Facultad de Estudios Superiores de Acatlán

Práctica 3:
Algoritmos de Clasificación
para la Predicción de Cáncer de Mama:

Regresión Logística
Análisis Discriminante Lineal (LDA)
K Vecinos más Cercanos (KNN)
Máquinas de Vector Soporte (SVM)
Naive Bayes Classifier (NBC)

Diplomado en Ciencia de Datos

Módulo II:
Modelado Supervisado

Profesora:
Act. Lorena Pineda Rodríguez

Alumno:
Francisco Roman Peña de la Rosa

Ciudad de México Agosto, 2021

Planteamiento del Problema

De la información proporcionada en la práctica tenemos:

Poniendo a prueba la experiencia del médico practicante nos dio como resultado la siguiente matriz de confusión con 171 registros.

[[56 10]

[13 92]]

Interpretación de los resultados:

Tenemos 171 pacientes en nuestro conjunto de validación.

- De los 66 pacientes que se predijo que **No** tenían cáncer, 13 se diagnosticaron sin cáncer cuando en realidad si tenían la enfermedad (**Error Tipo 1**).
- De los 105 pacientes que se predijo que, **Si** tenían cáncer, 10 se diagnosticaron con cáncer cuando en realidad no tenían la enfermedad (**Error Tipo 2**).

Interpretación de Resultados

Modelado Clases Desbalanceadas

Regresión Logística

De este modelo obtuvimos la siguiente matriz de confusión:

[[29, 0],

[0, 97]]

Interpretación de los resultados:

Tenemos 126 pacientes en nuestro conjunto de validación.

- De los 29 pacientes que se predijo que, **No** tenían cáncer, 0 se diagnosticaron sin cáncer cuando en realidad si tenían la enfermedad (**Error Tipo 1**).

- De los 92 pacientes que se predijo que, **Si** tenían cáncer, 0 se diagnosticaron con cáncer cuando en realidad no tenían la enfermedad (**Error Tipo 2**).

Análisis Discriminante Lineal (LDA)

De este modelo obtuvimos la siguiente matriz de confusión:

[[24, 4],

[2, 95]]

Interpretación de los resultados:

Tenemos 125 pacientes en nuestro conjunto de validación.

- De los 28 pacientes que se predijo que, **No** tenían cáncer, 2 se diagnosticaron sin cáncer cuando en realidad si tenían la enfermedad (**Error Tipo 1**).
- De los 97 pacientes que se predijo que, **Si** tenían cáncer, 4 se diagnosticaron con cáncer cuando en realidad no tenían la enfermedad (**Error Tipo 2**).

K Vecinos más Cercanos

De este modelo obtuvimos la siguiente matriz de confusión:

[[28, 1],

[0, 97]]

Interpretación de los resultados:

Tenemos 126 pacientes en nuestro conjunto de validación.

- De los 29 pacientes que se predijo que, **No** tenían cáncer, 0 se diagnosticaron sin cáncer cuando en realidad si tenían la enfermedad (**Error Tipo 1**).
- De los 97 pacientes que se predijo que, **Si** tenían cáncer, 1 se diagnosticaron con cáncer cuando en realidad no tenían la enfermedad (**Error Tipo 2**).

Máquinas de Vector Soporte

De este modelo obtuvimos la siguiente matriz de confusión:

[[9, 20],

[5, 92]]

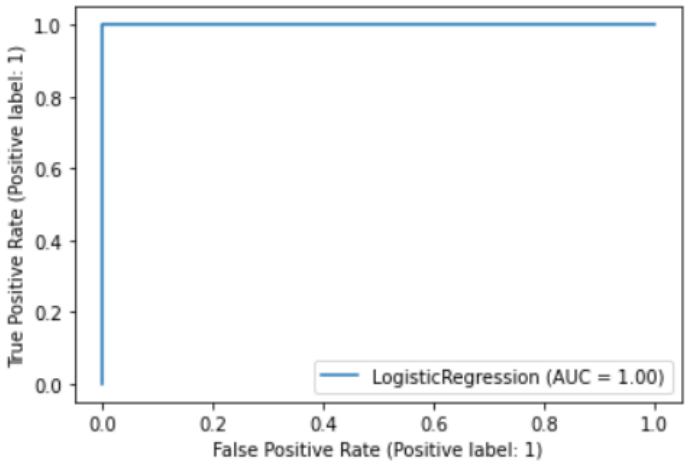
Interpretación de los resultados:

Tenemos 126 pacientes en nuestro conjunto de validación.

