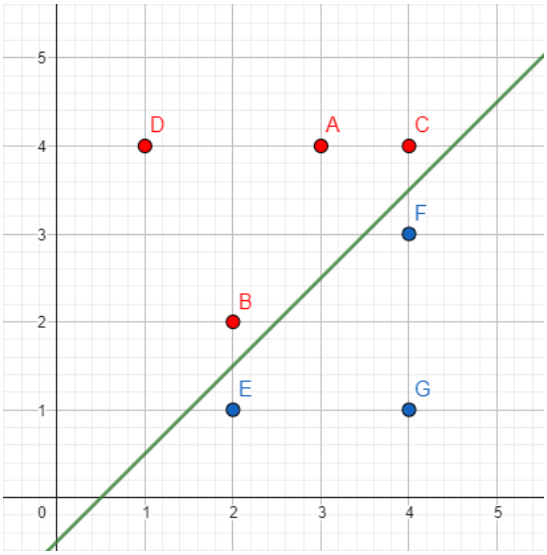
$$m = \frac{Y2 - Y1}{X2 - X1}$$

$$b = y - mx = 1.5 - 1 * 2 = -0.5$$

RECTA

$$f(x) = mx + b$$

$$f(x) = x - 0.5$$



2. Dé la regla de clasificación para el clasificador con margen máximo. Debe ser algo como lo siguiente: $z = (z_1, z_2)$ se clasifica como Rojo si $f(z_1) \geq z_2$ y si $f(z_1) < z_2$ se clasifica como Azul.

Para el clasificador con margen máximo basado en el hiperplano de separación óptimo, la regla de clasificación se puede establecer de la siguiente manera:

Dado un punto z = ($z$1, $z$2), se clasificará como "Rojo" si f($z$1) ≥ $z$2, y se clasificará como "Azul" si f($z$1) < $z$2.

En esta regla, f($z$1) representa el valor de la función f(x) en el punto $z$1 del eje x. Si el valor de f($z$1) es mayor o igual que el valor de $z$2 en el eje y, entonces el punto se clasifica como "Rojo". Por otro lado, si el valor de f($z$1) es menor que el valor de $z$2, entonces el punto se clasifica como "Azul".

Entonces se tiene que $z$ = ($z$1, $z$2)

Si $f$($z$1) ≥ $z$2 se clasifica como rojo y Si $f$($z$1) < $z$2 se clasifica como azul.

Ejemplo:

Si tenemos el punto z = (1, 2) y la ecuación del hiperplano de separación es f(x) = x - 0.5, podemos aplicar la regla de clasificación para determinar a qué clase pertenece.

Calculamos el valor de f($z$1):

f($z$1) = 1 - 0.5 = 0.5

Luego comparamos f($z$1) con z2:

0.5 < 2

Dado que 0.5 es menor que 2, según la regla de clasificación establecida, el punto z = (1, 2) se clasifica como "Azul".

Esta regla de clasificación nos permite asignar una clase (Rojo o Azul) a un punto de datos en función de su posición en relación con el hiperplano de separación óptimo.

3. Indique los vectores de soporte para el clasificador de margen máximo y encuentre las ecuaciones $h(x) = Mx + B$ y $g(x) = Mx + B$ de las rectas paralelas que definen el margen máximo.

Para encontrar los vectores de soporte y las ecuaciones de las rectas paralelas que definen el margen máximo en un clasificador de margen máximo, necesitamos utilizar el algoritmo de Support Vector Machine (SVM) y obtener los vectores de soporte y los coeficientes del hiperplano óptimo.

