Tecnológico de Monterrey

Tecnológico y de Estudios Superiores de Monterrey

Inteligencia artificial avanzada para la ciencia de datos I

Profesor: Jorge Adolfo Ramírez Uresti

Módulo 2 Uso de framework o biblioteca de aprendizaje máquina para la implementación de una solución

Liam Garay Monroy A01750632

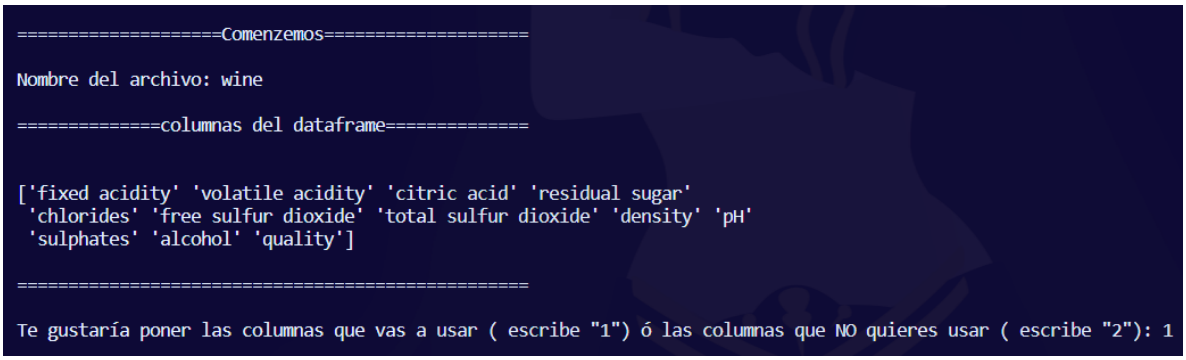
12 de Septiembre de 2022

Pruebas del algoritmo Random Forest Classifier

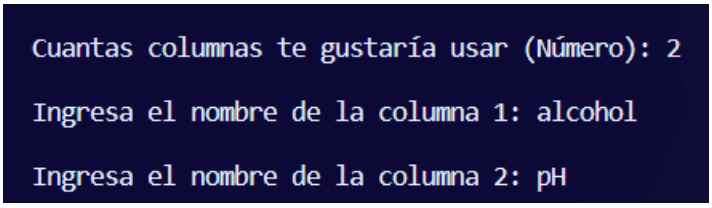
A continuación, se presentarán distintas pruebas del algoritmo con las cuales podremos revisar cual es el accuracy de nuestro modelo dependiendo de los distintos híper parámetros que utilicemos, así como las columnas disponibles, entre otras cosas.

Primera prueba:

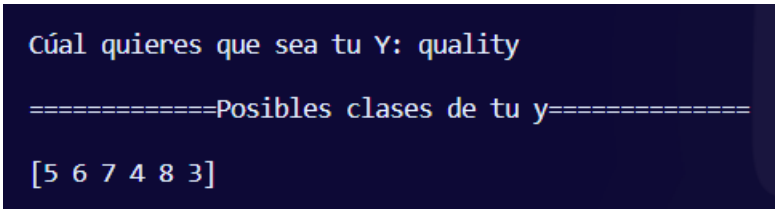
Primero se nos pide el nombre del archivo y si queremos especificar columnas a usar o a quitar



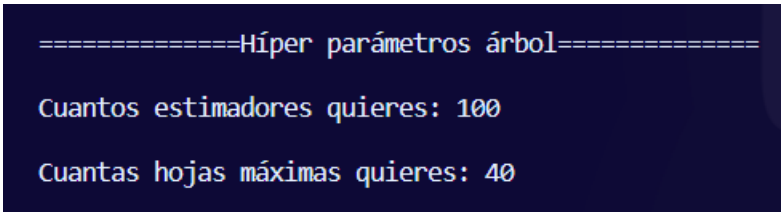
Se nos pide también el número de columnas que utilizaremos (si elegimos la opción 1), así como que especifiquemos cuales serán estas



También se nos pide especificar nuestra Y, y se nos despliega una lista con los posibles valores que puede tomar Y



