Análisis sobre el desempeño de un modelo de aprendizaje máquina

José María Ibarra Pérez, a01706970
Tecnológico de Monterrey
11 de septiembre de 2023

TC3006C.101 Inteligencia artificial avanzada para la ciencia de datos I

I. Introducción

La construcción de modelos de aprendizaje máquina se debe llevar a cabo bajo un mismo contexto de datos, es decir, los datos que entrenan y eventualmente evalúan al modelo mantienen el mismo formato; es necesario que para que se obtengan predicciones confiables se imputen al modelo datos conteniendo el mismo tipo de registros con los que este fue entrenado. Además, estos datos deben de ser adecuados para el tipo de modelo que se requiera, y deberán recibir diferentes tratamientos según si se busca, por ejemplo, una regresión o clasificación, o si se trata de un modelo supervisado a diferencia de uno no-supervisado. Para fines de este trabajo, se realizaron dos versiones de un algoritmo de clasificación binaria: una **máquina de soporte vectorial (SVM)**. La primera de estas dos versiones es la implementación del algoritmo desde cero, esto es, programación de todas las reglas y funciones dentro del algoritmo. Segundo, la implementación del algoritmo a través del marco de trabajo o librería de Python S*cikit-learn*. Se pretende comparar en este reporte las evaluaciones de dichos algoritmos, de forma que se identifiquen las diferencias de capacidades en cuestiones de generalización y generación de predicciones, esperando que la librería muestre mejores resultados. Además, en el caso de la implementación con la librería, se presentan técnicas que permiten ver la mejora del modelo ante su susceptibilidad hacia *overfitting* y *underfitting.*

II. Datos

Para este trabajo, se entrenaron una serie de modelos con dos diferentes conjuntos de datos. Primero, el conjunto de datos de *Iris Dataset*, el cual agrupa una serie de registros de mediciones de las principales características de flores iris, cada una perteneciente a uno de tres tipos de iris. Para la adaptación de los datos, fue necesario seleccionar una partición de este conjunto, dado que la SVM implementada es capaz de clasificar entre dos clases solamente. Se seleccionaron entonces 2 de las clases disponibles, como se muestra en la Figura 1. Segundo, se entrenó con el conjunto de datos de *Breast Cancer Dataset*, el cual agrupa una colección de registros de pacientes y si estos tenían cáncer o no. De esta forma, ambos los conjuntos de datos son apropiados para implementarse en el algoritmo de clasificación binaria.

```
if dataset_num == 1: #Seleccionar dos clases y dos atributos
    X = X[y != 2, :]
    y = y[y != 2]

y[y == 0] = -1  # Ajuste de clases (-1, 1)
```

Figura 1 Se selecciona la partición de datos perteneciente solamente a las primeras dos clases del conjunto de datos de Iris.

III. Entrenamiento y evaluación

Para el entrenamiento del modelo sin el uso de una librería se implementó el algoritmo por cada uno de sus partes. Este consiste en una serie de funciones que se encargan de preparar los datos, y calcular los diferentes componentes del algoritmo: funciones de costo y pérdida, función lineal para delimitar la diferencia de clases y el gradiente con el que se itera durante el entrenamiento. Por otro lado, el entrenamiento del modelo implementado con la librería es más sencillo a comparación, pues se limita al tratamiento de datos y *fit* con la librería. Ambos entrenamientos se hacen luego de una separación de datos en un conjunto de prueba y otro de entrenamiento, el cual para el modelo con librería, es dividido nuevamente en conjuntos de entrenamiento y validación con *K-fold cross validation.*

Se busca hacer evidente la generalización de los modelos. En el caso del modelo sin librería, se entrenan 3 modelos diferentes con cada uno de los conjuntos de datos mencionados (Iris, Cáncer), cada uno de estos modelos con porciones de entrenamiento diferentes, de manera que se pueda ver si el algoritmo implementado es capaz de generalizar con diferentes datos de entrenamiento. La figuras siguientes muestran los resultados de las seis corridas de entrenamiento, en pérdida contra época y métricas de precisión. La corrida para cuándo la pérdida aumenta considerablemente entre épocas.

