Trabajo Práctico Final

Modelo predictivo para la detección de cáncer de mama

Moreira Pablo

UNAB

Inferencia estadística y reconocimiento de patrones

Tabla de contenido

Capítulo 1

Introducción:

La información a analizar

En el presente informe se analizan los datos correspondientes a una base que recopila información obtenida del Dr. Wolberg sobre sus casos clínicos, la misma cuenta con 683 muestras donde vemos características en distintas variables de los análisis realizados a pacientes con tumores mamarios que pueden ser benignos o malignos.

La base de datos se descarga del siguiente link: https://www.kaggle.com/adhyanmaji31/breast-cancer-prediction?select=Breast+Cancer+Prediction.csv

Las variables que forman parte de los estudios que se realizan a este tipo de pacientes a tener en cuenta son:

- Número de código de muestra: número de identificación
- Espesor del grupo: evaluada en el rango de 1 a 10
- Uniformidad del tamaño de la célula: evaluada en el rango de 1 a 10
- Uniformidad de la forma de la célula: evaluada en el rango de 1 a 10
- Adhesión marginal: evaluada en el rango de 1 a 10
- Tamaño de célula epitelial única: evaluada en el rango de 1 a 10
- Núcleos desnudos: evaluada en el rango de 1 a 10
- Cromatina suave: evaluada en el rango de 1 a 10
- Nucleolos normales: evaluada en el rango de 1 a 10
- Mitosis: evaluada en el rango de 1 a 10
- Clase: evaluada en 0 o 1

Se busca un modelo predictivo que pueda clasificar de forma correcta si un tumor mamario es benigno o maligno utilizando como base los estudios realizados en las pacientes.

Capítulo 2

Análisis Exploratorio

En primera instancia vamos a descartar para el análisis el código de muestra porque no es relevante para el entrenamiento y testeo del modelo predictivo.

La clase es lo que determina si el tumor analizado es maligno o benigno identificando a los benignos con 0 y los malignos con 1.

La proporción entre la cantidad de datos de cada grupo es la que se ve en la

Tomamos como parámetro que si el resultado es benigno vamos a considerarlo como **negativo** y si es maligno como **positivo**.

El resto de las variables relacionadas a los distintos ítems pueden tomar valores entre 1 y 10

Conclusiones del análisis exploratorio

Al aplicar componentes principales notamos que las dos variables que más intervienen en la primera componente son:

- Uniformidad de la forma de la célula: Cuando esta variable aumenta su valor dentro del rango, aumenta la influencia negativa la primera componente
- Espesor del grupo: Cuando esta variable aumenta su valor dentro del rango, aumenta la influencia positiva en la primera componente

Mientras que en la segunda componente son:

- Mitosis: Cuando esta variable aumenta su valor dentro del rango, aumenta la influencia positiva en la segunda componente
- Nucleolos normales: Cuando esta variable aumenta su valor dentro del rango, aumenta la influencia negativa la segunda componente

Las componentes principales agrupan el 74% de la información (varianza) de los datos y de ellas se obtiene una marcada división (pero no definitiva) entre los casos negativos y positivos

Capítulo 3

Clasificación Supervisada

Con los datos de la base vamos a aplicar distintos modelos de entrenamiento para obtener clasificaciones que se ajusten con la mayor precisión posible a los datos utilizados, esto se realiza dividiendo los datos en dos subconjuntos, uno mayor que se utiliza para entrenar al modelo y uno menor que se utiliza como testeo.

Intentamos así minimizar el error de clasificación que puede verse reflejado en dos casos:

- Que clasifique como positivo y sea negativo (el caso de un falso positivo, marcaria como tumor maligno cuando en realidad es benigno)
- Que clasifique como negativo y sea positivo (en este caso es un falso negativo, da como resultado que el tumor es benigno cuando en realidad es maligno)

Las dos opciones se consideran como fallas y los detalles de las distintas pruebas realizadas se informan en el apéndice técnico.