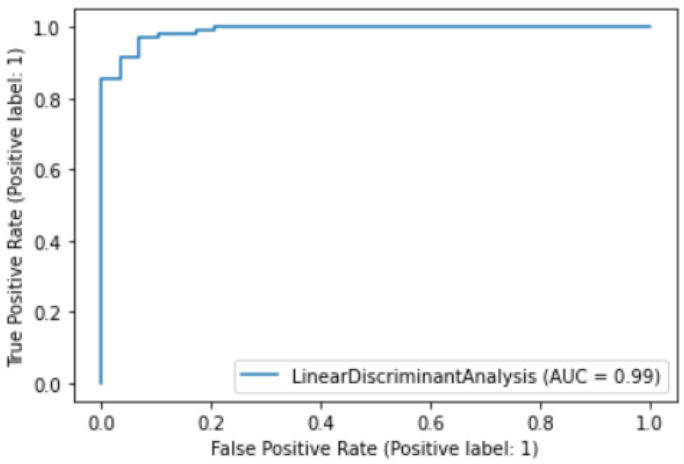
- De los 29 pacientes que se predijo que, **No** tenían cáncer, 5 se diagnosticaron sin cáncer cuando en realidad si tenían la enfermedad (**Error Tipo 1**).
- De los 97 pacientes que se predijo que, **Si** tenían cáncer, 20 se diagnosticaron con cáncer cuando en realidad no tenían la enfermedad (**Error Tipo 2**).

Curvas ROC – Modelos Sin Balanceo de Clases

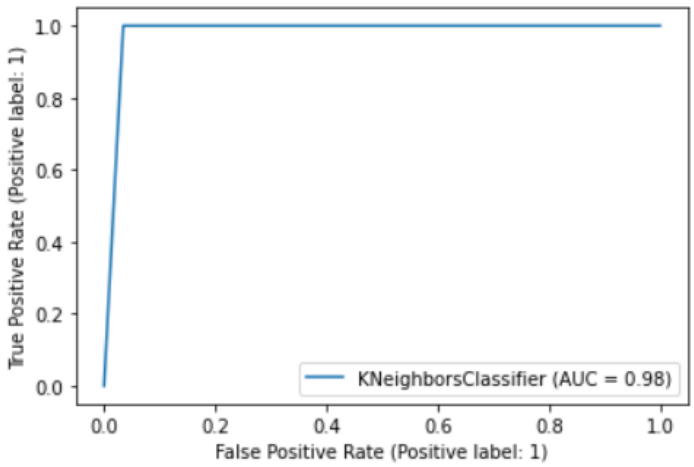
Regresión Logística



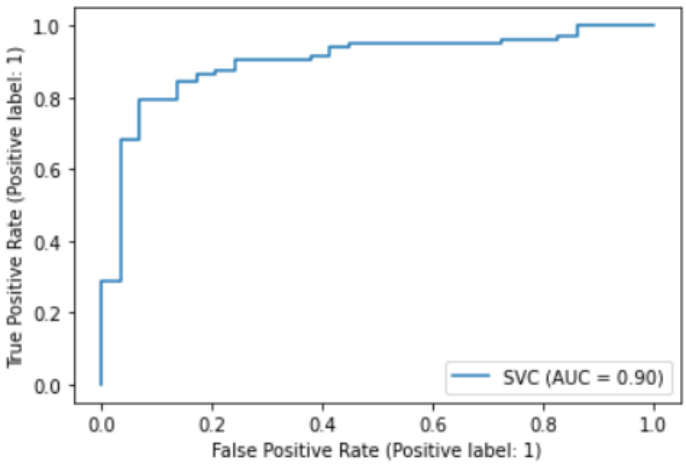
Análisis Lineal Discriminante (LDA)