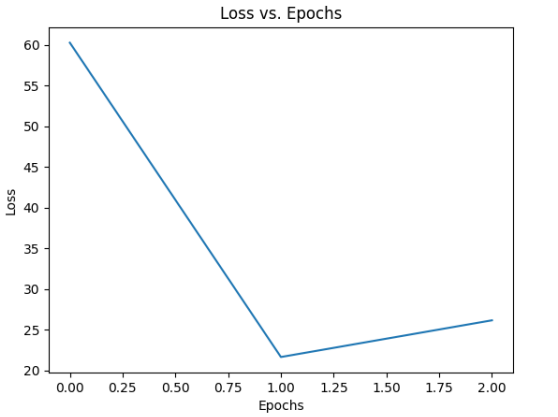


Figura 2.1 Pérdida contra épocas, Iris corrida 1

Métricas:

	precision	recall	f1-score	support
-1	0.91	1.00	0.95	10
1	1.00	0.95	0.97	20
accuracy			0.97	30
macro avg	0.95	0.97	0.96	30
weighted avg	0.97	0.97	0.97	30

Figura 2.2 Métricas de evaluación, Iris corrida 1

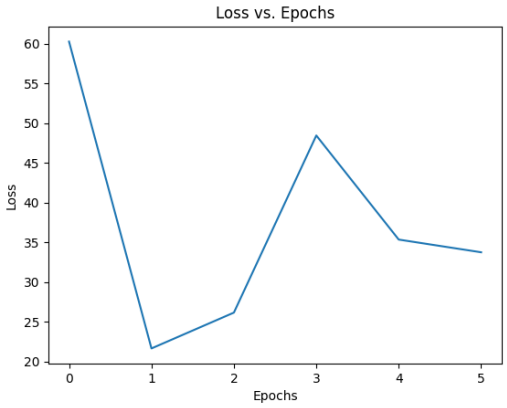


Figura 2.3 Pérdida contra épocas, Iris corrida 2

Métricas:

	precision	recall	f1-score	support
-1	0.95	1.00	0.97	18
1	1.00	0.92	0.96	12
accuracy			0.97	30
macro avg	0.97	0.96	0.96	30
weighted avg	0.97	0.97	0.97	30

Figura 2.4 Métricas de evaluación, Iris corrida 2

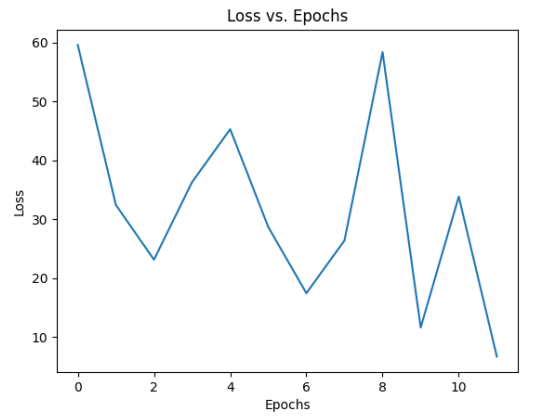


Figura 2.5 Pérdida contra épocas, Iris corrida 3

Métricas:

	precision	recall	f1-score	support
-1	1.00	0.44	0.62	18
1	0.55	1.00	0.71	12
accuracy			0.67	30
macro avg	0.77	0.72	0.66	30
weighted avg	0.82	0.67	0.65	30

Figura 2.6 Métricas de evaluación, Iris corrida 3

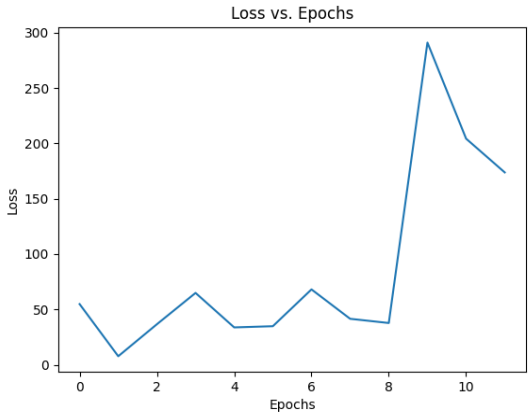


Figura 2.7 Pérdida contra épocas, Cáncer corrida 1

Métricas:

	precision	recall	f1-score	support
-1	0.91	0.71	0.80	59
1	0.86	0.96	0.91	112
accuracy			0.88	171
macro avg	0.89	0.84	0.86	171
weighted avg	0.88	0.88	0.87	171

