Estimación de anomalías de temperatura del aire en el departamento de Antioquia

Modelos supervisados

Valentina Sánchez Castaño

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Para recordar...

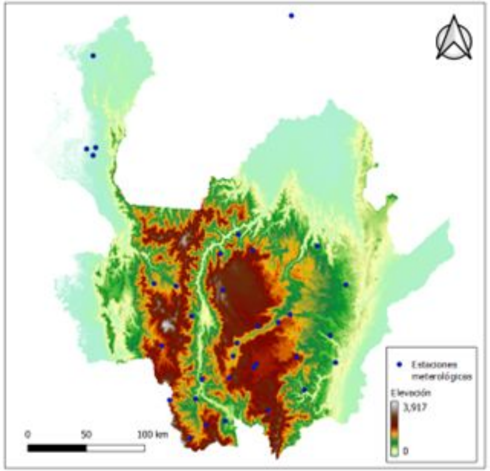
X

- Evaporación total (m of water equivalent)
- Temperatura del suelo (K)
- Cobertura de nubes (%)
- Velocidad del viento (m/s)
- Tipo de cobertura (-)
- NDVI (-)
- Modelo digital de elevación

Y

- Temperatura del aire

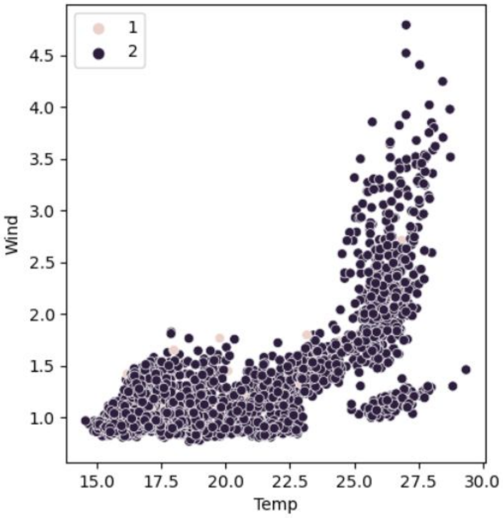
Datos desde 2003 hasta el 2014 y 36 estaciones

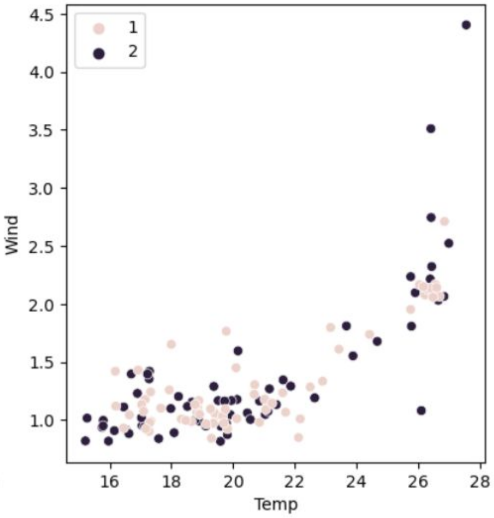


Técnicas de preprocesamiento

Random Under Sampler

sampling_strategy='majority'





0	1
5028	106

Desbalance

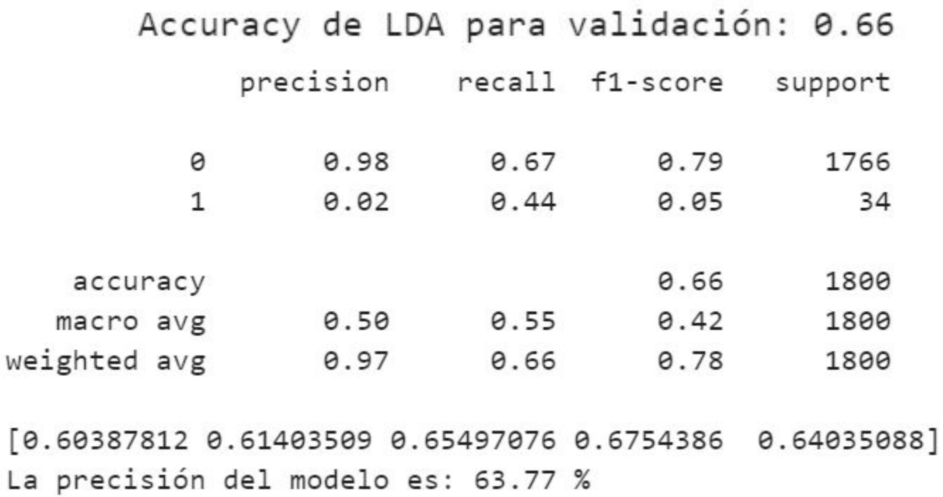
2.11% de los datos

I. Análisis discriminante lineal

Linear Discriminant Analysis

Test_size = 0.3

1175	591
19	15



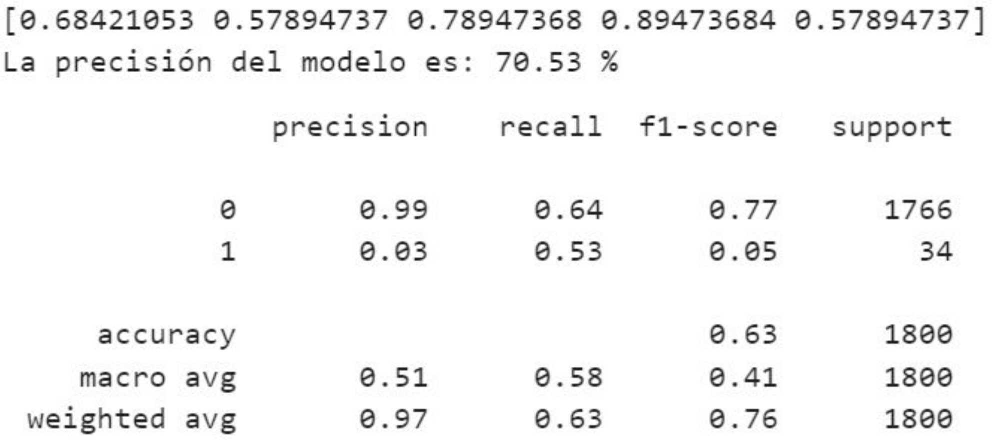
```
Accuracy de LDA para validación: 0.66
              precision    recall  f1-score   support

           0       0.98      0.67      0.79      1766
           1       0.02      0.44      0.05        34

    accuracy                           0.66      1800
   macro avg       0.50      0.55      0.42      1800
weighted avg       0.97      0.66      0.78      1800

[0.60387812 0.61403509 0.65497076 0.6754386  0.64035088]
La precisión del modelo es: 63.77 %
```

Solo se tiene un componente discriminante

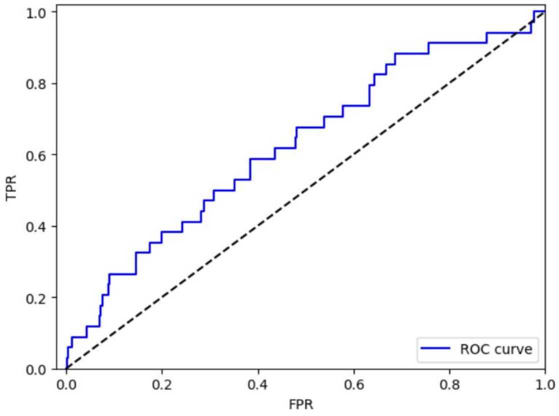
Linear Discriminant Analysis y KFold

Con KFold incrementa la precisión del modelo, pero no mejora en la predicción de I.