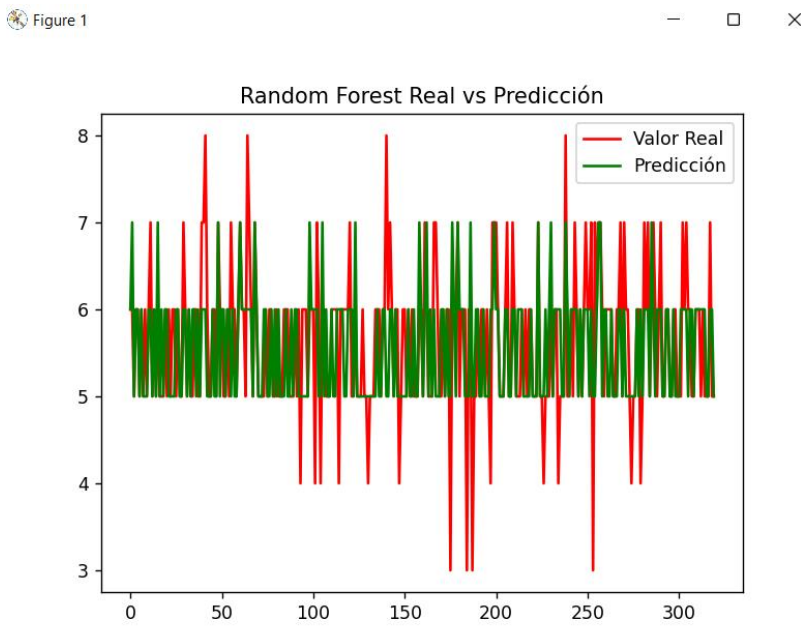
También podemos especificar los híper parámetros de nuestro árbol



Con esto se nos otorga un accuracy de nuestro modelo

```
======================Accuracy======================

random forest accuracy: 0.55
```

También se nos despliega un pequeño gráfico que nos muestra la diferencia entre el valor real de train y la predicción que se hizo con los datos de train



Posteriormente se nos pide que asignemos valores a las columnas que especificamos serían nuestra X

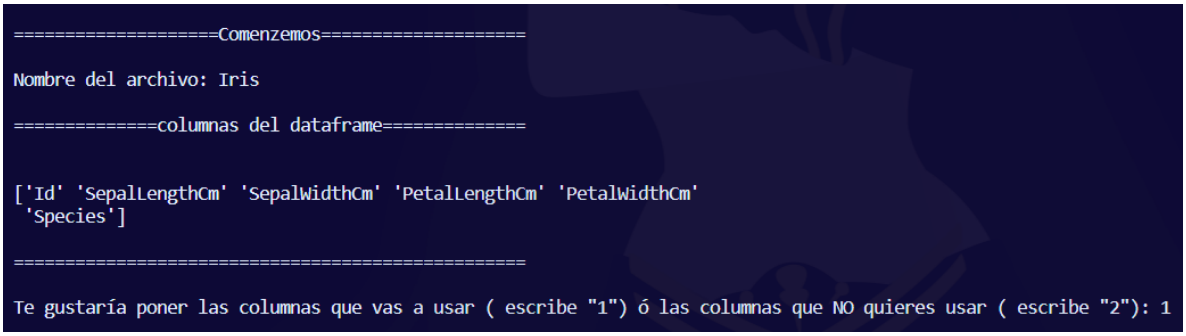
```
===============Valores de columnas================

Valor a asignar en la columna alcohol: 9.5

Valor a asignar en la columna pH: 3.38

==================Probabilidades==================
```

Y por último nos arroja las probabilidades sobre las clases a las que puede pertenecer nuestra predicción y la clase a la que se predice que pertenece teniendo en cuenta las probabilidades