K Vecinos más Cercanos



Máquinas Vector Soporte



Reportes de clasificación – Modelos Sin Balanceo de Clases

Regresión Logística

```
--Logistic Regression--


[[29  0]
 [ 0 97]]
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	29
1	1.00	1.00	1.00	97
accuracy			1.00	126
macro avg	1.00	1.00	1.00	126
weighted avg	1.00	1.00	1.00	126

Análisis Lineal Discriminante (LDA)

```
--Linear Discriminant Analysis - LDA--


[[25  4]
 [ 2 95]]
```

	precision	recall	f1-score	support
0	0.93	0.86	0.89	29
1	0.96	0.98	0.97	97
accuracy			0.95	126
macro avg	0.94	0.92	0.93	126
weighted avg	0.95	0.95	0.95	126

K Vecinos más Cercanos

```
--K Nearest Neighbors Classifier--


[[28  1]
 [ 0 97]]
```

	precision	recall	f1-score	support
0	1.00	0.97	0.98	29
1	0.99	1.00	0.99	97
accuracy			0.99	126
macro avg	0.99	0.98	0.99	126
weighted avg	0.99	0.99	0.99	126

Máquinas Vector Soporte

```
--SVM Kernel--


[[ 9 20]
 [ 5 92]]
```

	precision	recall	f1-score	support
0	0.64	0.31	0.42	29
1	0.82	0.95	0.88	97
accuracy			0.80	126
macro avg	0.73	0.63	0.65	126
weighted avg	0.78	0.80	0.77	126

- De los resultados obtenidos en los 4 modelos testeados, y presentados antes, podemos ver que el se desempeña para nuestro problema es la **Regresión Logística** con una precisión del 100%, y que además puede interpretarse de la matriz de confusión arrojada del conjunto de validación.
- Por otra parte, de la matriz de confusión proporcionada en el planteamiento del problema, tenemos las siguientes métricas:
  - Exactitud= ***(VP+VN)/(VP+FP+FN+VN)=*** 0.8455= 84.55%
  - Precisión= ***VP/(VP+FP)=*** 0.8484
- Contrastando las métricas del punto anterior vs las que obtuvimos en los diferentes modelos testeados, vemos lo siguiente:

- En la **regresión logística** se observa que la exactitud obtenida fue de 1 (100%) comparada con 0.8455 (84.55%) del médico practicante.

  Basado en lo anterior y tomando como referencia la interpretación de la matriz de confusión de los resultados del médico y el modelo, la recomendación es elegir este modelo como herramienta de ayuda para el diagnóstico de cáncer.
- Para el caso del **análisis lineal discriminante (LDA)**, se obtuvo una exactitud de 0.955 (95.5%) comparada con 0.8455 (84.55%) del médico prácticamente.

  Nuevamente se puede ver, que este modelo también tiene la capacidad de arrojar buenos resultados en el diagnóstico de cáncer y ser de gran utilidad como herramienta para el médico de planta.
- Por otra parte, en el caso del modelo de **K vecinos más cercanos**, tenemos la misma situación: un valor de exactitud de 0.955 (95.5%) comparado con el 0.8455 (84.55%) del médico practicante.

  Al igual que en **LDA**, este modelo también representa una herramienta robusta y confiable que el médico de planta puede elegir para el diagnóstico de cáncer.
- Finalmente, el modelo de **máquinas vector soporte** arroja una exactitud de 0.825 (82.5%) que resulta ser menor a 0.8455 (84.55%) del médico prácticamente.

  Para este caso en particular, la mejor alternativa es que el médico de planta reciba el apoyo del practicante para el diagnóstico de cáncer en los pacientes.

Interpretación de Resultados

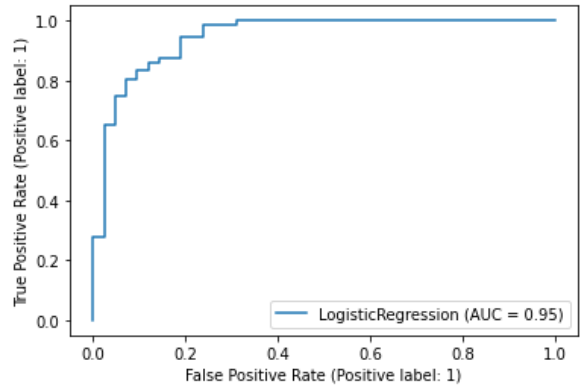
Modelado Clases Balanceadas