Figura 2.8 Métricas de evaluación, Cáncer corrida 1

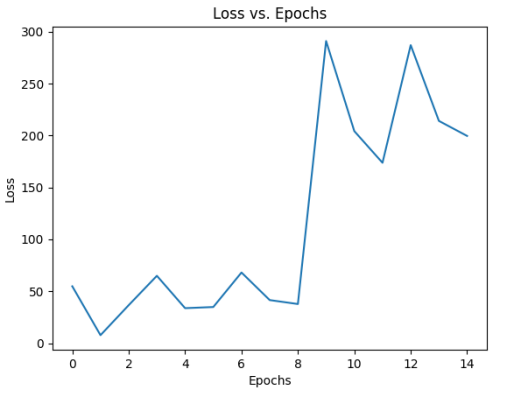


Figura 2.9 Pérdida contra épocas, Cáncer corrida 2

Métricas:

	precision	recall	f1-score	support
-1	1.00	0.56	0.72	70
1	0.77	1.00	0.87	101
accuracy			0.82	171
macro avg	0.88	0.78	0.79	171
weighted avg	0.86	0.82	0.80	171

Figura 2.10 Métricas de evaluación, Cáncer corrida 2

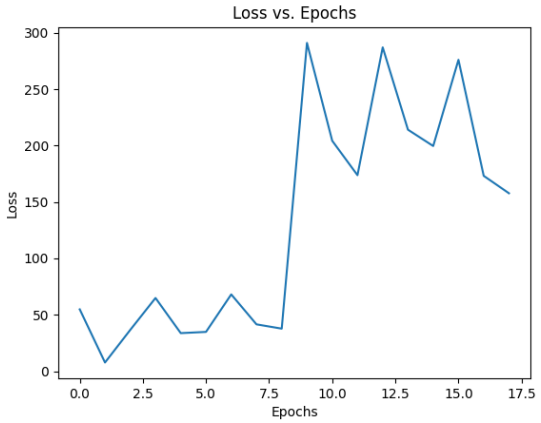


Figura 2.11 Pérdida contra épocas, Cáncer corrida 3

Métricas:

	precision	recall	f1-score	support
-1	1.00	0.48	0.65	58
1	0.79	1.00	0.88	113
accuracy			0.82	171
macro avg	0.90	0.74	0.77	171
weighted avg	0.86	0.82	0.80	171

Figura 2.12 Métricas de evaluación, Cáncer corrida 3

Se puede observar en las figuras el rendimiento del algoritmo de entrenamiento sin la implementación de la librería. El objetivo de correr varias veces el algoritmo con cada uno de los conjuntos de datos es, como se mencionó, observar la generalización del mismo. Se puede observar con las métricas que efectivamente el algoritmo construido es capaz de generar predicciones sin importar la sección de los datos, además que se observa que es peor clasificando en el segundo conjunto de datos, esto probablemente debido a que es un conjunto considerablemente más complejo que el primero.

En el caso del entrenamiento del algoritmo implementado con la librería, solamente se corre un modelo con cada uno de los conjuntos de datos, pues se asume que el modelo de la librería es capaz de generalizar correctamente sin importar el conjunto de datos. Las siguientes figuras muestran las matrices de confusión y evaluación general y de *cross validation* de cada uno de los conjuntos de datos.

