Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo. (Portafolio Análisis)

Facundo Vecchi A01283666

Introduccion

En el portafolio de implementacion se intento encontrar cual seria el mejor de 8 modelos para predecir el tipo de vino del Wine dataset. La conclusion de esa implementacion fue que el mejor modelo para el dataset de wine es el de Random Forest Classifier de Sklearn. El proposito de este analisis es utilizar un modelo y obtener su sesgo, varianza y diagnosticar su nivel de ajuste (underfit, fit u overfit) y finalmente aplicar tecnicas de refinacion

Modelo a utilizar

- **Regresion logistica**

Para este analisis, se utilizara el modelo de Regresion Logistica, aunque se concluyo que el mejor modelo es Random Forest, como obtuvo una precision de 100% no tiene mucho sentido utilizarlo para este analisis. Se escogio Regresion logisitica ya que es un modelo que vimos en clase y es mas simple de manipular los resultados con los parametros, ya que no depende de la aleatoridad.

Modelo base

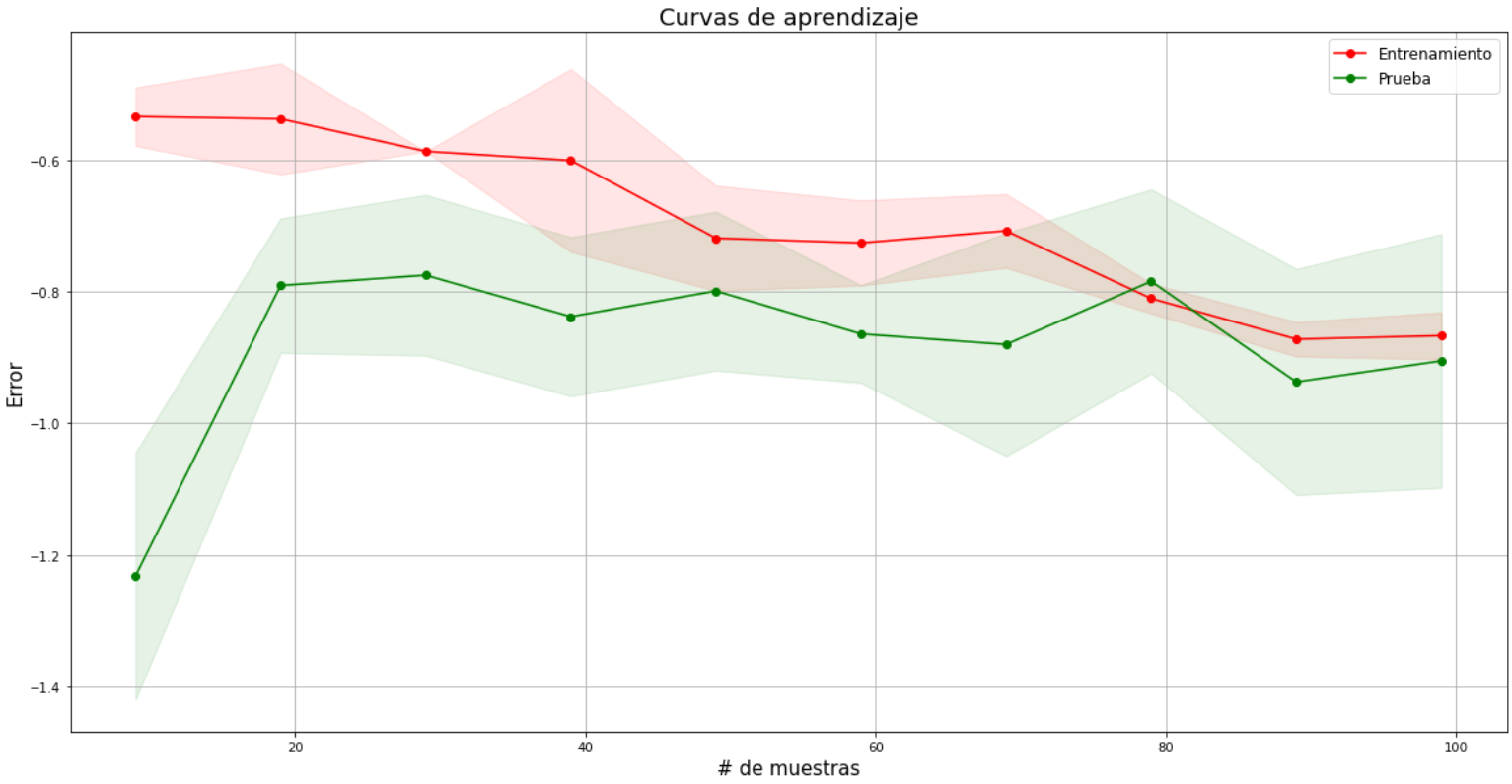
Primero vamos a observar el modelo base el cual tendra los siguiente parametros:

