Tecnológico de Monterrey

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

Materia:

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Profesor:

Docente: Ivan Mauricio Amaya Contreras

Alumno

Gilberto Ramos Salinas A01734128

Fecha

08 de Septiembre del 2022

Introducción

A lo largo de este módulo tuve la oportunidad de realizar diferentes implementaciones sobre el uso de Machine Learning en distintas bases de datos con el objetivo de mejorar dicha implementación en futuras entregas. Por ello, el análisis de este documento sera sobre un proyecto realizado con frameworks como pandas, numpy y sklearn. Este proyecto fue ejecutado con una base de datos proveniente de Kaggle denominada Iris, la cual contiene datos provenientes como la longitud cepal, ancho cepal, longitud del pétalo y ancho del pétalo. El propósito principal es generar una predicción del subtipo de la planta iris para facilitar el trabajo de clasificación.

Exploración de datos

Antes de realizar el análisis es necesario revisar el dataset para ver la calidad de la información. Esto se hace con el objetivo de observar si el dataset cuenta con espacios vacíos o con información incompleta. También se puede ver la relación que existe entre las variables para ver si existe una correlación grande entre las especies y sus atributos.
Esta visualización puede ser obtenida a través de los siguientes comandos:

El análisis debe de contener los siguientes elementos:

```
##Checar valores no nulos
df.notnull().sum()
```

```
Id              150
SepalLengthCm   150
SepalWidthCm    150
PetalLengthCm   150
PetalWidthCm    150
Species         150
No_Species      150
dtype: int64
```

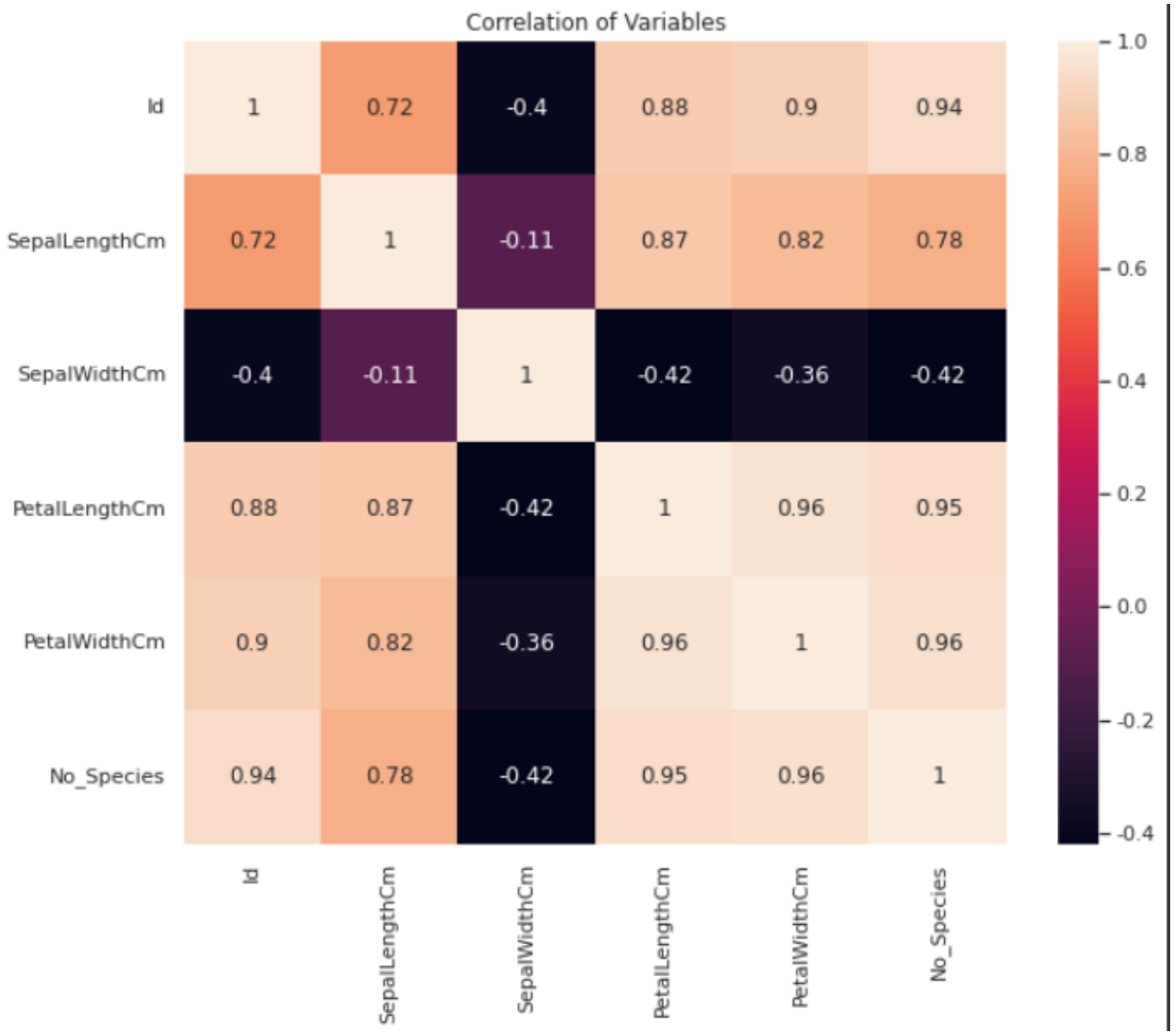
```
[71] ##Checar valores nulos
     df.isnull().sum()
```

```
Id              0
SepalLengthCm   0
SepalWidthCm    0
PetalLengthCm   0
PetalWidthCm    0
Species         0
No_Species      0
dtype: int64
```

```
##Checar blancos
df.isna().sum()
```

```
Id              0
SepalLengthCm   0
SepalWidthCm    0
PetalLengthCm   0
PetalWidthCm    0
Species         0
No_Species      0
dtype: int64
```