- Encontrar h(x)

PENDIENTE

Usando el punto B (2,2)

$$m = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{4-2}{4-2} = 1$$

Se tiene h(x)= mx + b

$b = y - mx = 2 - 1 * 2 = 0$

h(x)=x

Encontrar g(x)

Usando los puntos E (2,1) y F (4,3)

PENDIENTE

Usando el punto B (2,1)

$$m = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{3-1}{4-2} = 1$$

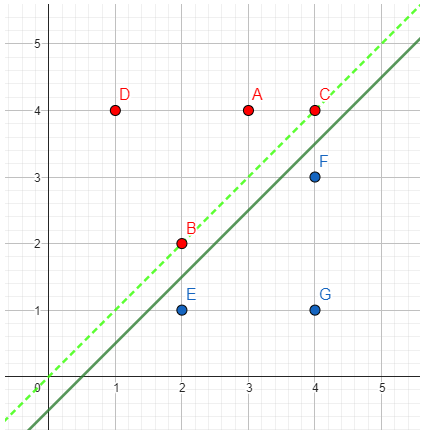
Se tiene

$$b = y - mx = 2 - 1 * 1 = -1$$

$$g(x) = mx + b$$

$$g(x) = x - 1$$

4. En su dibujo indique el margen para el hiperplano (recta) de margen máximo.




5. Explique porqué un ligero movimiento de la séptima observación (fila) NO afectaría el hiperplano (recta) de separación.

La posición de un punto individual tiene un impacto limitado en la ubicación y orientación del hiperplano de separación. Esto se debe a que el objetivo principal del SVM es maximizar el margen entre las clases, y para lograrlo, se centra en los puntos más cercanos al hiperplano de separación, conocidos como vectores de soporte. Estos vectores de soporte determinan la posición y orientación del hiperplano. Un ligero movimiento en una única observación generalmente no cambiará la posición relativa de los vectores de soporte y, por lo tanto, no afectará significativamente el hiperplano de separación. El algoritmo SVM es resistente a cambios pequeños en los datos y se enfoca en los puntos más críticos para la clasificación.

Sin embargo, si ese ligero movimiento resultara en que la séptima observación se convirtiera en un vector de soporte, entonces sí podría tener un impacto en el hiperplano de separación, ya que los vectores de soporte determinan la posición del hiperplano. En ese caso, el hiperplano se ajustaría para adaptarse al nuevo conjunto de vectores de soporte y maximizar el margen entre las clases.

Los puntos B, E, C y F se utilizan para establecer el hiperplano con el margen mínimo y máximo, ya que están en el límite de la intersección entre los puntos azules y rojos. Si se produce un ligero movimiento en el séptimo punto, G, no afectaría el hiperplano de separación, ya que G se encuentra dentro del área de los puntos azules y no está en el límite. En este caso, G no sería considerado un vector de soporte y, por lo tanto, su movimiento no tendría un impacto significativo en la posición y orientación del hiperplano de separación. El hiperplano de separación se determina principalmente por los puntos límite que definen el margen entre las clases.

6. Dibuje un hiperplano (recta) se separación pero que NO es el hiperplano óptimo de separación y calcule la ecuación para este hiperplano (recta) de la forma $q(x) = Mx + B$.

Para dibujar un hiperplano de separación que no sea el hiperplano óptimo, podemos seleccionar una recta que separe las clases, pero no maximice el margen entre ellas. Sin embargo, es importante tener en cuenta que este hiperplano no será el óptimo y puede resultar en una menor capacidad de generalización.

Si tomamos los puntos P1(2, 1.8) y P2(4, 3.8) para calcular la ecuación de un hiperplano de separación, podemos utilizar estos puntos y la fórmula de la pendiente para encontrar los valores de M y B.