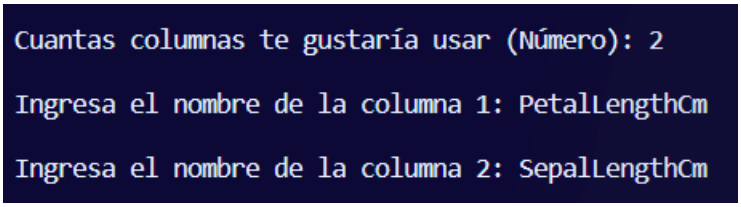
```
==================Probabilidades==================
Valores de Y: [3 4 5 6 7 8]

probabilidad de las clases [[9.5, 3.38]] [[9.80660078e-04 4.00261401e-02 6.26038585e-01 2.98690001e-01
  3.41509446e-02 1.13668803e-04]]

================Predicción final================
tu predicción de la columna quality tiene un valor de: [5]
```

Prueba con otro archivo (Iris) hacemos lo mismo que con el anterior:

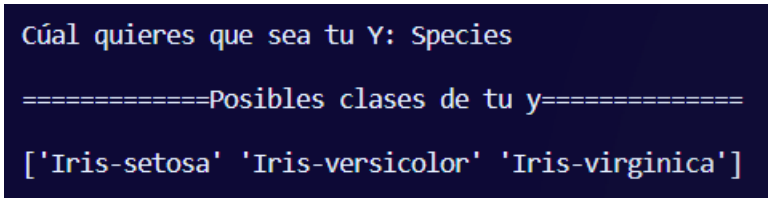
Primero se nos pide el nombre del archivo y si queremos especificar columnas a usar o a quitar



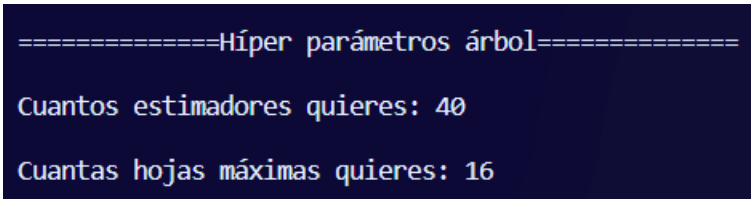
Se nos pide también el número de columnas que utilizaremos (si elegimos la opción 1), así como que especifiquemos cuales serán estas



También se nos pide especificar nuestra Y, y se nos despliega una lista con los posibles valores que puede tomar Y



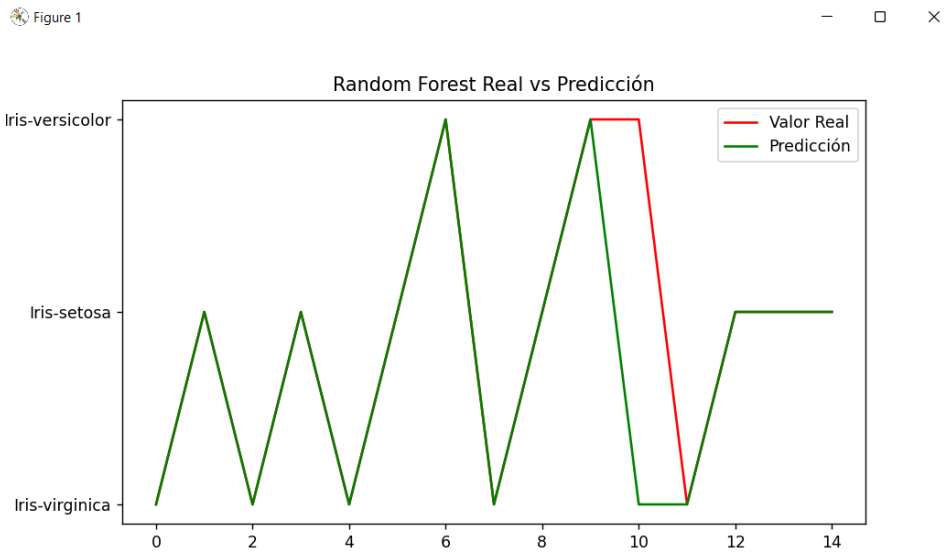
También podemos especificar los híper parámetros de nuestro árbol



Con esto se nos otorga un accuracy de nuestro modelo

```
====================Accuracy====================

random forest accuracy: 0.9333333333333333
```

También se nos despliega un pequeño gráfico que nos muestra la diferencia entre el valor real de train y la predicción que se hizo con los datos de train