1122	644
16	18

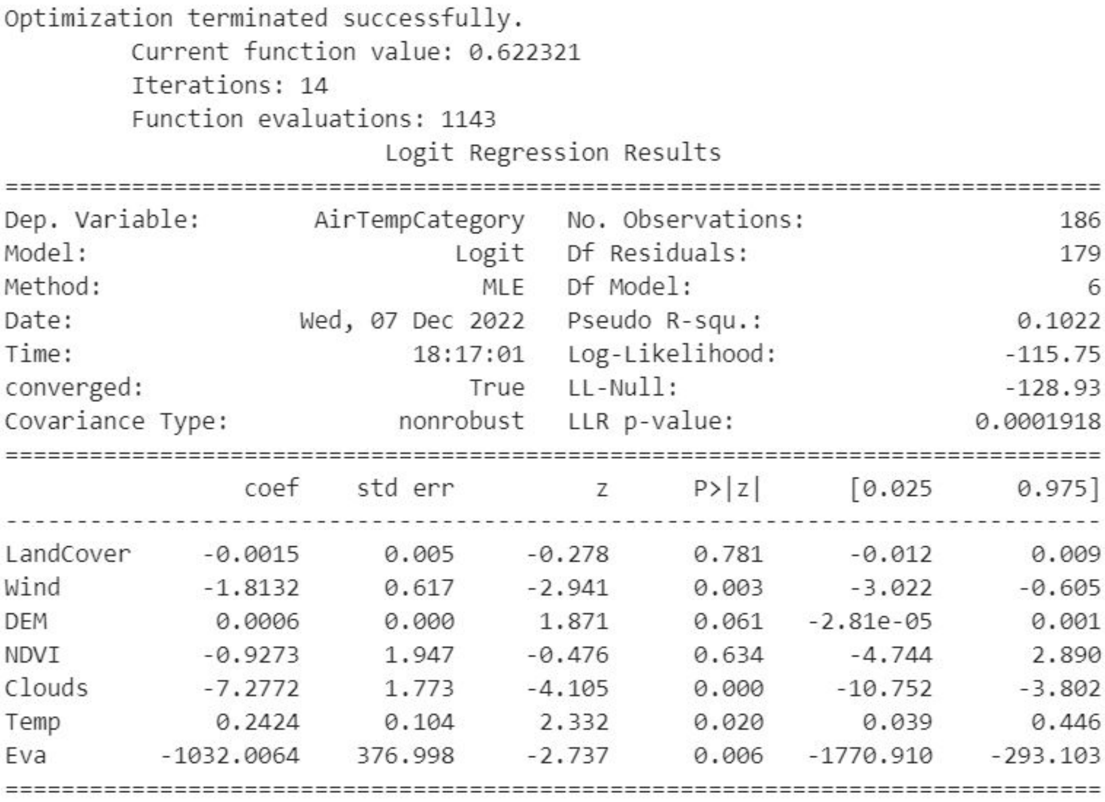


```
[0.68421053 0.57894737 0.78947368 0.89473684 0.57894737]
La precisión del modelo es: 70.53 %
              precision    recall  f1-score   support

           0       0.99      0.64      0.77      1766
           1       0.03      0.53      0.05        34

    accuracy                           0.63      1800
   macro avg       0.51      0.58      0.41      1800
weighted avg       0.97      0.63      0.76      1800
```



2. Regresión logística

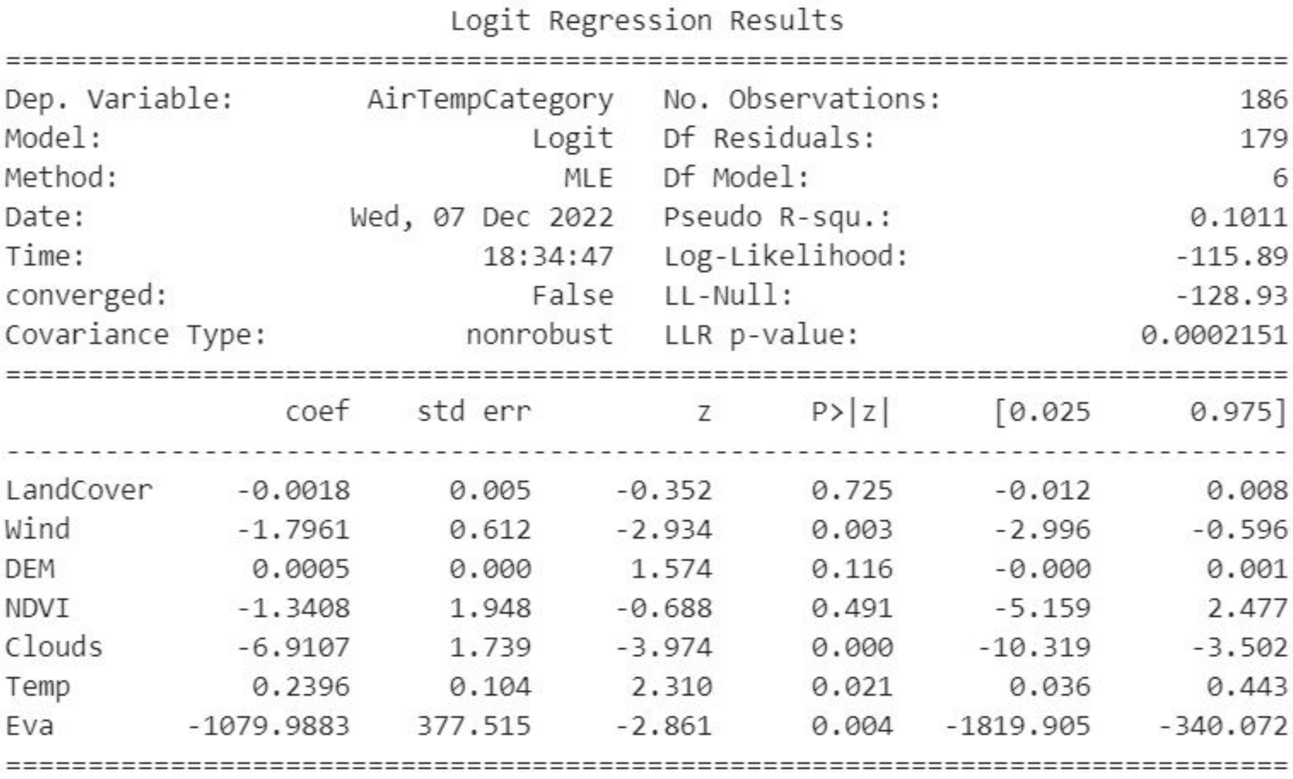
Regresión Logística con Statsmodels



```
Optimization terminated successfully.
         Current function value: 0.622321
         Iterations: 14
         Function evaluations: 1143
                           Logit Regression Results
==============================================================================
Dep. Variable:        AirTempCategory   No. Observations:                  186
Model:                          Logit   Df Residuals:                      179
Method:                           MLE   Df Model:                            6
Date:                Wed, 07 Dec 2022   Pseudo R-squ.:                  0.1022
Time:                        18:17:01   Log-Likelihood:                -115.75
converged:                       True   LL-Null:                       -128.93
Covariance Type:            nonrobust   LLR p-value:                 0.0001918
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
LandCover     -0.0015      0.005     -0.278      0.781      -0.012       0.009
Wind          -1.8132      0.617     -2.941      0.003      -3.022      -0.605
DEM            0.0006      0.000      1.871      0.061   -2.81e-05       0.001
NDVI          -0.9273      1.947     -0.476      0.634      -4.744       2.890
Clouds        -7.2772      1.773     -4.105      0.000     -10.752      -3.802
Temp           0.2424      0.104      2.332      0.020       0.039       0.446
Eva        -1032.0064    376.998     -2.737      0.006   -1770.910    -293.103
==============================================================================
```