```
0.9470899470899471
0.9161290322580645
[[30 12]
 [ 1 71]]
```

--Logistic Regression--

```
[[30 12]
 [ 1 71]]
```

	precision	recall	f1-score	support
0	0.97	0.71	0.82	42
1	0.86	0.99	0.92	72
accuracy			0.89	114
macro avg	0.91	0.85	0.87	114
weighted avg	0.90	0.89	0.88	114

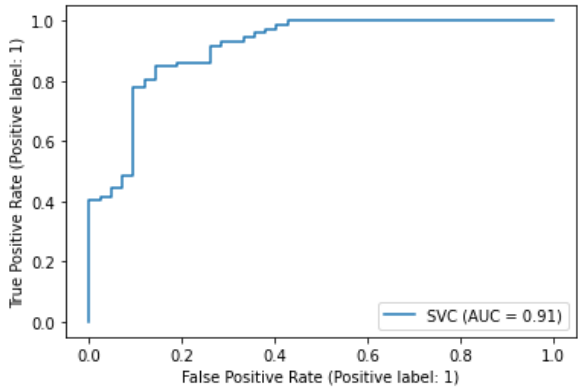



```
0.9107142857142856
0.8774193548387097
[[27 15]
 [ 4 68]]
```

--Máquinas de Vector Soporte--

```
[[27 15]
 [ 4 68]]
```

	precision	recall	f1-score	support
0	0.87	0.64	0.74	42
1	0.82	0.94	0.88	72
accuracy			0.83	114
macro avg	0.85	0.79	0.81	114
weighted avg	0.84	0.83	0.83	114

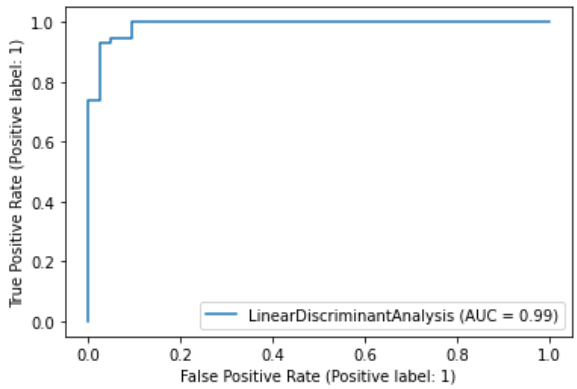



```
0.9894179894179894
0.9600000000000001
[[36  6]
 [ 0 72]]
```

--Linear Discriminant Analysis LDA--

```
[[36  6]
 [ 0 72]]
```

	precision	recall	f1-score	support
0	1.00	0.86	0.92	42
1	0.92	1.00	0.96	72
accuracy			0.95	114
macro avg	0.96	0.93	0.94	114
weighted avg	0.95	0.95	0.95	114

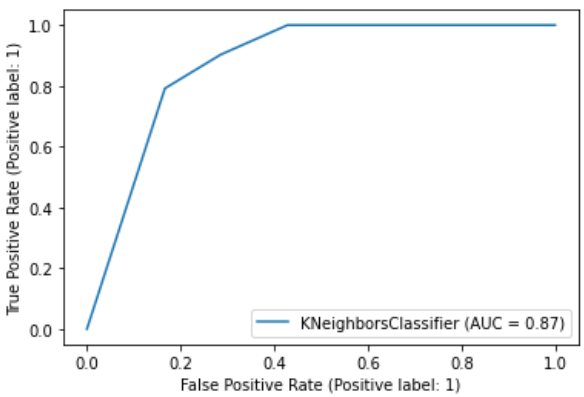