- Caracteristicas a utilizarze:
  - Cantidad de caracteristicas: 1
  - Nombres:
    - Alcohol
- Parametros relevantes por defecto de Regresion Lineal:
  - penalty = l2
  - solver = lbfgs
  - max_iter = 100
  - C = 1.0

Se empieza utilizando una sola caracteristica y los parametros por defecto para tener espacio de mejora en el modelo.

Resultados

- Sesgo Alto
- Varianza baja
- Underfit



Error total promedio	Sesgo promedio	Varianza promedio
0.298426	0.296296	0.076759

Como se puede observar en la grafica, ambas lineas de entrenamiento y prueba se acercan casi que al instante mientras aumenta el numero de muestras, esto es un claro indicador de que el modelo tiene un sesgo alto. Para confirmar lo visto en la grafica, se calculo el sesgo y varianza promedio del modelo, el cual tambien se muestra alto indicando que este modelo tiene underfit.

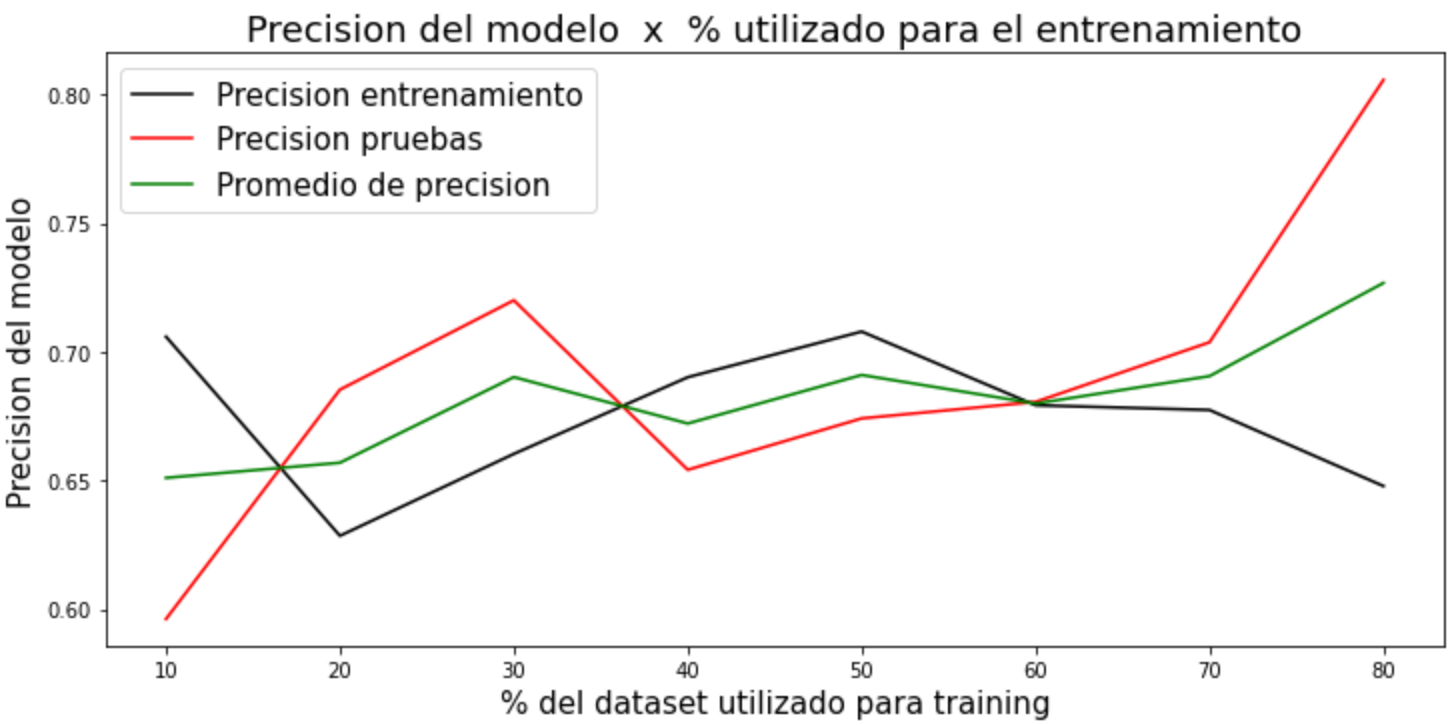
Splits de entrenamiento y pruebas

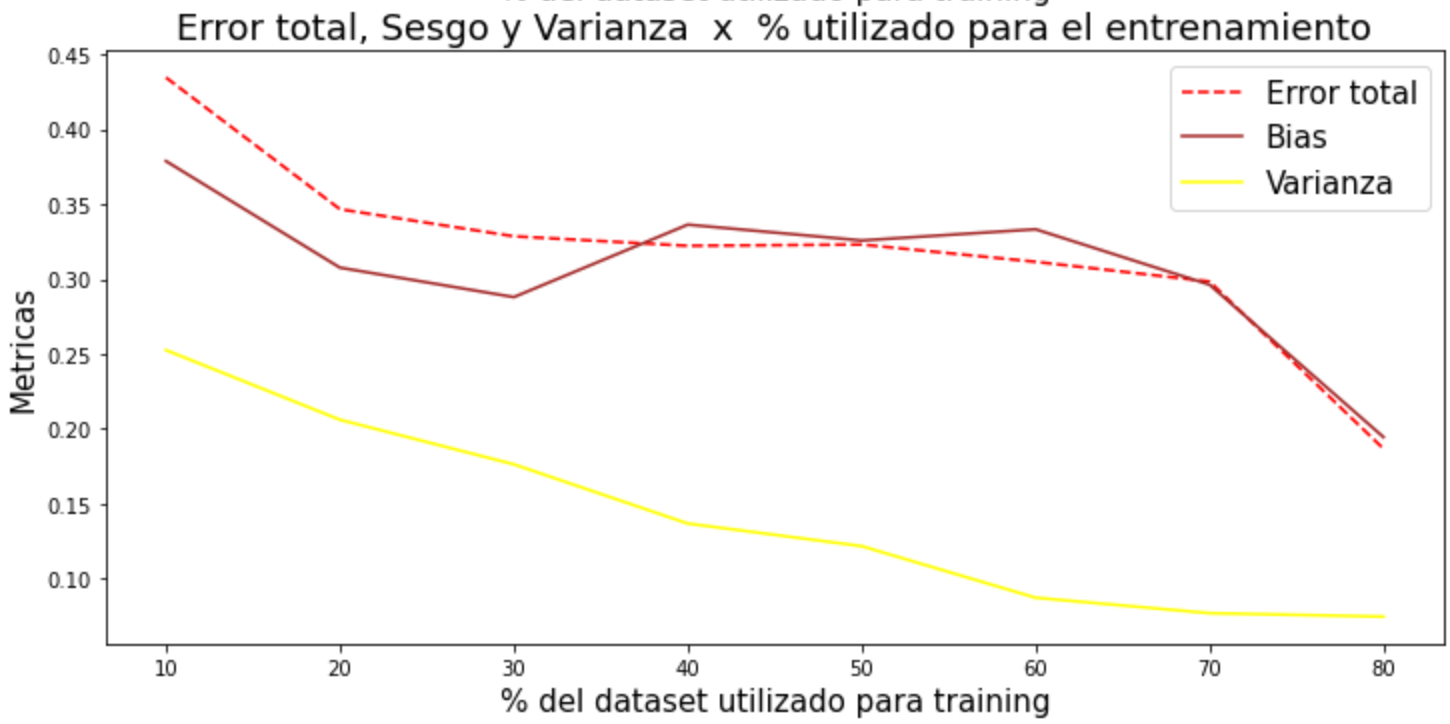
Antes de empezar a modificar los parametros, veremos cuanto cambia la precision del modelo, junto con el sesgo y su varianza cuando se utilizan diferentes porcentajes para dividir el dataset.
Para lograr esto, se va a calcular y graficar la precision del modelo, su error total, varianza y sesgo utilizando un porcentaje de 10% a 80% de entrenamiento en pasos de 10%. (En total se entrenarian 8 modelos ya que no tiene mucho sentido incrementar el porcentaje de entrenamiento aun mas ya que no se tendrian datos para validar el modelo)

Resultados

- Split a utilizar: 80%-20%

	Cantidad de muestras							
	10%-90%	20%-80%	30%-70%	40%-60%	50%-50%	60%-40%	70%-30%	80%-20%
Entrenamiento	17	35	53	71	89	106	124	142
Prueba	161	143	125	107	89	72	54	36





Observando las graficas podemos concluir que el mejor split parece ser el de 80%-20% ya que es en donde el modelo tiene la mayor precision promedio y la menor varianza, sesgo y error total, por lo que se concluye que este split es el mas estable y el que mejor utiliza el modelo para ajustarse al dataset.