Se obtuvieron valores P favorables para unas variables, las otras no son suficientes para tener una buena predicción

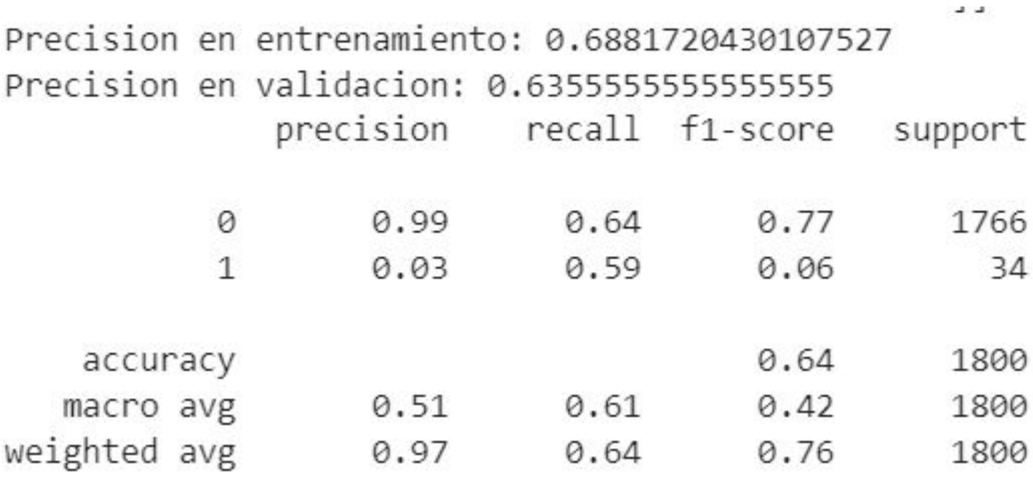
Regresión Logística con Statsmodels



```
                           Logit Regression Results
==============================================================================
Dep. Variable:        AirTempCategory   No. Observations:                  186
Model:                          Logit   Df Residuals:                      179
Method:                           MLE   Df Model:                            6
Date:                Wed, 07 Dec 2022   Pseudo R-squ.:                  0.1011
Time:                        18:34:47   Log-Likelihood:                -115.89
converged:                      False   LL-Null:                       -128.93
Covariance Type:            nonrobust   LLR p-value:                 0.0002151
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
LandCover     -0.0018      0.005     -0.352      0.725      -0.012       0.008
Wind          -1.7961      0.612     -2.934      0.003      -2.996      -0.596
DEM            0.0005      0.000      1.574      0.116      -0.000       0.001
NDVI          -1.3408      1.948     -0.688      0.491      -5.159       2.477
Clouds        -6.9107      1.739     -3.974      0.000     -10.319      -3.502
Temp           0.2396      0.104      2.310      0.021       0.036       0.443
Eva        -1079.9883    377.515     -2.861      0.004   -1819.905    -340.072
==============================================================================
```

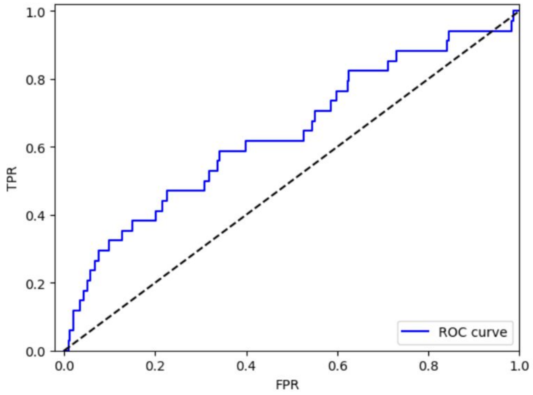
Empleando el parámetro de máximas interacciones se disminuye el P-value, haciéndolas más significativas a la hora de predecir.

Regresión Logística con SKlearn

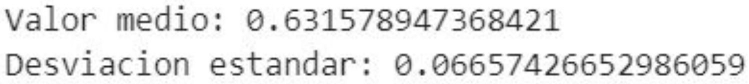


```
Precision en entrenamiento: 0.6881720430107527
Precision en validacion: 0.6355555555555555
              precision    recall  f1-score   support

           0       0.99      0.64      0.77      1766
           1       0.03      0.59      0.06        34

    accuracy                           0.64      1800
   macro avg       0.51      0.61      0.42      1800
weighted avg       0.97      0.64      0.76      1800
```

Penalidad L1 (Lasso)

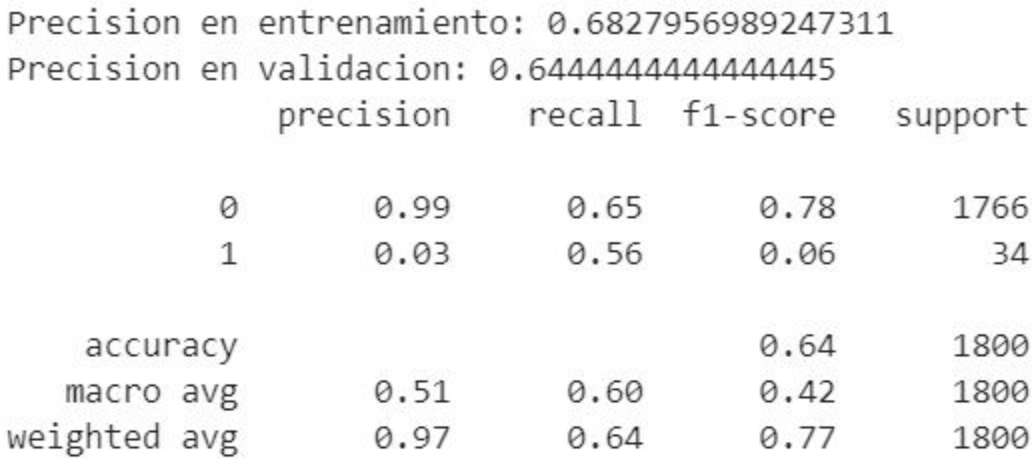


El KFold no mejora la predicción para la clase 1



```
Valor medio: 0.631578947368421
Desviacion estandar: 0.06657426652986059
```

Regresión Logística con SKlearn



```
Precision en entrenamiento: 0.6827956989247311
Precision en validacion: 0.6444444444444445
              precision    recall  f1-score   support

           0       0.99      0.65      0.78      1766
           1       0.03      0.56      0.06        34

    accuracy                           0.64      1800
   macro avg       0.51      0.60      0.42      1800
weighted avg       0.97      0.64      0.77      1800
```