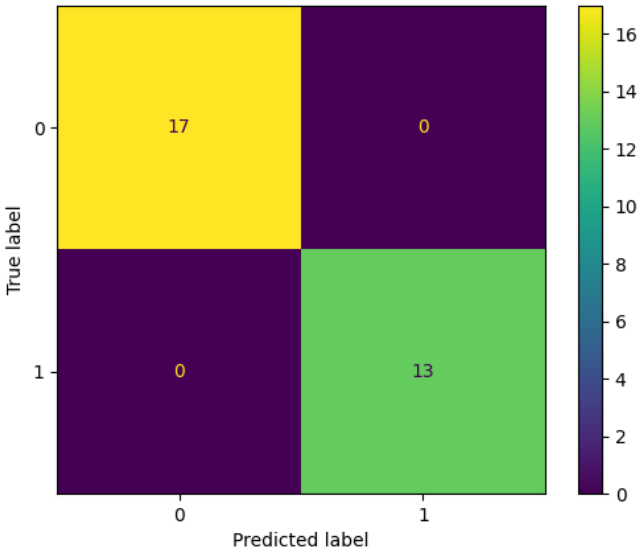


Figura 3.1 Matriz de confusión, Iris

```
10-fold cross-validation score: [1. 1. 1. 1. 1. 1. 1. 1.
 1. 1.]
Mean cv score: 1.0
              precision    recall  f1-score   support

          -1       1.00      1.00      1.00        17
           1       1.00      1.00      1.00        13

    accuracy                           1.00        30
   macro avg       1.00      1.00      1.00        30
weighted avg       1.00      1.00      1.00        30
```

Figura 3.2 Evaluación general y CV, Iris

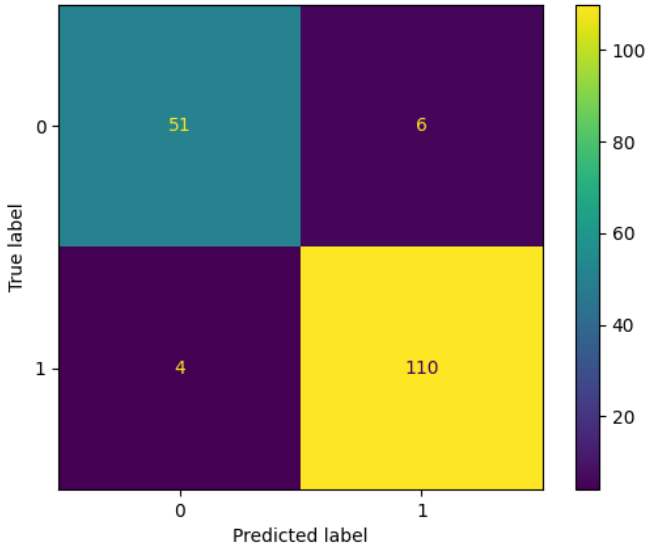


Figura 3.3 Matriz de confusión, Cáncer

```
10-fold cross-validation score: [0.925      0.95       0
.95       0.975      0.925      0.95
 1.         0.975      0.8974359  0.97435897]
Mean cv score: 0.9521794871794873
              precision    recall  f1-score   support

          -1       0.93      0.89      0.91        57
           1       0.95      0.96      0.96       114

    accuracy                           0.94       171
   macro avg       0.94      0.93      0.93       171
weighted avg       0.94      0.94      0.94       171
```

Figura 3.4 Evaluación general y CV, Cáncer

Podemos ver la comparación entre los dos distintos conjuntos de datos, donde de nuevo podemos apreciar que el modelo es mejor clasificando los datos del primero conjunto de datos. En el segundo conjunto de datos, sin embargo, podemos ver que aunque el modelo tiene métricas muy buenas, no es capaz de clasificar perfectamente entre las clases. Las precisiones a lo largo de los *folds* de CV muestran evidencia de que el modelo es, efectivamente, capaz de generalizar con diferentes porciones de los datos.

IV. Análisis de sesgo y varianza

Aun cuando existe evidencia de la generalización, es prudente generar un análisis de sesgo y varianza, donde la primera refiere a la capacidad generalizar la información provista y generar previsiones que capturen la naturaleza correcta de los datos. La segunda refiere a la susceptibilidad de los datos de aprender únicamente la los datos provistos, cuando no es capaz de aprender la naturaleza verdadera de los datos y aprende solamente los patrones de los datos de entrenamiento con alta sensibilidad.