Posteriormente se nos pide que asignemos valores a las columnas que especificamos serían nuestra X

```
==============Valores de columnas===============

Valor a asignar en la columna PetalLengthCm: 4.5

Valor a asignar en la columna SepalLengthCm: 1.3
```

Y por último nos arroja las probabilidades sobre las clases a las que puede pertenecer nuestra predicción y la clase a la que se predice que pertenece teniendo en cuenta las probabilidades

```
==================Probabilidades==================
Valores de Y: ['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']

probabilidad de las clases [[4.5, 1.3]] [[0.325 0.2   0.475]]

=================Predicción final=================
tu predicción de la columna Species tiene un valor de: ['Iris-virginica']
```

Con el ejemplo anterior podemos variar los híper parámetros de nuestro árbol y ver cuanto eso repercute en nuestro accuracy

```
=============Híper parámetros árbol=============

Cuantos estimadores quieres: 1

Cuantas hojas máximas quieres: 3

====================Accuracy====================

random forest accuracy: 1.0
```

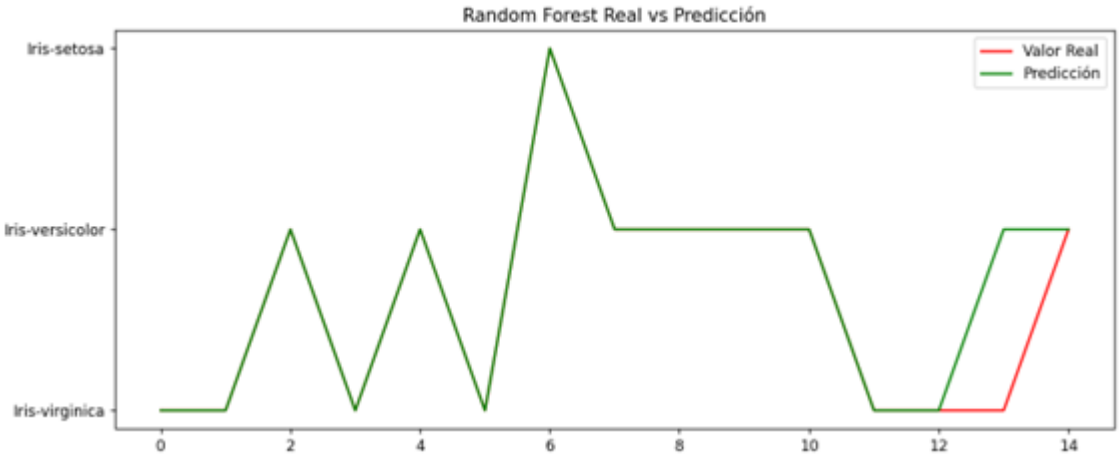
Y nuestra predicción tanmbien cambia a pesar de ser los mismos datos

```
==============Valores de columnas===============

Valor a asignar en la columna PetalLengthCm: 4.5

Valor a asignar en la columna SepalLengthCm: 1.3

==================Probabilidades==================
Valores de Y: ['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']

probabilidad de las clases [[4.5, 1.3]] [[0.         0.94871795 0.05128205]]

=================Predicción final=================
tu predicción de la columna Species tiene un valor de: ['Iris-versicolor']
```

Ahora con híper parámetros diferentes un tanto más grandes

```
==============Híper parámetros árbol==============

Cuantos estimadores quieres: 400

Cuantas hojas máximas quieres: 100

====================Accuracy=====================

random forest accuracy: 0.9333333333333333
```

Podemos apreciar que el accuracy se mantuvo, pero la gráfica cambio en cuanto a que predicciones erraron