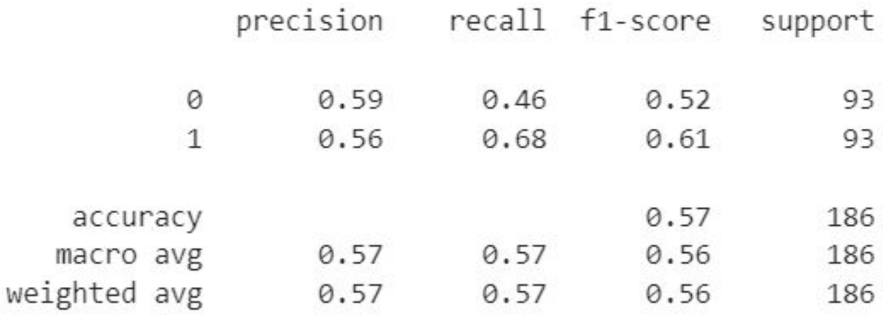
Penalidad L2
(Ridge Regression)

Disminuye el recall

3. K-nearest neighbors

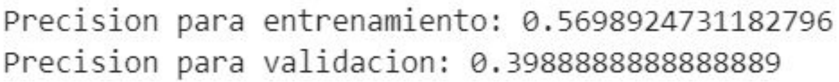
KNeighborsClassifier

N_neighbors = 15



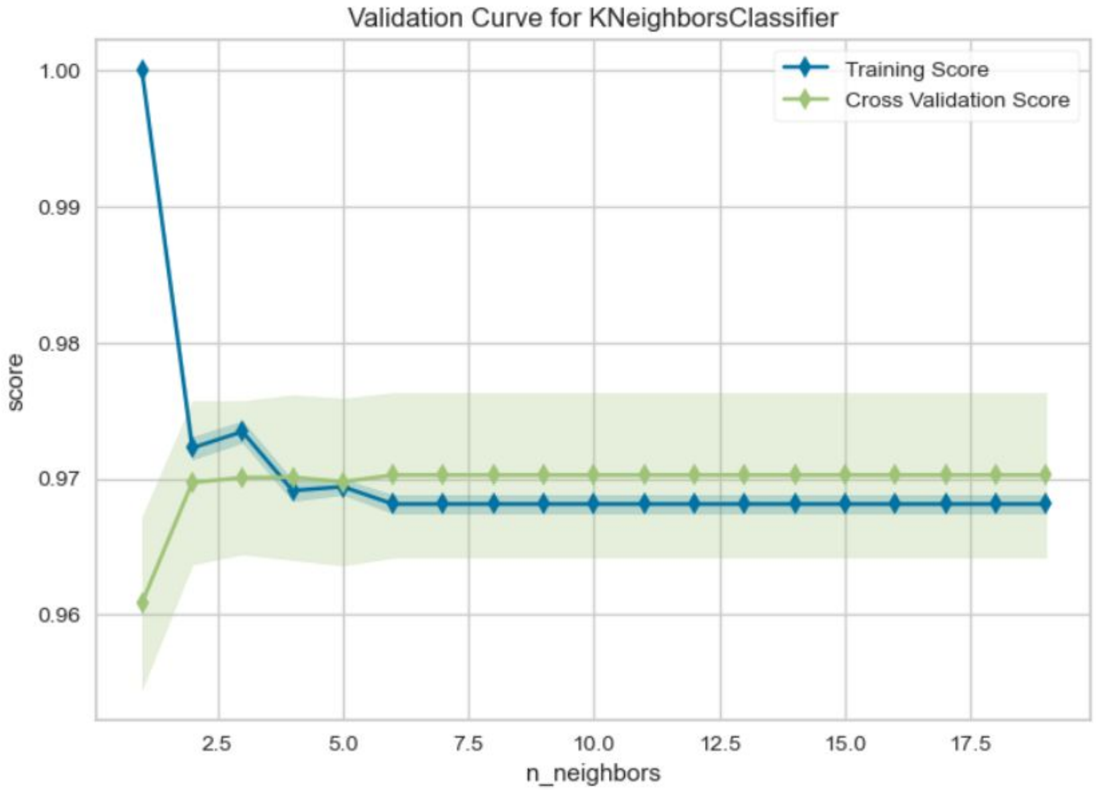
```
              precision    recall  f1-score   support

           0       0.59      0.46      0.52        93
           1       0.56      0.68      0.61        93

    accuracy                           0.57       186
   macro avg       0.57      0.57      0.56       186
weighted avg       0.57      0.57      0.56       186
```

43	50
30	63



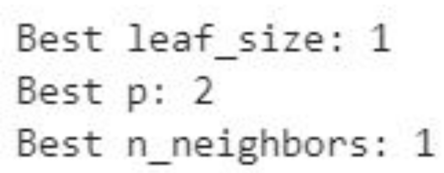
```
Precision para entrenamiento: 0.5698924731182796
Precision para validacion: 0.3988888888888889
```

Validation Curve

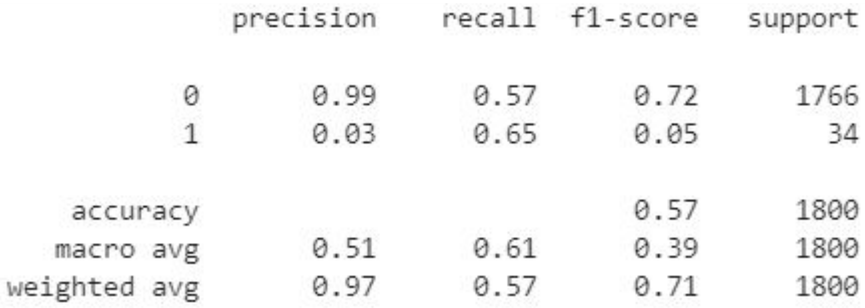


Óptimo → K = 3

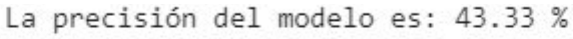
GridSearchCV



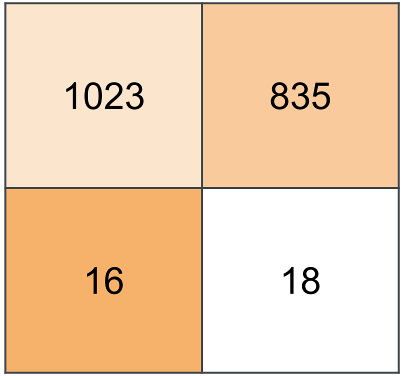
```
Best leaf_size: 1
Best p: 2
Best n_neighbors: 1
```



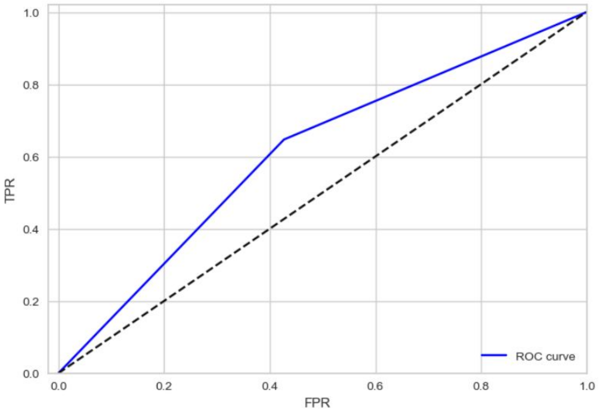
```
              precision    recall  f1-score   support

           0       0.99      0.57      0.72      1766
           1       0.03      0.65      0.05        34

    accuracy                           0.57      1800
   macro avg       0.51      0.61      0.39      1800
weighted avg       0.97      0.57      0.71      1800
```



```
La precisión del modelo es: 43.33 %
```



