**GUIÓN APC 1B:**

**Diapositiva 1: Portada**

Buenos días a todos, somos [nombres], nuestro grupo es GA608 y vamos a hacer una presentación explicando cómo hemos aplicado modelos de clasificación para detectar sitios web fraudulentos.

**Diapositiva 2:**

Los pasos que hemos seguido para realizar la práctica han sido:

-Analizar los datos que venían en la base de datos, para poder detectar qué contiene, qué tipos de datos son y sobre qué variable aplicar la clasificación.

-Utilizar distintos clasificadores y técnicas de selección.

-Analizar el rendimiento de los distintos clasificadores que usamos.

-Como hemos resuelto el problema.

-Y finalmente las conclusiones que hemos sacado con la realización de esta práctica.

**Diapositiva 3:**

En esta tabla están todas las variables que contenía nuestra base de datos. Cada una de estas variables podría tomar el valor de -1, 0 o 1.

-1 indicaba que era una característica fraudulenta

0 indicaba que era una característica sospechosa

y 1 para indicar que no es una web fraudulenta.

Nuestra variable objetivo será la que tenemos al final llamada Result, que nos indica si una web con todo el conjunto de variables anteriores es fraudulenta o no.

**Diapositiva 4:**

Para los diferentes clasificadores que hemos utilizado, hemos utilizado distintas técnicas de selección para realizar el entrenamiento.

La Básica, que consiste en separar un porcentaje de datos para entrenamiento y otro porcentaje de los datos para validación (en nuestro caso hemos hecho un 80, 70 y 50 % de entrenamiento).

El K-fold, que consiste en partir los datos en K elementos, y realizar una iteración para obtener un error, y utilizar como error total la media de todos los errores obtenidos en las K iteraciones.

Finalmente tenemos el Leave One Out, que consiste en utilizar una única muestra de test y realizar una iteración con cada una de las muestras.

**Diapositiva 5:**

Esta es la tabla con los resultados en precisión que hemos obtenido utilizando un clasificador de regresión logística con las siguientes características.

Como podemos observar, el método de selección que nos está dando un resultado de precisión más alto es el de K-Fold, donde tenemos prácticamente la misma precisión utilizando desde K=3 a K=6.

**Diapositiva 6:**

En esta dispositiva tenemos los resultados obtenidos con un clasificador SVM , utilizando distintos métodos de selección y utilizando 3 kernels distintos, el Linear, Poly y rbf.

El kernel que nos da un mejor resultado de precisión es el Linear, con un 86% de precisión y utilizando el método de selección K-Fold.

**Diapositiva 7:**

Para realizar el código, hemos utilizado jupyter notebook, porque creemos que es cómodo trabajar con celdas.

**Diapositiva 8:**

Y las librerías que hemos utilizado son las básicas, numpy para tratar los matrices, sklearn para poder utilizar los clasificadores y también nos ha ayudado a realizar la selección con K-fold ,pandas para la lectura de datos, y librerías para representar gráficamente.

**Diapositiva 9:**

Para hacer un buen análisi de el rendimiento del clasificador, utilizamos otro tipo de medidas como el F1 score, para ver qué diferencia había de los datos que obtenemos utilizando como medida de rendimiento la precisión a los que obtenemos utilizando el F1 score. Para cada uno de los métodos de selección utilizamos los distintos average que permite la función f1 score de sklearn. En este caso, el mejor resultado lo obtenemos utilizando con clasificador de regresión logística utilizando un average macro.

**Diapositiva 10:**

Hemos realizado distintas gráficas de precision-recall y de curva ROC. En esta mostramos las gráficas obtenidas con el clasificador que nos daba un mejor resultado teniendo en cuenta el F1 score.

**Diapositiva 11:**

Y en esta gráfica el clasificador que tenía un mejor resultado teniendo en cuenta la precisión como valor de rendimiento.

**Diapositiva 12:**

Estas gráficas son una curva ROC para cada una de las clases utilizando un SVM linear. Como observamos, la clase 0 que es la segunda, es la que tiene un peor rendimiento ya que los true positive no son tan altos como en las otras dos clases, mientras que la clase 1 y -1 tienen unos valores de true positive bastante altos.

**Diapositiva 13:**

Como pruebas de visualización utilizamos el SVM con distintas configuraciones. Una vez llegados a este punto, nos planteamos que quizá planteando equivocadamente los valores a clasificar, ya que por la distribución de las muestras creíamos que una tercera clase que era la de sospechoso no se estaba visualizando correctamente. Pedimos ayuda en el foro y nos comentaron que lo mejor era utilizar en la variable objetivo los valores de 1 para identificar si es legítima o -1 para identificar si era fraudulenta, y no tener en cuenta los valores 0. Al realizar esto, la clasificación quedaba en 2 colores, y obteniamos las gráficas mostradas en la diapositiva. Realizamos unas cuantas gráficas más con distintos pares de características.

**Diapositiva 14:**

Como dificultades, sobretodo fue el hecho de plantear la clasificación de los datos, por el hecho de que nuestra base de datos contenía tres posibles valores, no teníamos claro si debíamos hacer una clasificación binaria o con más clases.

Hemos conseguido un modelo que tiene un F1 score aceptable y una clasificación con un buen rendimiento.

**Diapositiva 15:**

Y en cuanto a la actividad, el trabajo en equipo nos ha facilitado la faena. En esta práctica creemos que hemos tenido un objetivo claro y bastante soporte para realizar la práctica. El aspecto positivo ha sido aprender a aplicar diferentes técnicas para realizar un modelo de clasificación. Y quizá si tenemos que comentar algo, hemos tenido un poco de dudas para saber que medida de rendimiento escoger para determinar si un modelo es mejor que otro.

**Diapositiva 16:**

Gracias por vuestra atención