```
0.8741732804232804
0.87248322147651
[[30 12]
 [ 7 65]]
```

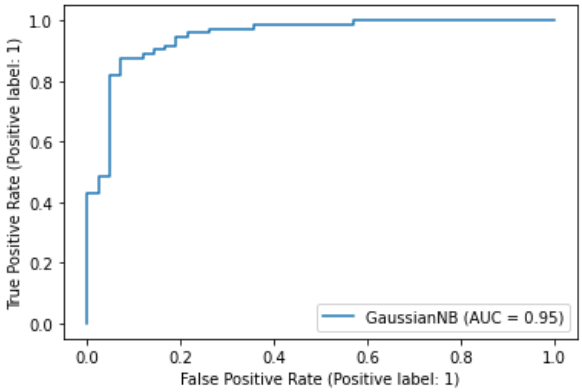
--K Vecinos más Cernanos--

```
[[30 12]
 [ 7 65]]
```

	precision	recall	f1-score	support
0	0.81	0.71	0.76	42
1	0.84	0.90	0.87	72
accuracy			0.83	114
macro avg	0.83	0.81	0.82	114
weighted avg	0.83	0.83	0.83	114



```
0.9480820105820106
0.9041095890410958
[[34  8]
 [ 6 66]]


--Naive Bayes Classifier--


[[34  8]
 [ 6 66]]

              precision    recall  f1-score   support

           0       0.85      0.81      0.83        42
           1       0.89      0.92      0.90        72

    accuracy                           0.88       114
   macro avg       0.87      0.86      0.87       114
weighted avg       0.88      0.88      0.88       114
```



Después de testear los 5 modelos, podemos darnos cuenta de algunos puntos muy importantes:

- En primera instancia notamos en el EDA que la target esta desbalanceada, porque tenemos una menor cantidad de diagnósticos de cáncer.
- Es necesario balancear las clases y testear nuevamente los modelos.
- Los resultados obtenidos en los modelos antes de balancear la target son perfectos/casi perfectos en algunos.
- De los resultados obtenidos antes de balancear target, se observa que el modelo está aprendiendo más de la "Clase 0" y como consecuencia será difícil que generalice.

Conclusiones:

Se analizaron 5 modelos candidatos a ser utilizados como herramienta de ayuda al médico de planta, para el diagnóstico de cáncer contrastado con los resultados arrojados por la matriz de confusión del médico practicante.

- **Resultados**:

Médico Practicante: Acc: 0.85

Regresión Logística: Acc: 0.89 y AUC_ROC_Score: 0.95
Análisis Discriminante Lineal (LDA): Acc: 0.95 y AUC_ROC_Score: 0.98
Máquinas de Vector Soporte (SVM): Acc: 0.83 y AUC_ROC_Score: 0.91
K Vecinos más Cercanos (KNN): Acc: 0.83 y AUC_ROC_Score: 0.87
Naive Bayes Classifier (NBC): Acc: 0.88 y AUC_ROC_Score: 0.95

A partir de lo anterior, podemos elegir los siguientes modelos como una buena herramienta de ayuda al médico de planta para el diagnóstico de cáncer:

1.- **Análisis Discriminante Lineal (LDA)**: Acc: 0.95 y AUC_ROC_Score: 0.98
2.- **Regresión Logística**: Acc: 0.89 y AUC_ROC_Score: 0.95
3.- **Naive Bayes Classifier (NBC)**: Acc: 0.88 y AUC_ROC_Score: 0.95 (*******)

De donde los modelos **1** y **2** resultan ser la mejor alternativa, pues si bien el Naive Bayes Classifier también tuvo un muy buen desempeño comparado a los resultados del médico practicante. El hecho de asumir independencia entre las variables predictoras resulta ser un problema ya que en la práctica es casi imposible que obtengamos un conjunto de predictores que sean completamente independientes. Además, dentro del contexto clínico de este problema es muy importante que el diagnóstico sea el correcto para cada paciente.