1023	835
16	18

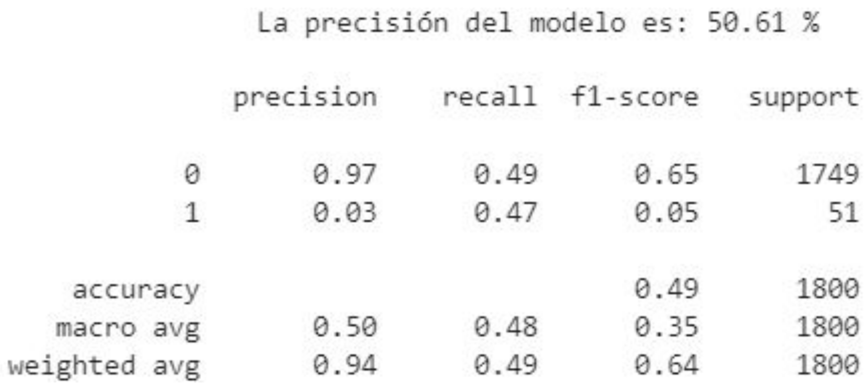


4. Support Vector Machine

SVC

Kernel = sigmoid
Kfold = 3
Probability =True

862	887
27	24

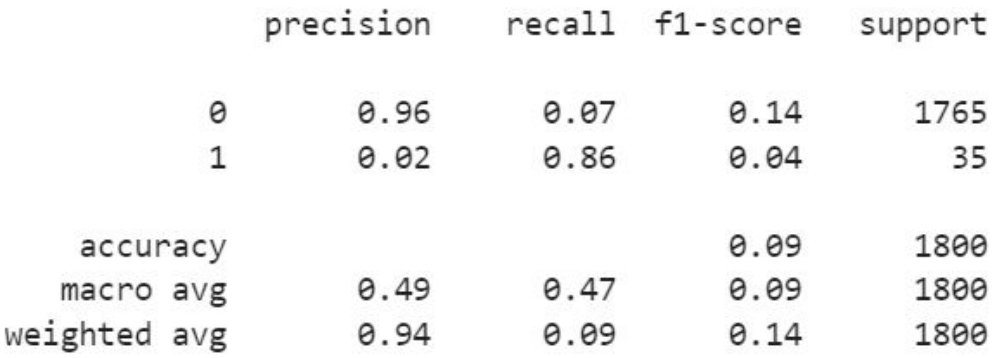


```
La precisión del modelo es: 50.61 %

              precision    recall  f1-score   support

           0       0.97      0.49      0.65      1749
           1       0.03      0.47      0.05        51

    accuracy                           0.49      1800
   macro avg       0.50      0.48      0.35      1800
weighted avg       0.94      0.49      0.64      1800
```

5. Redes neuronales

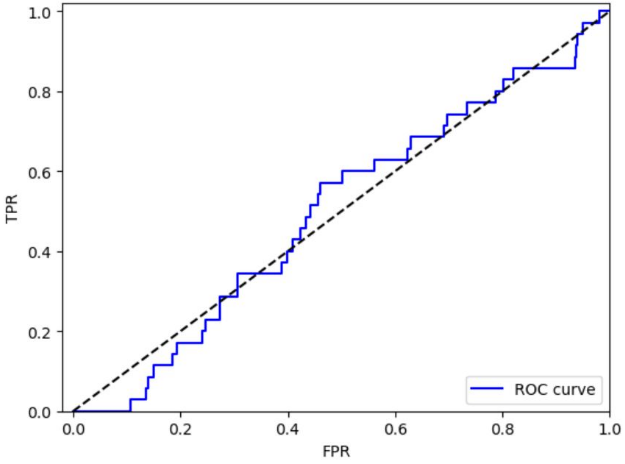
MLPClassifier

Kernel = sigmoid
Kfold = 3
Probability =True



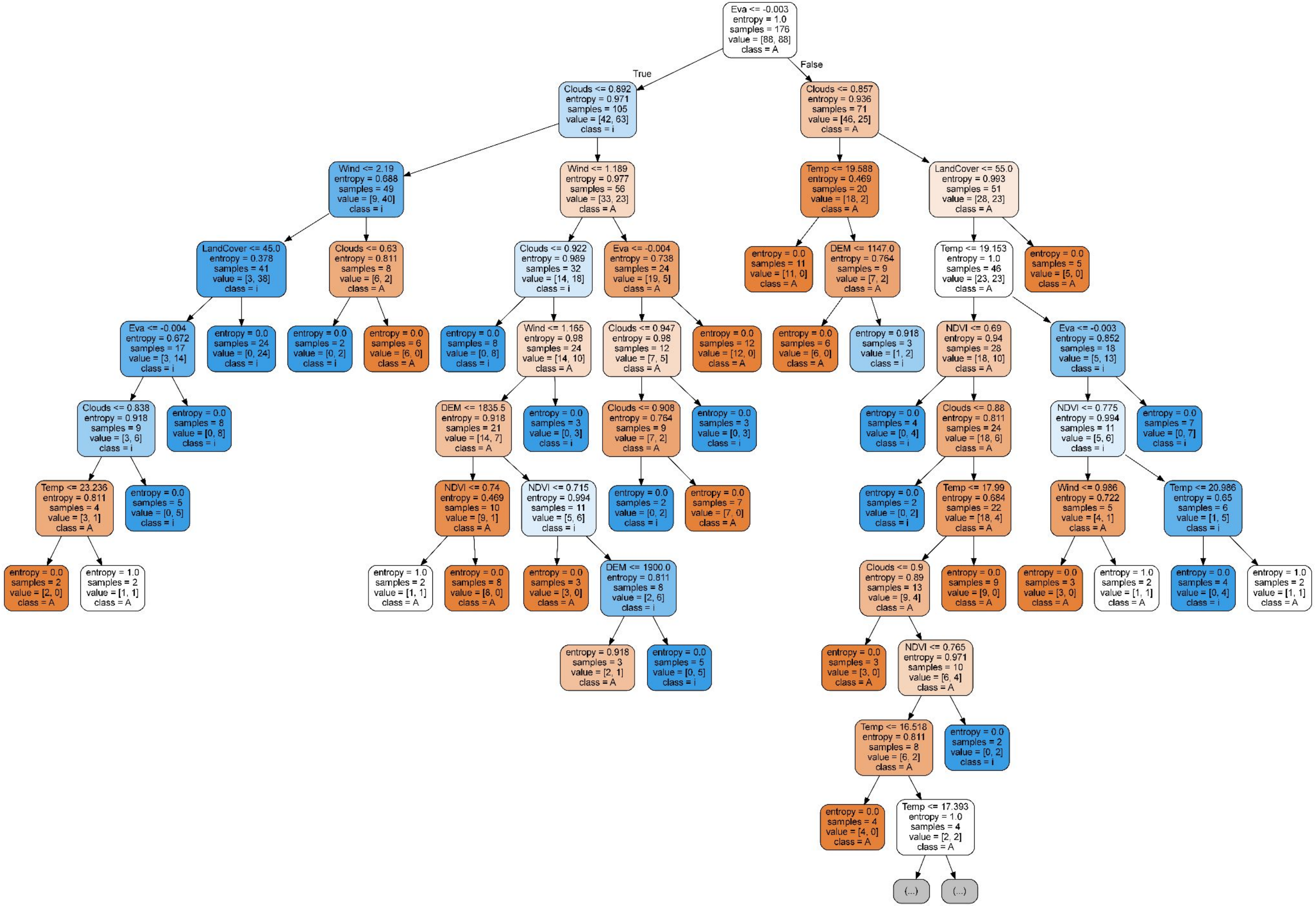
```
              precision    recall  f1-score   support

           0       0.96      0.07      0.14      1765
           1       0.02      0.86      0.04        35

    accuracy                           0.09      1800
   macro avg       0.49      0.47      0.09      1800
weighted avg       0.94      0.09      0.14      1800
```

