Trabalho_Final_ML_Cludia_Matheus

July 23, 2019

1 Trabalho prático: Machine Learning I

1.1 Claudia Oliveira; Matheus Reck de Oliveira

Neste trabalho, avaliamos um dataset com classificações de tipos predominantes de área florestal do **Roosevelt National Forest**, no norte do Colorado, nos Estados Unidos. Apresentamos uma análise exploratória do dataset, e avaliamos algoritmos para predição de cobertura de área florestal de segmentos de 30m x 30m da área do parque.

```
In [1]: import pandas as pd
        import matplotlib.pyplot as plt
        from sklearn.model_selection import cross_val_score
        pd.set_option('display.max_columns', 100)
```

1.2 Análise Exploratória

1.2.1 Amostra dos Dados

```
In [2]: dataset = pd.read_csv('C:/dev/PosUniritter/MachineLearning/Uniritter-ML_1_TrabalhoPrati

        # seta as variaveis que possuem os dados do dataset para o treinamento de todos os mod
        X = dataset.drop(['Cover_Type'], axis=1)
        y = dataset['Cover_Type']

        # Início da Analise exploratória do dataset
        print(dataset.head())
```

```
   Elevation  Aspect  Slope  Horizontal_Distance_To_Hydrology  \
0       2596      51      3                               258
1       2804     139      9                               268
2       2785     155     18                               242
3       2595      45      2                               153
4       2579     132      6                               300

   Vertical_Distance_To_Hydrology  Horizontal_Distance_To_Roadways  \
0                               0                              510
1                              65                             3180
2                             118                             3090
```

```
3                            -1                             391
4                           -15                              67

   Hillshade_9am  Hillshade_Noon  Hillshade_3pm  \
0            221             232            148
1            234             238            135
2            238             238            122
3            220             234            150
4            230             237            140

   Horizontal_Distance_To_Fire_Points  Wilderness_Area1  Wilderness_Area2  \
0                                6279                 1                 0
1                                6121                 1                 0
2                                6211                 1                 0
3                                6172                 1                 0
4                                6031                 1                 0

   Wilderness_Area3  Wilderness_Area4  Soil_Type1  Soil_Type2  Soil_Type3  \
0                 0                 0           0           0           0
1                 0                 0           0           0           0
2                 0                 0           0           0           0
3                 0                 0           0           0           0
4                 0                 0           0           0           0

   Soil_Type4  Soil_Type5  Soil_Type6  Soil_Type7  Soil_Type8  Soil_Type9  \
0           0           0           0           0           0           0
1           0           0           0           0           0           0
2           0           0           0           0           0           0
3           0           0           0           0           0           0
4           0           0           0           0           0           0

   Soil_Type10  Soil_Type11  Soil_Type12  Soil_Type13  Soil_Type14  \
0            0            0            0            0            0
1            0            0            1            0            0
2            0            0            0            0            0
3            0            0            0            0            0
4            0            0            0            0            0

   Soil_Type15  Soil_Type16  Soil_Type17  Soil_Type18  Soil_Type19  \
0            0            0            0            0            0
1            0            0            0            0            0
2            0            0            0            0            0
3            0            0            0            0            0
4            0            0            0            0            0

   Soil_Type20  Soil_Type21  Soil_Type22  Soil_Type23  Soil_Type24  \
0            0            0            0            0            0
1            0            0            0            0            0
```

```
2              0              0              0              0              0
3              0              0              0              0              0
4              0              0              0              0              0

   Soil_Type25  Soil_Type26  Soil_Type27  Soil_Type28  Soil_Type29  \
0            0            0            0            0            1
1            0            0            0            0            0
2            0            0            0            0            0
3            0            0            0            0            1
4            0            0            0            0            1

   Soil_Type30  Soil_Type31  Soil_Type32  Soil_Type33  Soil_Type34  \
0            0            0            0            0            0
1            0            0            0            0            0
2            1            0            0            0            0
3            0            0            0            0            0
4            0            0            0            0            0

   Soil_Type35  Soil_Type36  Soil_Type37  Soil_Type38  Soil_Type39  \
0            0            0            0            0            0
1            0            0            0            0            0
2            0            0            0            0            0
3            0            0            0            0            0
4            0            0            0            0            0

   Soil_Type40  Cover_Type
0            0           5
1            0           2
2            0           2
3            0           5
4            0           2
```