Detallar las variables que más influyen al momento de la clasificación (que tira más para maligno y más para benigno)

Capítulo 4

Conclusiones de la Clasificación Supervisada

Hemos podido comprobar experimentalmente que el clasificador que obtiene mejores resultados, según nuestras métricas, es el de Bayes ingenuo utilizando el punto de corte por defecto.

De esta forma logramos minimizar tanto los falsos negativos como los falsos positivos a su mínima expresión, garantizando la correcta clasificación en un 97% para los casos positivos y 95% para los negativos.

A pesar de esto cabe señalar que todos los clasificadores utilizados dan resultados similares

En el apéndice técnico se detallan los métodos utilizados y sus métricas

Apéndices técnicos

Influencia de las distintas variables en Componentes principales:

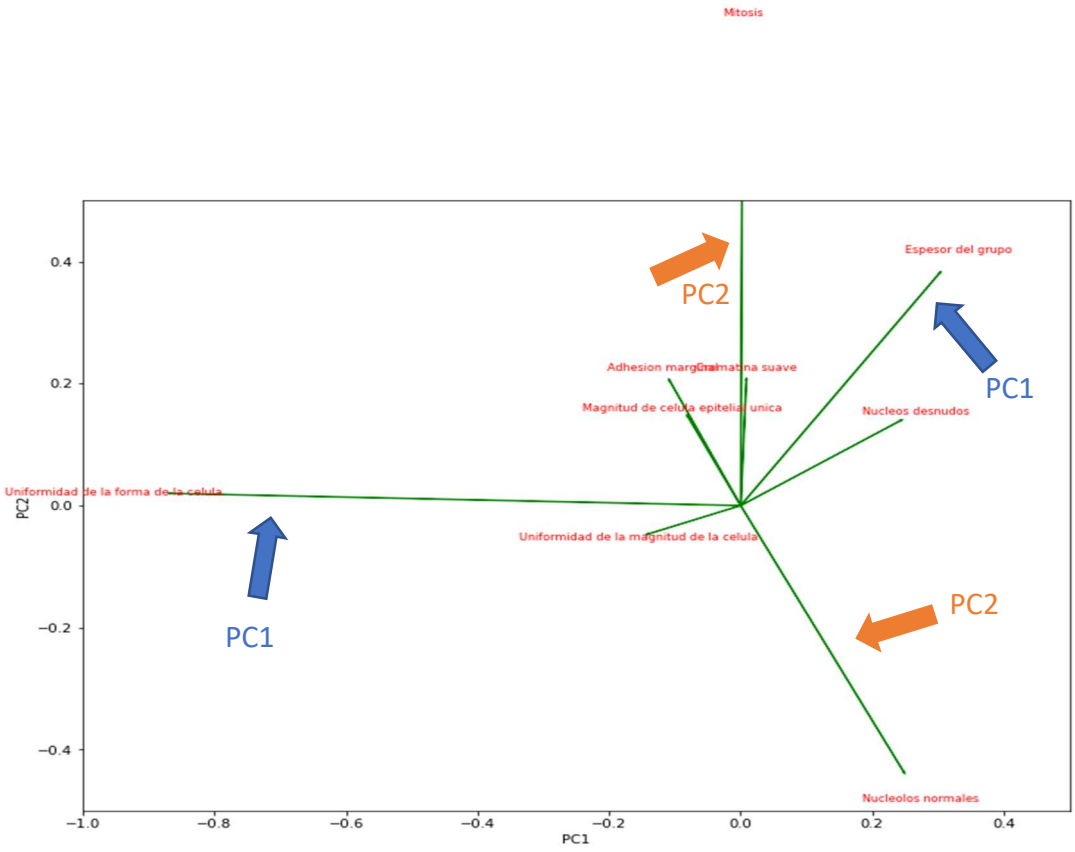
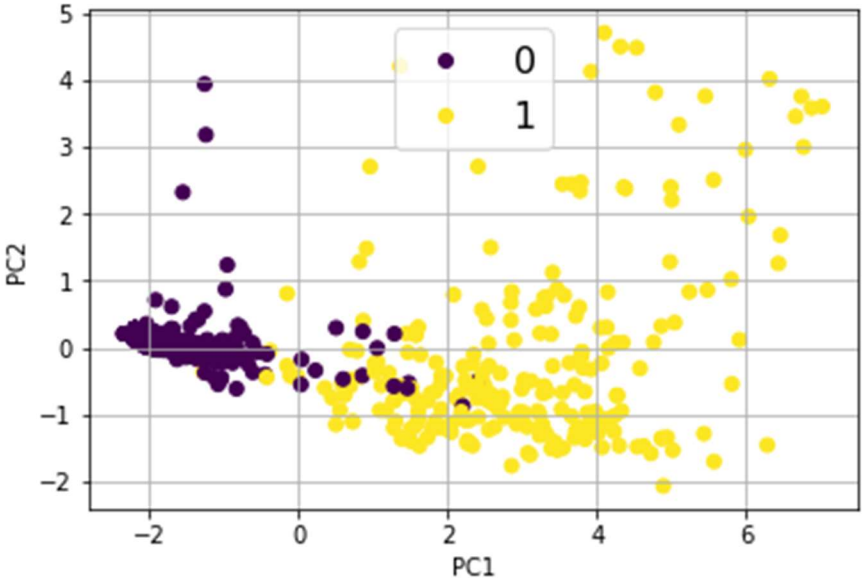
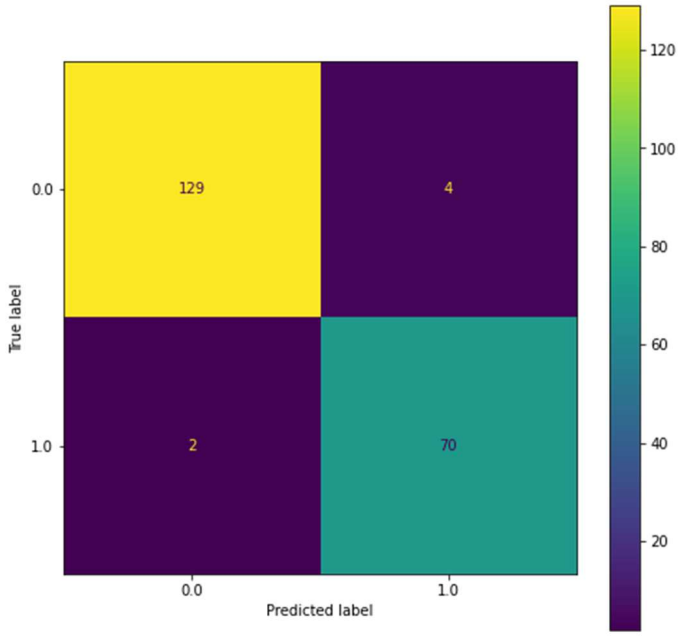