134	1634
5	30

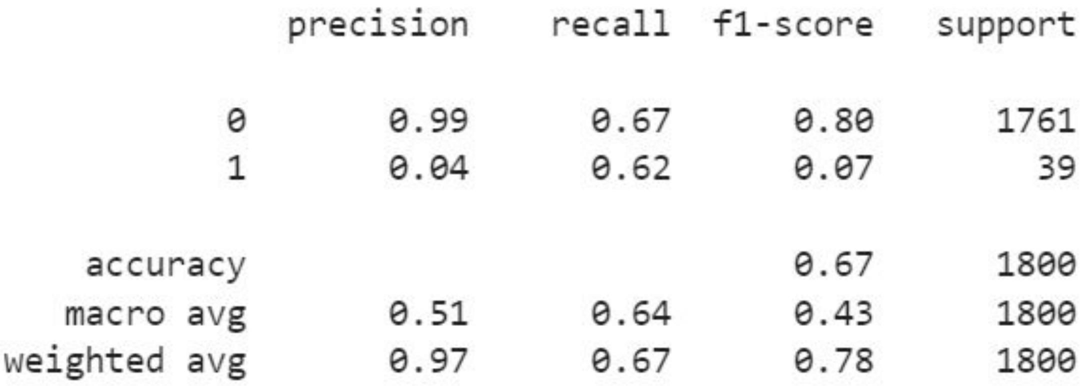


6. Ensambles

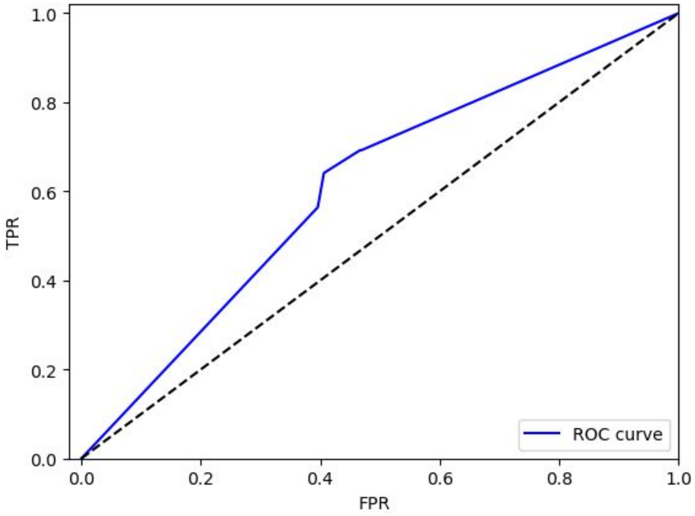
Decision Tree Classifier


Eva <= -0.003 entropy = 1.0 samples = 176 value = [88, 88] class = A
True
False
Clouds <= 0.892 entropy = 0.971 samples = 105 value = [42, 63] class = i
Clouds <= 0.857 entropy = 0.936 samples = 71 value = [46, 25] class = A
Wind <= 2.19 entropy = 0.688 samples = 49 value = [9, 40] class = i
Wind <= 1.189 entropy = 0.977 samples = 56 value = [33, 23] class = A
Temp <= 19.588 entropy = 0.469 samples = 20 value = [18, 2] class = A
LandCover <= 55.0 entropy = 0.993 samples = 51 value = [28, 23] class = A
LandCover <= 45.0 entropy = 0.378 samples = 41 value = [3, 38] class = i
Clouds <= 0.63 entropy = 0.811 samples = 8 value = [6, 2] class = A
Clouds <= 0.922 entropy = 0.989 samples = 32 value = [14, 18] class = i
Eva <= -0.004 entropy = 0.738 samples = 24 value = [19, 5] class = A
entropy = 0.0 samples = 11 value = [11, 0] class = A
DEM <= 1147.0 entropy = 0.764 samples = 9 value = [7, 2] class = A
Temp <= 19.153 entropy = 1.0 samples = 46 value = [23, 23] class = A
entropy = 0.0 samples = 5 value = [5, 0] class = A
Eva <= -0.004 entropy = 0.672 samples = 17 value = [3, 14] class = i
entropy = 0.0 samples = 24 value = [0, 24] class = i
entropy = 0.0 samples = 2 value = [0, 2] class = i
entropy = 0.0 samples = 6 value = [6, 0] class = A
entropy = 0.0 samples = 8 value = [0, 8] class = i
Wind <= 1.165 entropy = 0.98 samples = 24 value = [14, 10] class = A
Clouds <= 0.947 entropy = 0.98 samples = 12 value = [7, 5] class = A
entropy = 0.0 samples = 12 value = [12, 0] class = A
entropy = 0.0 samples = 6 value = [6, 0] class = A
entropy = 0.918 samples = 3 value = [1, 2] class = i
NDVI <= 0.69 entropy = 0.94 samples = 28 value = [18, 10] class = A
Eva <= -0.003 entropy = 0.852 samples = 18 value = [5, 13] class = i
Clouds <= 0.838 entropy = 0.918 samples = 9 value = [3, 6] class = i
entropy = 0.0 samples = 8 value = [0, 8] class = i
DEM <= 1835.5 entropy = 0.918 samples = 21 value = [14, 7] class = A
entropy = 0.0 samples = 3 value = [0, 3] class = i
Clouds <= 0.908 entropy = 0.764 samples = 9 value = [7, 2] class = A
entropy = 0.0 samples = 3 value = [0, 3] class = i
entropy = 0.0 samples = 4 value = [0, 4] class = i
Clouds <= 0.88 entropy = 0.811 samples = 24 value = [18, 6] class = A
NDVI <= 0.775 entropy = 0.994 samples = 11 value = [5, 6] class = i
entropy = 0.0 samples = 7 value = [0, 7] class = i
Temp <= 23.236 entropy = 0.811 samples = 4 value = [3, 1] class = A
entropy = 0.0 samples = 5 value = [0, 5] class = i
NDVI <= 0.74 entropy = 0.469 samples = 10 value = [9, 1] class = A
NDVI <= 0.715 entropy = 0.994 samples = 11 value = [5, 6] class = i
entropy = 0.0 samples = 2 value = [0, 2] class = i
entropy = 0.0 samples = 7 value = [7, 0] class = A
entropy = 0.0 samples = 2 value = [0, 2] class = i
Temp <= 17.99 entropy = 0.684 samples = 22 value = [18, 4] class = A
Wind <= 0.986 entropy = 0.722 samples = 5 value = [4, 1] class = A
Temp <= 20.986 entropy = 0.65 samples = 6 value = [1, 5] class = i
entropy = 0.0 samples = 2 value = [2, 0] class = A
entropy = 1.0 samples = 2 value = [1, 1] class = A
entropy = 1.0 samples = 2 value = [1, 1] class = A
entropy = 0.0 samples = 8 value = [8, 0] class = A
entropy = 0.0 samples = 3 value = [3, 0] class = A
DEM <= 1900.0 entropy = 0.811 samples = 8 value = [2, 6] class = i
Clouds <= 0.9 entropy = 0.89 samples = 13 value = [9, 4] class = A
entropy = 0.0 samples = 9 value = [9, 0] class = A
entropy = 0.0 samples = 3 value = [3, 0] class = A
entropy = 1.0 samples = 2 value = [1, 1] class = A
entropy = 0.0 samples = 4 value = [0, 4] class = i
entropy = 1.0 samples = 2 value = [1, 1] class = A
entropy = 0.918 samples = 3 value = [2, 1] class = A
entropy = 0.0 samples = 5 value = [0, 5] class = i
entropy = 0.0 samples = 3 value = [3, 0] class = A
NDVI <= 0.765 entropy = 0.971 samples = 10 value = [6, 4] class = A
Temp <= 16.518 entropy = 0.811 samples = 8 value = [6, 2] class = A
entropy = 0.0 samples = 2 value = [0, 2] class = i
entropy = 0.0 samples = 4 value = [4, 0] class = A
Temp <= 17.393 entropy = 1.0 samples = 4 value = [2, 2] class = A
(...)
(...)