Por medio del siguiente diagrama de calor podemos observar que unas variables presentan una mayor correlación que otras. Sin embargo, esto no fue considerado determinante para



Separación y evaluación del modelo con un conjunto de prueba y un conjunto de validación (Train/Test/Validation).

La separación es obtenida a través de la librería sklearns con X_train, x_test, Y_train, y_test. La validación es obtenida por el método K Fold, el cual revisa la veracidad de un modelo machine learning. Para este caso al tratarse de la base de datos Iris se decidió implementar un modelo de regresión logística.

```
[ ]  from sklearn.model_selection import train_test_split
     X_train, X_test, y_train, y_test = train_test_split(x, y, random_state=42)
```

```
Reporte de Clasificación para  Regresion Logistica

              precision    recall  f1-score   support

           0       1.00      1.00      1.00        15
           1       1.00      1.00      1.00        11
           2       1.00      1.00      1.00        12

    accuracy                           1.00        38
   macro avg       1.00      1.00      1.00        38
weighted avg       1.00      1.00      1.00        38
```

```python
from sklearn.model_selection import KFold
from sklearn.metrics import make_scorer
from sklearn.model_selection import cross_validate

kfold = KFold(n_splits=30, shuffle=True, random_state=42)
scorer = make_scorer(accuracy_score)

veracity = cross_validate(neural_netowrk, x, y, cv=kfold, scoring=scorer)
```