Refinacion del modelo

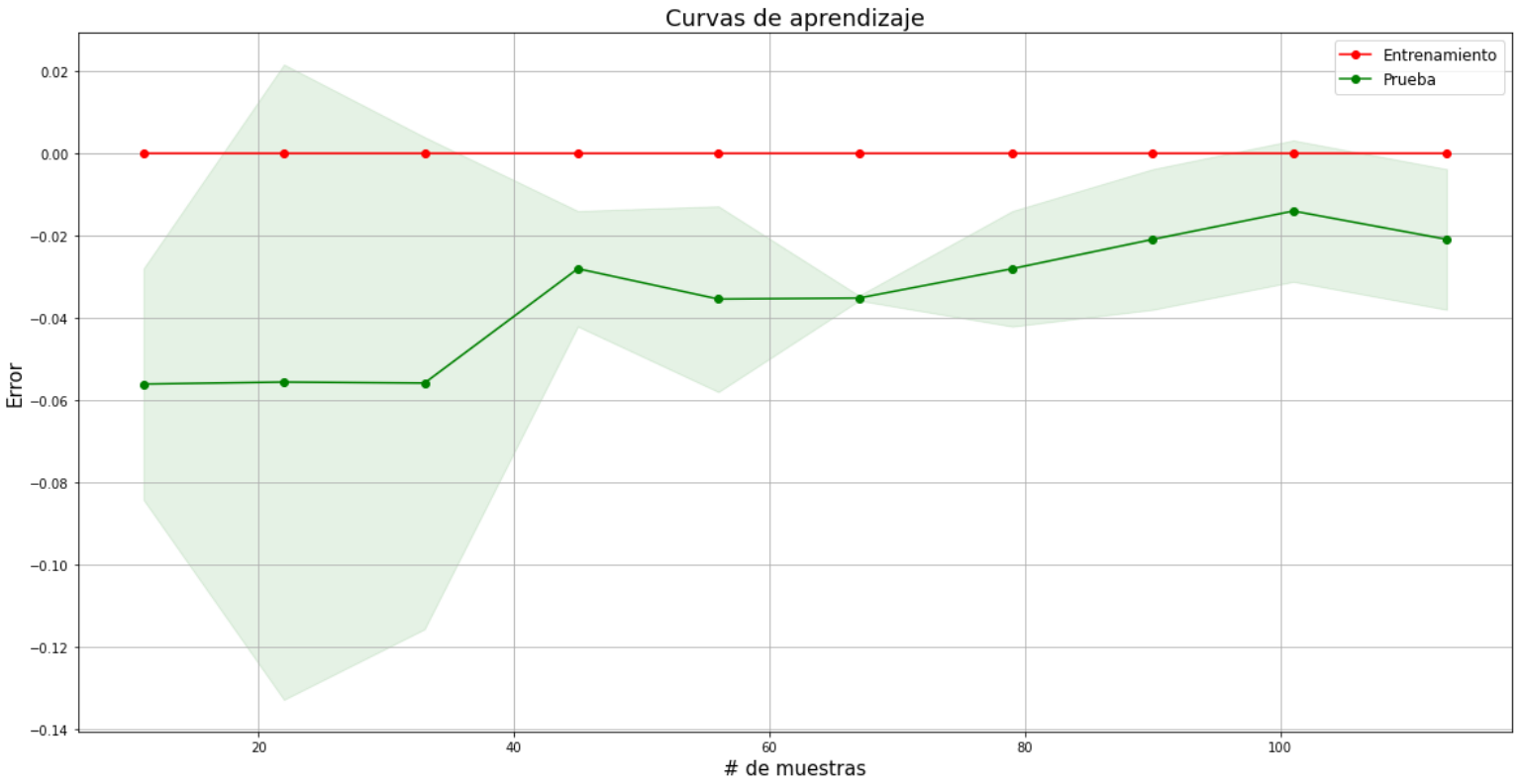
Utilizando lo descubierto anteriormente se utilizaran tecnicas de refinacion para mejorar la precision, error, sesgo y varianza del modelo. Sabiendo que el modelo tiene un sesgo alto, indicando underfitting, se van a incrementar la cantidad de caracteristicas que se utilizaran para entrenar el modelo y se modificaran los parametros del modelo para buscar mejorar el modelo aun mas.