Decision Tree Classifier

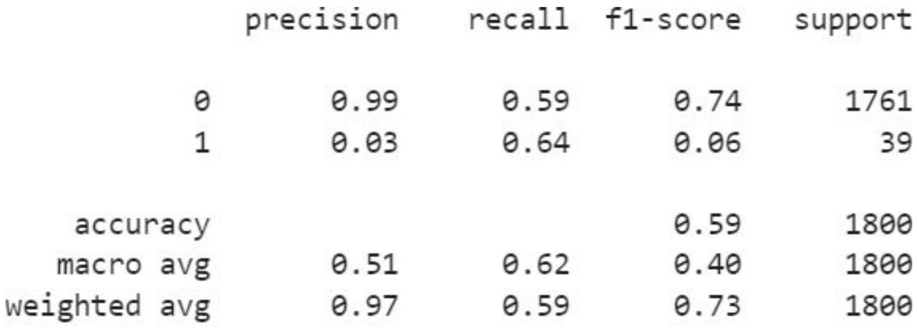


```
              precision    recall  f1-score   support

           0       0.99      0.67      0.80      1761
           1       0.04      0.62      0.07        39

    accuracy                           0.67      1800
   macro avg       0.51      0.64      0.43      1800
weighted avg       0.97      0.67      0.78      1800
```

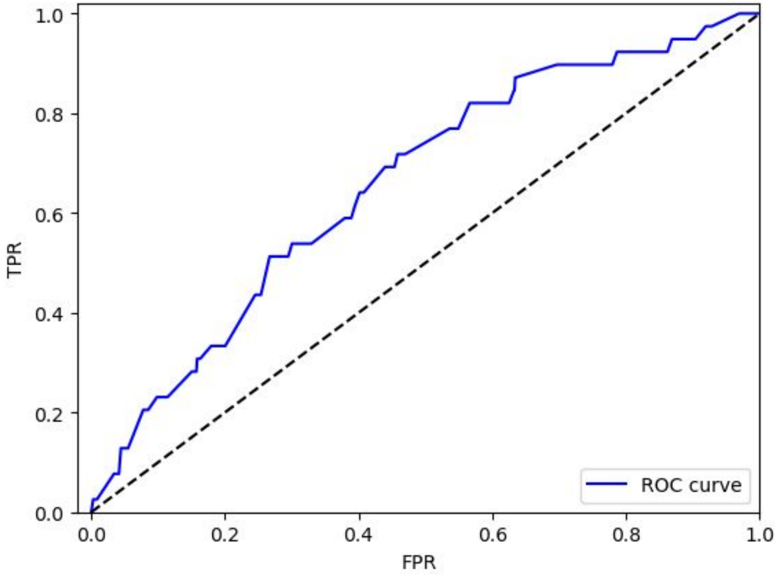


min_samples_split=2
min_samples_leaf=2

Bagging Classifier

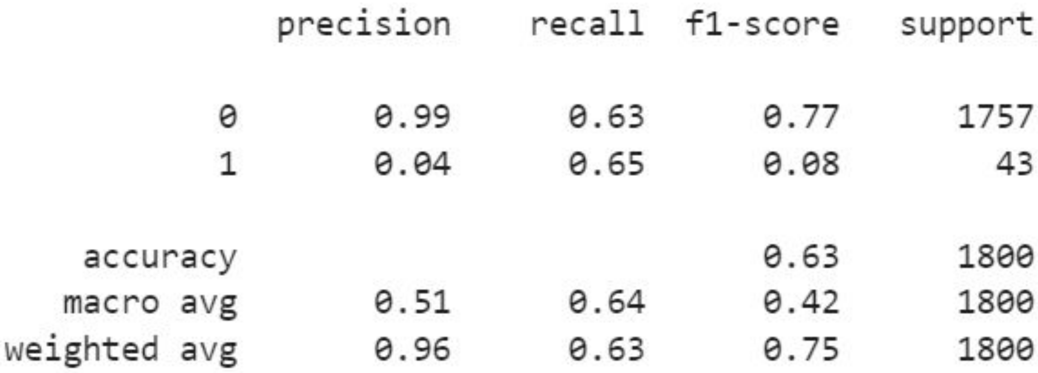


```
              precision    recall  f1-score   support

           0       0.99      0.59      0.74      1761
           1       0.03      0.64      0.06        39

    accuracy                           0.59      1800
   macro avg       0.51      0.62      0.40      1800
weighted avg       0.97      0.59      0.73      1800
```