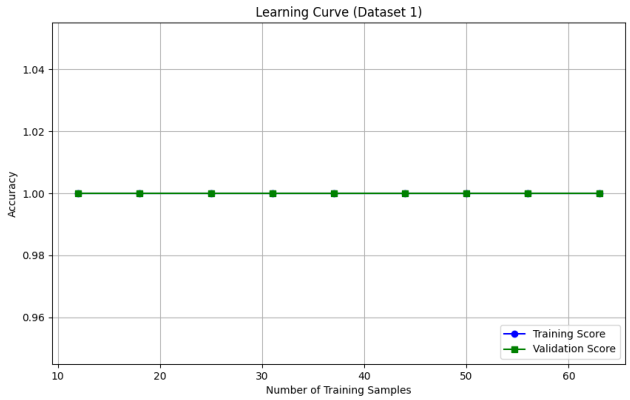


Figura 4.1 Precisión contra muestras de entrenamiento, Iris

La Figura 4.1 muestra el comportamiento de las precisiones de entrenamiento y validación a lo largo que se aumentan las muestras de entrenamiento del modelo. Para los datos de Iris se muestra que el modelo mantiene su precisión perfecta, y aunque esto podría atribuirse a un caso de varianza alta (*overfitting)*, es probable que se deba a la poca complejidad de los datos. Podemos decir que hay evidencia de alta varianza y bajo sesgo.

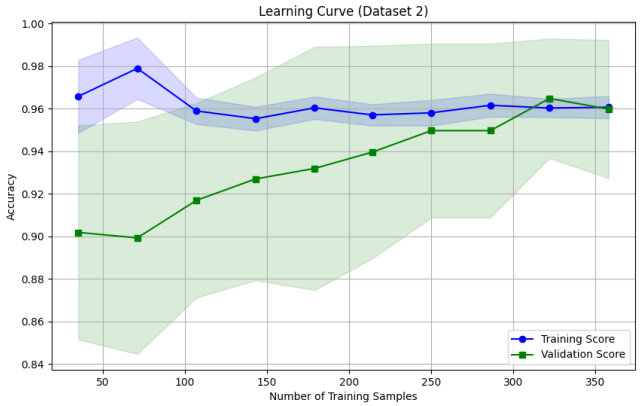
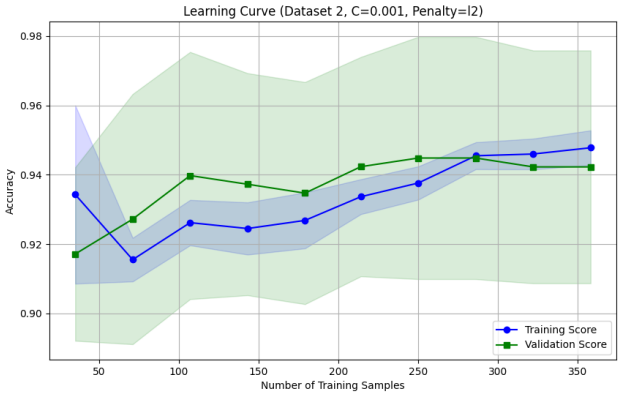


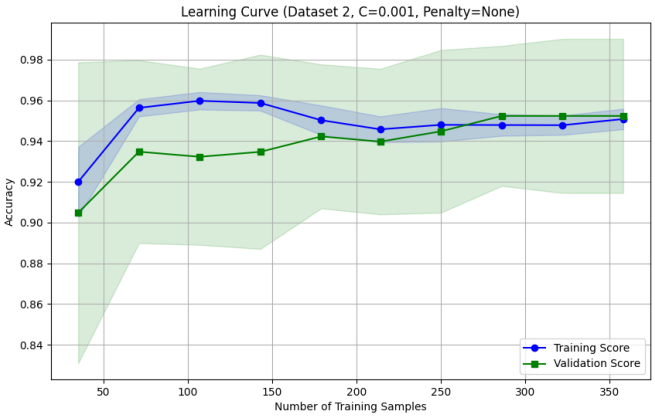
Figura 4.2 Precisión contra muestras de entrenamiento, Cáncer