Gráfico de componentes principales



Podemos apreciar que los casos negativos se agrupan en su mayoría hacia el margen inferior izquierdo, mientras que los positivos se desplazan hacia la derecha. Hay que considerar también que la cantidad de negativos es mucho mayor, es por eso que están mucho más concentrados

Datos y métricas del clasificador elegido (Bayes ingenuo con punto de corte por defecto)



Este modelo predictivo marca dentro del dataset solo dos falsos negativos y cuatro falsos positivos

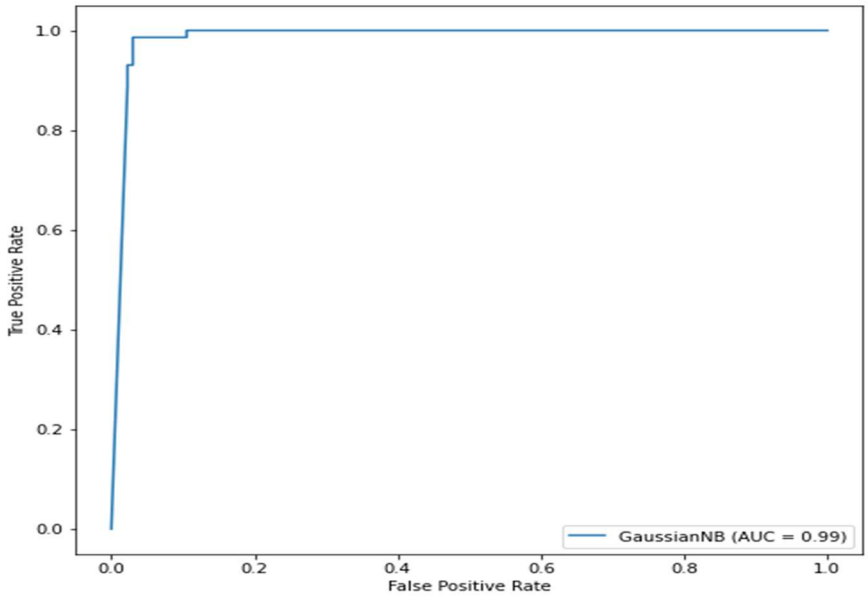
```
Métricas sobre datos de TEST
Accuracy:  0.97
Recall:  0.97
Precision:  0.95
F1:  0.96
Classification Report:
              precision    recall  f1-score   support

         0.0       0.98      0.97      0.98       133
         1.0       0.95      0.97      0.96        72

    accuracy                           0.97       205
   macro avg       0.97      0.97      0.97       205
weighted avg       0.97      0.97      0.97       205
```

Nos centramos en el resultado del recall ya que es el que mide la precisión en el calculo de los falsos negativos que, como vimos anteriormente, este error haría que un caso de tumor maligno fuera diagnosticado como benigno.

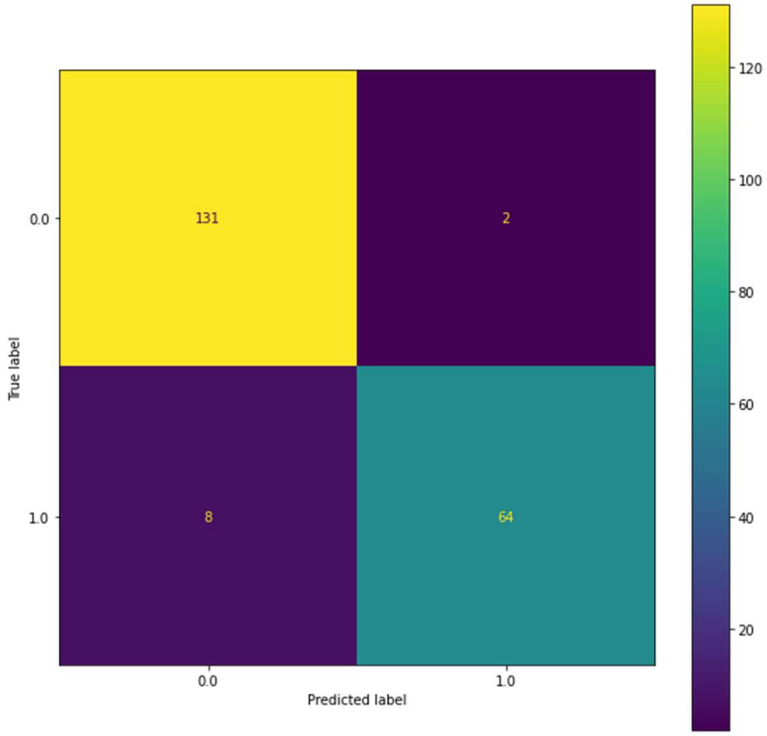
Evaluación en TEST por Curva ROC



Para llegar a esta elección se evaluaron otros métodos de clasificación que, si bien obtuvieron resultados aceptables, no llegaron a los valores anteriormente expuestos.

Al variar el punto de corte a 0.7 se produce una disminución del recall en 0.01%. Esto significa que realiza la clasificación prácticamente igual salvo que comete un error más en la clasificación de positivos obteniendo un falso negativo adicional.