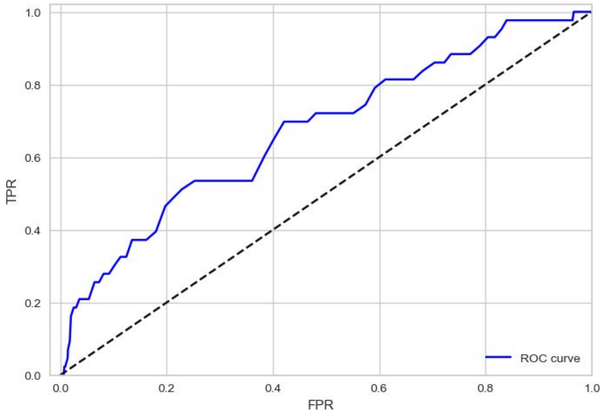


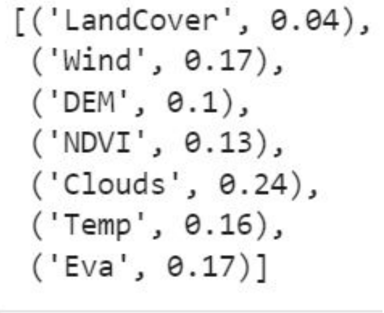
base_estimator=dtc
n_estimators=10

Random Forest Classifier



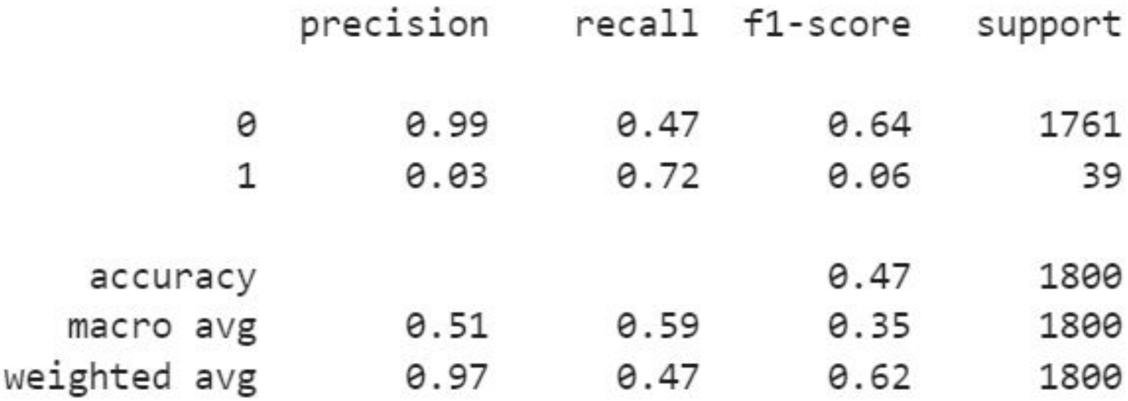
```
              precision    recall  f1-score   support

           0       0.99      0.63      0.77      1757
           1       0.04      0.65      0.08        43

    accuracy                           0.63      1800
   macro avg       0.51      0.64      0.42      1800
weighted avg       0.96      0.63      0.75      1800
```



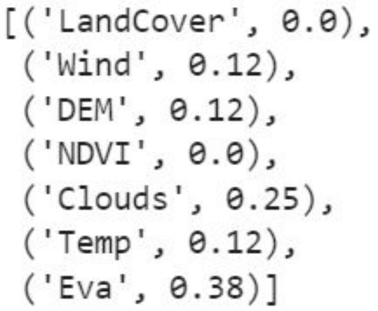


```
[('LandCover', 0.04),
 ('Wind', 0.17),
 ('DEM', 0.1),
 ('NDVI', 0.13),
 ('Clouds', 0.24),
 ('Temp', 0.16),
 ('Eva', 0.17)]
```

Ada Boost Classifier

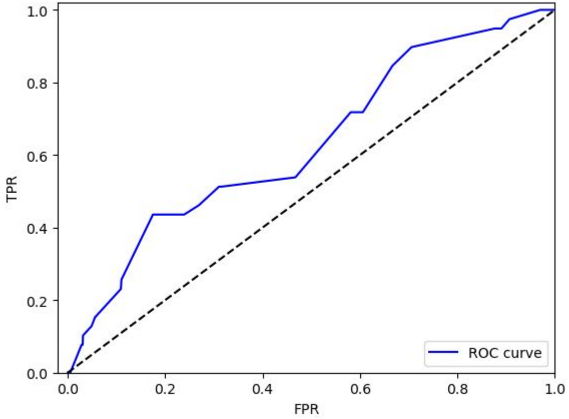


```
              precision    recall  f1-score   support

           0       0.99      0.47      0.64      1761
           1       0.03      0.72      0.06        39

    accuracy                           0.47      1800
   macro avg       0.51      0.59      0.35      1800
weighted avg       0.97      0.47      0.62      1800
```

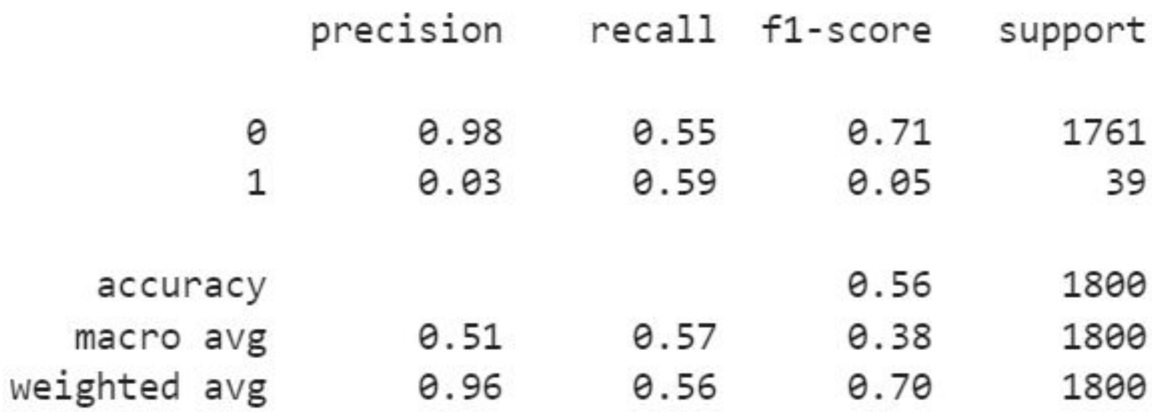


```
[('LandCover', 0.0),
 ('Wind', 0.12),
 ('DEM', 0.12),
 ('NDVI', 0.0),
 ('Clouds', 0.25),
 ('Temp', 0.12),
 ('Eva', 0.38)]
```

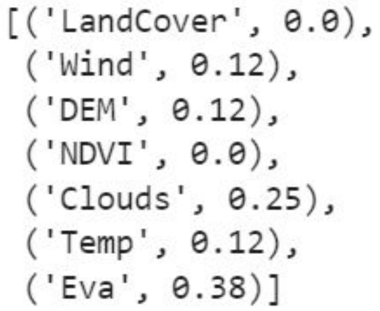
N_estimators = 8
random_state = 1
Score = 0.577



Gradient Boosting Classifier

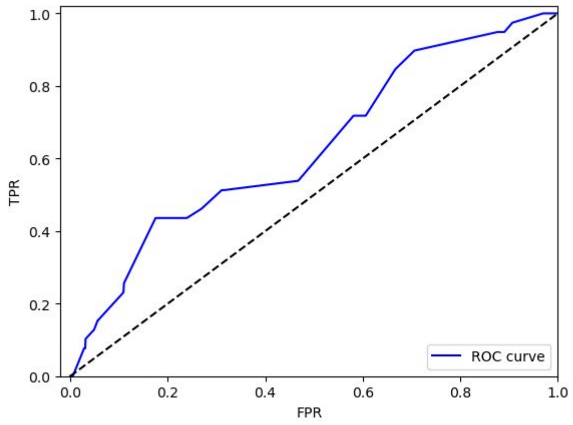


```
              precision    recall  f1-score   support

           0       0.98      0.55      0.71      1761
           1       0.03      0.59      0.05        39

    accuracy                           0.56      1800
   macro avg       0.51      0.57      0.38      1800
weighted avg       0.96      0.56      0.70      1800
```



```
[('LandCover', 0.0),
 ('Wind', 0.12),
 ('DEM', 0.12),
 ('NDVI', 0.0),
 ('Clouds', 0.25),
 ('Temp', 0.12),
 ('Eva', 0.38)]
```

N_estimators = 2
random_state = 1
Score = 0.55



Resumen

	Precision	Recall
Linear Discriminant Analysis - Kfold	0.03	0.53
Regresión Logística - LI	0.03	0.59
KNeighborsClassifier	0.56	0.68
SupportVectorMachine	0.03	0.47
MLPClassifier	0.02	0.86
Decision Tree Classifier	0.04	0.62
Bagging Classifier	0.03	0.64
Random Forest Classifier	0.04	0.65
Ada Boost Classifier	0.03	0.72
Gradient Boosting Classifier	0.03	0.59

Muchas Gracias