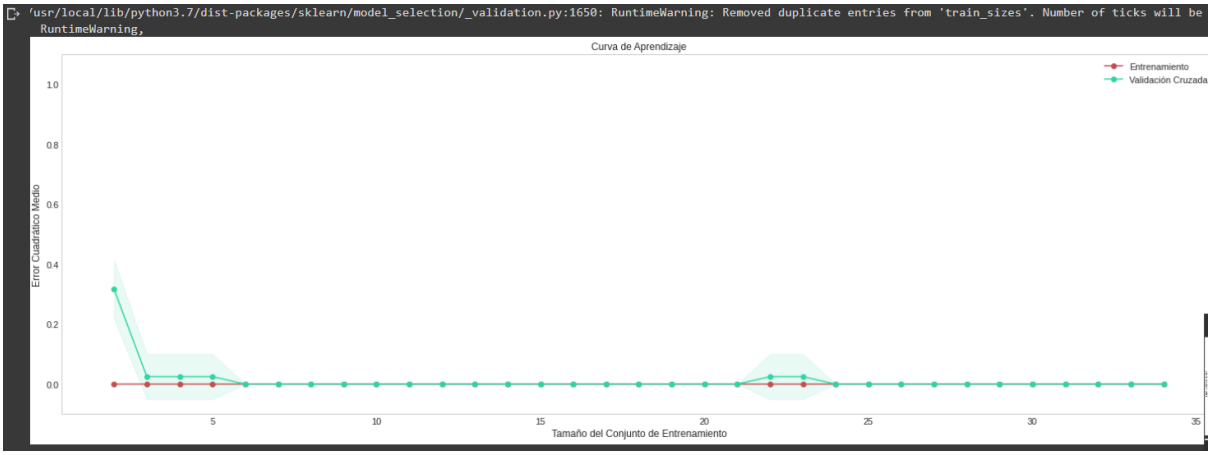
```
Accuracy of the Linear Regretion model with k-fold cross validation
K-fold accuracies:  [1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1.]
Mean : 1.0
Variance:  0.0
Standard deviation:  0.0
Bias:  0.0
```

Diagnóstico y explicación el grado de bias o sesgo: bajo medio alto

Anteriormente se puede apreciar como los resultados del K Fold demuestran como la diferencia entre el promedio del valor predicho contra el real es 0. Demostrando que el modelo comprende con exactitud la información. Sin embargo, en la vida real no todos los modelos dan este grado de predicción ya que los datos pueden ser más complejos, pero al tratarse de algo tan simple como una clasificación el modelo fue capaz de aprenderlo.

Diagnóstico y explicación el grado de varianza: bajo medio alto

La varianza como tal nos explica la distribución de los datos respecto a la media. Por ello, la varianza en este caso nos da 0 ya que esto puede ser únicamente calculado cuando la media es diferente de 1. Esto demuestra que no existe distribución dispersa de datos respecto a la media.



Diagnóstico y explicación el nivel de ajuste del modelo: underfitt fitt overfitt

Basándonos en los resultados previos se pudo apreciar que la veracidad del modelo es de 1 indicando que el modelo no presenta un overfitting o un underfitting.

Basándote en lo encontrado en tu análisis utiliza técnicas de regularización o ajuste de parámetros para mejorar el desempeño de tu modelo y documenta en tu reporte cómo mejoró este.

En el caso de este modelo, los resultados de las predicciones fueron excelentes por parte del modelo de regresión logística. Sin embargo, este caso no se puede aplicar para otros estudios. Por ello, en el caso de que las variables hubieran dañado el entrenamiento se pudiera implementar el uso de hiperparametros, los cuales se obtienen a través de un análisis que nos regresa las variables que realicen un mejor entrenamiento en el modelo.

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Predicted Value	Real Values
0	1	5.1	3.5	1.4	0.2	1.0	1.0
1	2	4.9	3.0	1.4	0.2	0.0	0.0
2	3	4.7	3.2	1.3	0.2	2.0	2.0
3	4	4.6	3.1	1.5	0.2	1.0	1.0
4	5	5.0	3.6	1.4	0.2	1.0	1.0
5	6	5.4	3.9	1.7	0.4	0.0	0.0
6	7	4.6	3.4	1.4	0.3	1.0	1.0
7	8	5.0	3.4	1.5	0.2	2.0	2.0
8	9	4.4	2.9	1.4	0.2	1.0	1.0
9	10	4.9	3.1	1.5	0.1	1.0	1.0
10	11	5.4	3.7	1.5	0.2	2.0	2.0
11	12	4.8	3.4	1.6	0.2	0.0	0.0
12	13	4.8	3.0	1.4	0.1	0.0	0.0
13	14	4.3	3.0	1.1	0.1	0.0	0.0