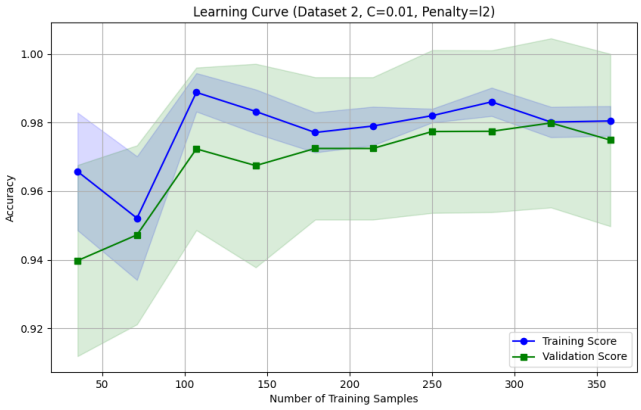
La Figura 4.2 muestra la misma evolución a lo largo del entrenamiento, donde podemos observar que mientras se aumentan las muestras de entrenamiento, las precisiones de entrenamiento y validación se acercan, lo que es indicio de buen rendimiento del modelo. En este caso podemos decir que hay evidencia de sesgo bajo y varianza baja.

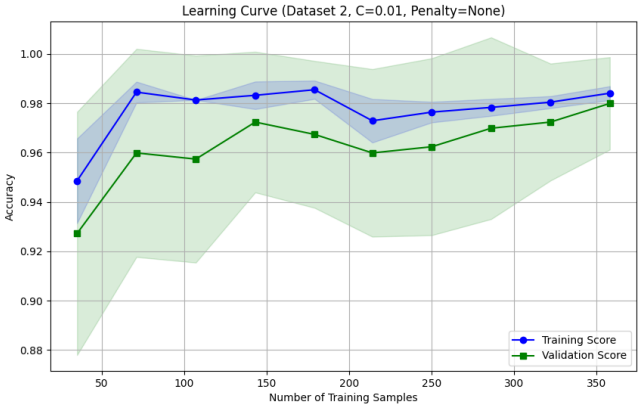
V. Mejoramiento del modelo

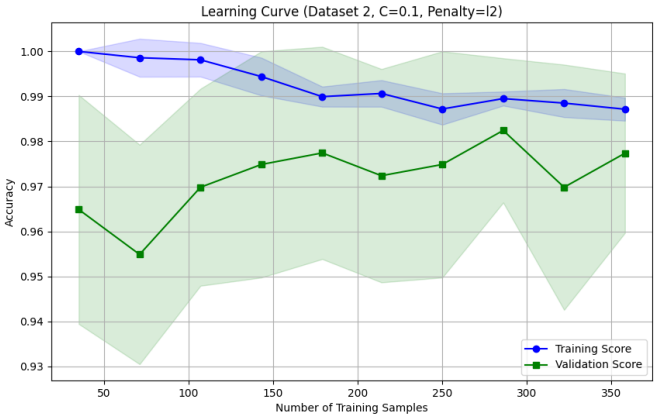
Existen diferentes técnicas durante el entrenamiento de un modelo que permiten monitorear el mismo, de manera que se evite el sesgo y varianza alta.