Calculamos la pendiente M:

$$m = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{3.8 - 1.8}{4 - 2} = 1$$

Luego, utilizamos uno de los puntos, por ejemplo, P1(2, 1.8), y la pendiente M para calcular el valorde B:

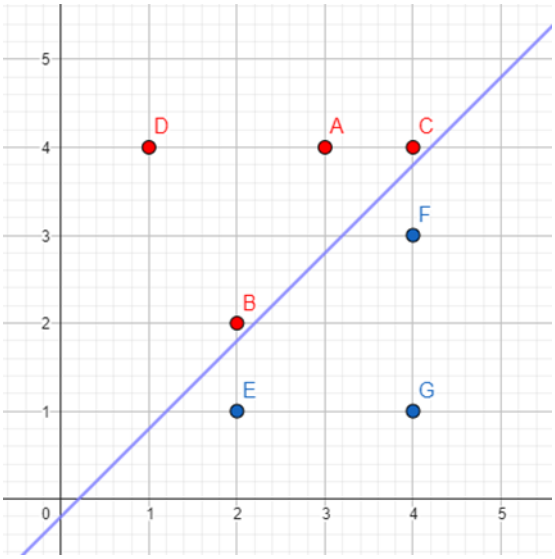
$$b = y - mx = 1.8 - 1 * 2 = -0.2$$

Por lo tanto, la ecuación del hiperplano de separación sería:

$$q(x) = mx + b$$

$$q(x) = x - 0.2$$

Este hiperplano de separación se ha determinado utilizando los puntos P1 y P2, y representa una línearecta que separa los puntos de las dos clases. Sin embargo, es importante destacar que este hiperplanono es necesariamente el hiperplano óptimo de separación, ya que para determinar el hiperplano óptimose requiere un algoritmo de aprendizaje automático, como el SVM, que maximiza el margen entre lasclases.



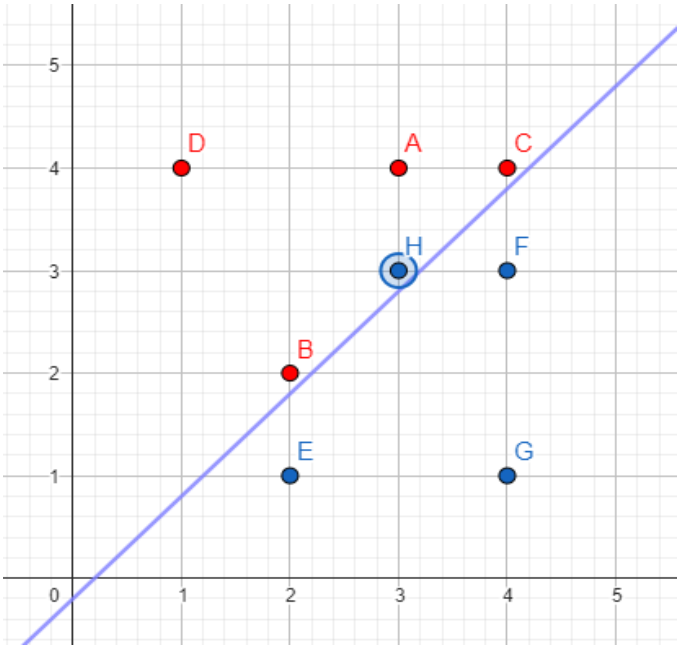

7. Dibuje una observación (fila) adicional de manera que las dos clases ya NO sean separables por un hiperplano (recta).

Si deseamos agregar una observación adicional de manera que las dos clases ya no sean separables por un hiperplano (recta), podemos agregar un punto que esté en la región de intersección entre las dos clases.

Para hacer que las dos clases no sean separables por un hiperplano, podemos agregar un punto adicional en la región de intersección, por ejemplo

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 3 | 4 | Rojo |
| 2 | 2 | Rojo |

| | | |
|---|---|---|
| 4 | 4 | Rojo |
| 1 | 4 | Rojo |
| 2 | 1 | Azul |
| 4 | 3 | Azul |
| 4 | 1 | Azul |
| 3 | 3 | Rojo |




Al agregar el punto (3, 3) como una observación adicional, ahora las dos clases ya no pueden ser separadas perfectamente por un hiperplano (recta). Esto se debe a que hay un punto rojo en la región de los puntos azules.

Es importante destacar que cuando las clases ya no son separables por un hiperplano, se pueden utilizar otros enfoques de clasificación, como algoritmos no lineales, para lograr una mejor separación.