Incremento de caracteristicas

- Caracteristicas a utilizarze:
  - Cantidad de caracteristicas: 13
  - Caracteristicas:
    - Alcohol
    - Malic acid
    - Ash
    - Alcalinity of ash
    - Magnesium
    - Total phenols
    - Flavanoids
    - Nonflavanoid phenols
    - Proanthocyanins
    - Color intensity
    - Hue
    - OD280/OD315 of diluted wines
    - Proline
- Parametros relevantes por defecto de Regresion Lineal:
  - penalty = l2
  - solver = lbfgs
  - max_iter = 100
  - C = 1.0

Resultados

- Sesgo bajo
- Varianza baja
- Fit



```
 Error total promedio  Sesgo promedio  Varianza promedio
              0.00625             0.0            0.00625
```

Como se puede observar en la grafica, el error en el set de entrenamiento se ha vuelto cero y el error del set de prueba ha disminuido en general. Tambien se puede observar como la linea de prueba se va lentamente acercando hacia la linea de entrenamiento mientras el numero de muestras va incrementando hasta estabilizarse alrededor de un error de -0.02, indicando que el modelo tiene un buen fit al dataset. Asimismo podemos confirmar lo que nos muestra la grafica con las calculaciones del error total, sesgo y varianza promedio, las cuales nos muestran como el error es muy bajo, el sesgo promedio es 0 y la varianza promedio tambien es muy baja.

Parametros del modelo

Ya que se sabe cual es el mejor split de entrenamiento y prueba y las caracteristicas a utilizarse para el dataset se empezara a probar con diferentes parametros del modelo con el proposito de reducir aun mas la varianza y el sesgo e incrementar la precision simultaneamente.
Para lograr esto, vamos a concentrarnos en 4 parametros del modelo de regresion logistica de Sklearn:

- solver : Algoritmo utilizado en la optimización
- C : Inverso de la fuerza de regularización
- max_iter : Número máximo de iteraciones necesarias para que los solucionadores converjan
- penalty : Especificar la norma del penalty

Con el proposito de facilitar la busqueda de estos parametros, se hara uso de un GridSearchCV para encontrar la mejor combinacion de parametros

```
Mejores parametros: {'C': 0.08858667904100823, 'max_iter': 100, 'penalty': 'l2', 'solver': 'newton-cg', 'verbose': 0}
```

Tras correr el GridSearchCV, este nos indica que, de los parametros que le dimos para probar, la mejor combinacion es la siguiente:

- penalty : l2
- solver : newton-cg
- max_iter : 100
- C : 0.08858667904100823

Resultados y Conclusiones

El mejor modelo (Refinado):

- Caracteristicas a utilizarze:
  - Cantidad de caracteristicas: 13
  - Caracteristicas:
    - Alcohol
    - Malic acid
    - Ash
    - Alcalinity of ash
    - Magnesium
    - Total phenols

        - Flavanoids
        - Nonflavanoid phenols
        - Proanthocyanins
        - Color intensity
        - Hue
        - OD280/OD315 of diluted wines
        - Proline
- Parametros relevantes por defecto de Regresion Lineal:
    - penalty = l2
    - solver = newton-cg
    - max_iter = 100
    - C = 0.08858667904100823

Comparativa general de sesgo y varianza

	Sesgo	**Varianza**
Base	Alto	Baja
Refinado	Bajo	Baja

Comparativa detallada de los modelos

```
          Precision  Error total esperado promedio  Sesgo promedio  Varianza promedio
base       0.722222                       0.298426        0.296296           0.076759
refinado   0.972222                       0.006528             0.0           0.006528
```

Como se puede observar, al añadir mas caracteristicas, probar con diferentes tamaños de splits y probar con diferentes combinaciones de parametros el modelo mejora en varios aspectos. Al incrementar el tamaño del split de entrenamiento podemos reducir la varianza signficativamente y al añadir mas caracteristicas podemos reducir el sesgo aun mas.