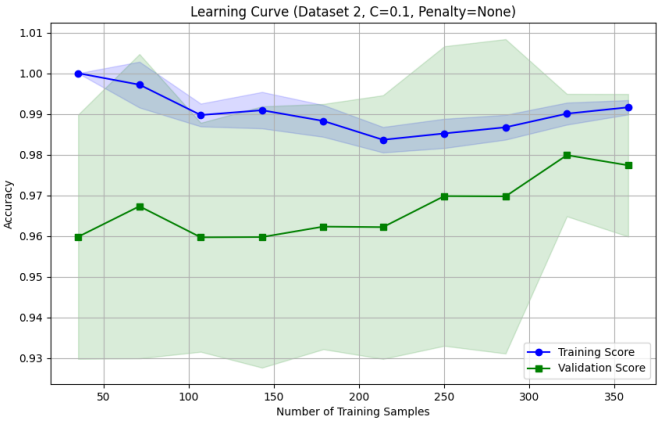


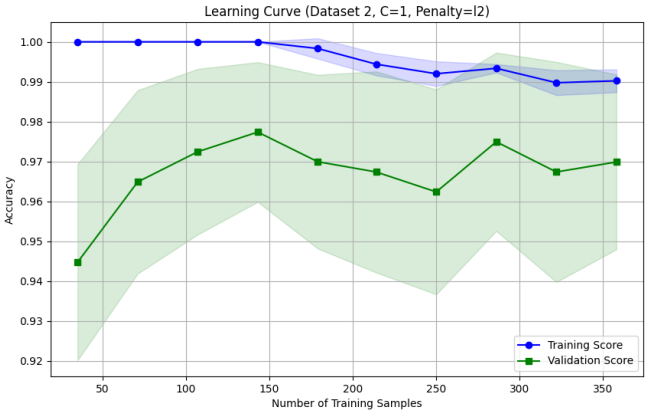


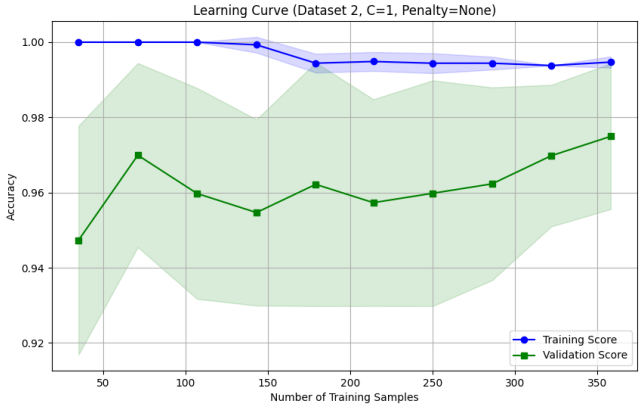












Las figuras anteriores muestran el cambio que sufre el entrenamiento del modelo cuando se comienza a variar con

las técnicas de regularización. En las gráficas se aprecian tres de estas técnicas: primero, el uso de *cross validation*, segundo, la cambio del parámetro de regularización *C*, y finalmente la regularización *L2*. El cambio del parámetro *C* refiere a la fuerza con la que el modelo busca maximizar la distancia entre las clases, aunque quite importancia a los errores de clasificación. Esto quiere decir que un *C* muy bajo podría resultar en sesgo alto, mientras que un valor alto podría resultar en varianza alta. En el caso de la regularización *L2*, esta refiere a una técnica que favorece los coeficientes bajos, pero no cero. Es decir, penaliza los coeficientes de altos de nuestra función de decisión.

En las gráficas podemos ver que mientras crece el valor de *C*, el rendimiento comienza a ser más pobre y el modelo comienza a sobre-entrenarse. Además, podemos notar la diferencia entre la aplicación de la regularización *L2*, donde los entrenamientos que penalizan loes coeficientes altos tienen ligeramente mejor desempeño.

VI. Conclusiones

Finalmente, observamos a través de diferentes análisis la diferencia de rendimiento según el tipo de algoritmo que usemos para entrenar, donde podemos claramente afirmar que el modelo implementado a través de una librería fue considerablemente mejor clasificando que el algoritmo manual. Además, el análisis del comportamiento del entrenamiento según el conjunto de datos y las técnicas de regularización nos permiten dar evidencia de las susceptibilidades que tienen estos modelados hacia la varianza y sesgo, además de ver el efecto que tienen estas técnicas para combatirlo.

Es posible, entonces, ver que el uso en conjunto de estas técnicas y herramientas nos permite generar modelos más precisos y confiables a la hora de implementarlos en problemáticas reales y conjuntos de datos posiblemente más complejos.

VII. Referencias