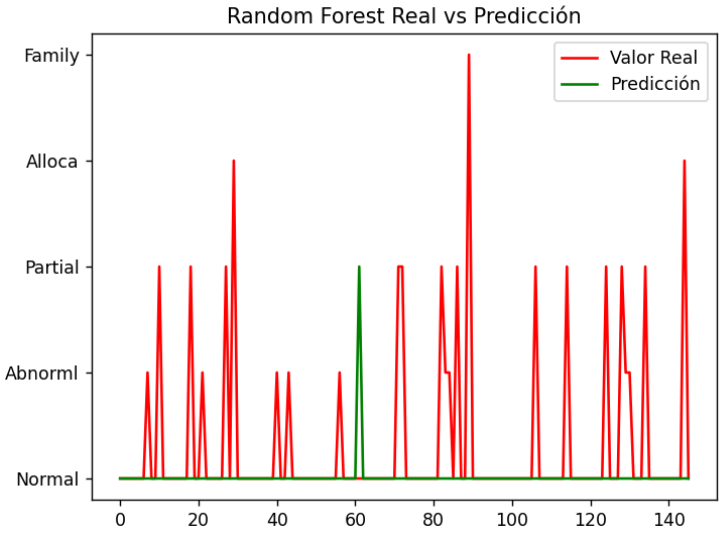


En conclusión: los modelos dependen de muchas cosas, no solo del modelo en sí, sino también en los híper parámetros que nosotros coloquemos, como se encuentre nuestro dataset, si es que este tiene valores sesgados o no, es importante mencionar que para analizar datos de una forma más concisa se requiere el analizar los datos de forma profunda incluso antes de entra al modelo, realizar una limpieza adecuada y estandarizar valores, con la finalidad de ayudar a nuestro modelo a realizar mejores predicciones y dependiendo de dicha preparación podremos esperar unos resultados u otros.

Prueba con otro data set

```
==================Comenzemos==================

Nombre del archivo: train

=============columnas del dataframe=============


['Id' 'MSSubClass' 'MSZoning' 'LotFrontage' 'LotArea' 'Street' 'Alley'
 'LotShape' 'LandContour' 'Utilities' 'LotConfig' 'LandSlope'
 'Neighborhood' 'Condition1' 'Condition2' 'BldgType' 'HouseStyle'
 'OverallQual' 'OverallCond' 'YearBuilt' 'YearRemodAdd' 'RoofStyle'
 'RoofMatl' 'Exterior1st' 'Exterior2nd' 'MasVnrType' 'MasVnrArea'
 'ExterQual' 'ExterCond' 'Foundation' 'BsmtQual' 'BsmtCond' 'BsmtExposure'
 'BsmtFinType1' 'BsmtFinSF1' 'BsmtFinType2' 'BsmtFinSF2' 'BsmtUnfSF'
 'TotalBsmtSF' 'Heating' 'HeatingQC' 'CentralAir' 'Electrical' '1stFlrSF'
 '2ndFlrSF' 'LowQualFinSF' 'GrLivArea' 'BsmtFullBath' 'BsmtHalfBath'
 'FullBath' 'HalfBath' 'BedroomAbvGr' 'KitchenAbvGr' 'KitchenQual'
 'TotRmsAbvGrd' 'Functional' 'Fireplaces' 'FireplaceQu' 'GarageType'
 'GarageYrBlt' 'GarageFinish' 'GarageCars' 'GarageArea' 'GarageQual'
 'GarageCond' 'PavedDrive' 'WoodDeckSF' 'OpenPorchSF' 'EnclosedPorch'
 '3SsnPorch' 'ScreenPorch' 'PoolArea' 'PoolQC' 'Fence' 'MiscFeature'
 'MiscVal' 'MoSold' 'YrSold' 'SaleType' 'SaleCondition' 'SalePrice']

================================================

Te gustaría poner las columnas que vas a usar ( escribe "1") ó las columnas que NO quieres usar ( escribe "2"): 1
```

```
=====================Accuracy=====================

random forest accuracy: 0.8287671232876712
```



```
==============Valores de columnas===============

Valor a asignar en la columna OverallQual: 5

Valor a asignar en la columna OverallCond: 5

=================Probabilidades=================
Valores de Y: ['Abnorml' 'AdjLand' 'Alloca' 'Family' 'Normal' 'Partial']

probabilidad de las clases [[5, 5]] [[0.08955513 0.00419853 0.01235771 0.0117561  0.83130409 0.05082843]]

================Predicción final================
tu predicción de la columna SaleCondition tiene un valor de: ['Normal']
```