Clasificador Discriminante lineal con punto de corte por defecto



Los falsos negativos de este modelo aumentan mucho con respecto al anterior, sus métricas son:

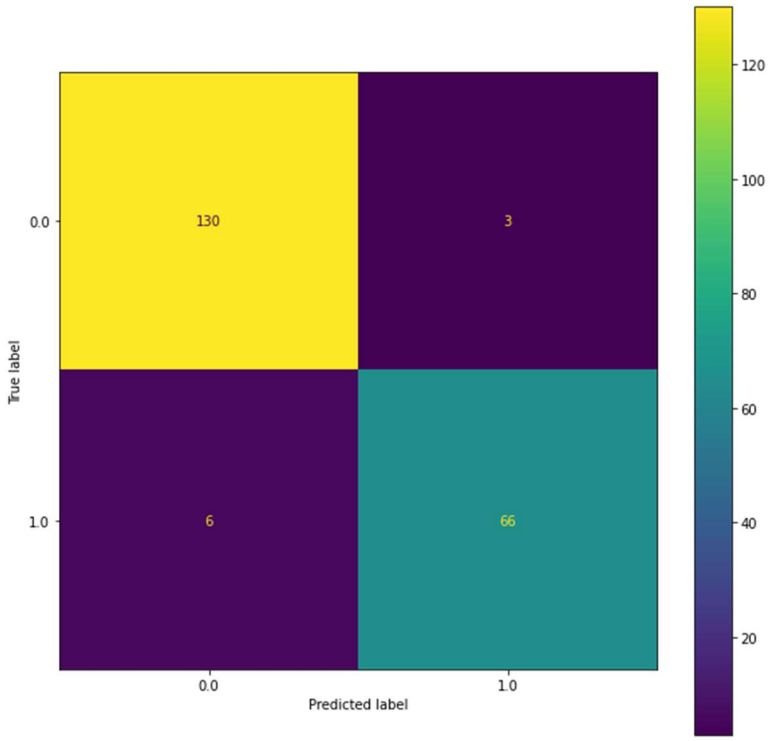
```
Métricas sobre datos de TEST sin punto de corte arbitrario
Accuracy:  0.95
Recall:  0.89
Precision:  0.97
F1:  0.93
Classification Report:
              precision    recall  f1-score   support

         0.0       0.94      0.98      0.96       133
         1.0       0.97      0.89      0.93        72

    accuracy                           0.95       205
   macro avg       0.96      0.94      0.95       205
weighted avg       0.95      0.95      0.95       205
```

Notamos la considerable baja en el recall que indica el error al clasificar los falsos negativos.

Clasificador Regresión Logística con punto de corte por defecto



En este caso vemos un aumento considerable de los falsos negativos (se triplican con respecto al modelo de Bayes) y una leve baja de los falsos positivos. En las métricas vemos una reducción del recall debido a esto.

```
Métricas sobre datos de TEST sin punto de corte arbitrario
```

```
Accuracy:  0.96
Recall:  0.92
Precision:  0.96
F1:  0.94
Classification Report:
              precision    recall  f1-score   support

         0.0       0.96      0.98      0.97       133
         1.0       0.96      0.92      0.94        72

    accuracy                           0.96       205
   macro avg       0.96      0.95      0.95       205
weighted avg       0.96      0.96      0.96       205
```

Clasificador Regresión Logística con punto de corte 0.7

En este caso detallamos únicamente las métricas para ver que la disminución en el recall es mayor al subir el punto de corte:

```
Métricas sobre datos nuevos de TEST, punto de corte arbitrario 0.7
Accuracy:  0.95
Recall:  0.88
Precision:  0.97
F1:  0.92
Classification Report:
              precision    recall  f1-score   support

         0.0       0.94      0.98      0.96       133
         1.0       0.97      0.88      0.92        72

    accuracy                           0.95       205
   macro avg       0.95      0.93      0.94       205
weighted avg       0.95      0.95      0.95       205
```

Si subimos el punto de corte a 0.8 el recall sigue cayendo en ese caso llega a 0.86

Como conclusión adicional, en todos los modelos al aumentar el punto de corte afectamos la clasificación de los falsos negativos, que siempre aumenta.

Futuros trabajos

Para ampliar los resultados sería conveniente que la base de datos pudiera ser ampliada con análisis de distintos profesionales y en un rango de tiempo más amplio. De esta forma y con una base más grande se puede lograr optimizar más aún la clasificación obtenida.