1.2.2 Informações Estatísticas

```
In [3]: shape = dataset.shape
        print("Total de Linhas e Colunas: ", shape)
        sumario = dataset.describe()
        print("Informações estatisticas do Dataset: ")
        print(sumario)
```

```
Total de Linhas e Colunas:  (581002, 55)
Informações estatisticas do Dataset:
           Elevation         Aspect          Slope  \
count  581002.000000  581002.000000  581002.000000
mean     2959.371136     155.657158      14.103702
std       279.980764     111.913616       7.488241
min      1859.000000       0.000000       0.000000
```

```
25%      2809.000000      58.000000       9.000000
50%      2996.000000     127.000000      13.000000
75%      3163.000000     260.000000      18.000000
max      3858.000000     360.000000      66.000000

       Horizontal_Distance_To_Hydrology  Vertical_Distance_To_Hydrology  \
count                     581002.000000                   581002.000000
mean                         269.429880                       46.419222
std                          212.549971                       58.295524
min                            0.000000                     -173.000000
25%                          108.000000                        7.000000
50%                          218.000000                       30.000000
75%                          384.000000                       69.000000
max                         1397.000000                      601.000000

       Horizontal_Distance_To_Roadways  Hillshade_9am  Hillshade_Noon  \
count                    581002.000000  581002.000000   581002.000000
mean                       2350.165252     212.145838      223.318806
std                        1559.257261      26.769947       19.768789
min                           0.000000       0.000000        0.000000
25%                        1106.000000     198.000000      213.000000
50%                        1997.000000     218.000000      226.000000
75%                        3328.000000     231.000000      237.000000
max                        7117.000000     254.000000      254.000000

       Hillshade_3pm  Horizontal_Distance_To_Fire_Points  Wilderness_Area1  \
count  581002.000000                       581002.000000     581002.000000
mean      142.528609                         1980.292908          0.448869
std        38.274526                         1324.176031          0.497379
min         0.000000                            0.000000          0.000000
25%       119.000000                         1024.000000          0.000000
50%       143.000000                         1710.000000          0.000000
75%       168.000000                         2550.000000          1.000000
max       254.000000                         7173.000000          1.000000

       Wilderness_Area2  Wilderness_Area3  Wilderness_Area4     Soil_Type1  \
count     581002.000000     581002.000000     581002.000000  581002.000000
mean           0.051435          0.436074          0.063621       0.005217
std            0.220884          0.495897          0.244077       0.072039
min            0.000000          0.000000          0.000000       0.000000
25%            0.000000          0.000000          0.000000       0.000000
50%            0.000000          0.000000          0.000000       0.000000
75%            0.000000          1.000000          0.000000       0.000000
max            1.000000          1.000000          1.000000       1.000000

          Soil_Type2     Soil_Type3     Soil_Type4     Soil_Type5  \
count  581002.000000  581002.000000  581002.000000  581002.000000
mean        0.012952       0.008298       0.021336       0.002749
```

```
std         0.113067       0.090713       0.144500       0.052356
min         0.000000       0.000000       0.000000       0.000000
25%         0.000000       0.000000       0.000000       0.000000
50%         0.000000       0.000000       0.000000       0.000000
75%         0.000000       0.000000       0.000000       0.000000
max         1.000000       1.000000       1.000000       1.000000

          Soil_Type6     Soil_Type7     Soil_Type8     Soil_Type9  \
count  581002.000000  581002.000000  581002.000000  581002.000000
mean        0.011315       0.000181       0.000308       0.001974
std         0.105768       0.013442       0.017550       0.044388
min         0.000000       0.000000       0.000000       0.000000
25%         0.000000       0.000000       0.000000       0.000000
50%         0.000000       0.000000       0.000000       0.000000
75%         0.000000       0.000000       0.000000       0.000000
max         1.000000       1.000000       1.000000       1.000000

         Soil_Type10    Soil_Type11    Soil_Type12    Soil_Type13  \
count  581002.000000  581002.000000  581002.000000  581002.000000
mean        0.056168       0.021356       0.051585       0.030002
std         0.230247       0.144569       0.221188       0.170592
min         0.000000       0.000000       0.000000       0.000000
25%         0.000000       0.000000       0.000000       0.000000
50%         0.000000       0.000000       0.000000       0.000000
75%         0.000000       0.000000       0.000000       0.000000
max         1.000000       1.000000       1.000000       1.000000

         Soil_Type14    Soil_Type15    Soil_Type16    Soil_Type17  \
count  581002.000000  581002.000000  581002.000000  581002.000000
mean        0.001029       0.000005       0.004897       0.005890
std         0.032066       0.002272       0.069805       0.076519
min         0.000000       0.000000       0.000000       0.000000
25%         0.000000       0.000000       0.000000       0.000000
50%         0.000000       0.000000       0.000000       0.000000
75%         0.000000       0.000000       0.000000       0.000000
max         1.000000       1.000000       1.000000       1.000000

         Soil_Type18    Soil_Type19    Soil_Type20    Soil_Type21  \
count  581002.000000  581002.000000  581002.000000  581002.000000
mean        0.003268       0.006921       0.015936       0.001442
std         0.057077       0.082903       0.125229       0.037951
min         0.000000       0.000000       0.000000       0.000000
25%         0.000000       0.000000       0.000000       0.000000
50%         0.000000       0.000000       0.000000       0.000000
75%         0.000000       0.000000       0.000000       0.000000
max         1.000000       1.000000       1.000000       1.000000

         Soil_Type22    Soil_Type23    Soil_Type24    Soil_Type25  \
```

count	581002.000000	581002.000000	581002.000000	581002.000000
mean	0.057440	0.099401	0.036623	0.000816
std	0.232682	0.299200	0.187834	0.028551
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000

	Soil_Type26	Soil_Type27	Soil_Type28	Soil_Type29 \
count	581002.000000	581002.000000	581002.000000	581002.000000
mean	0.004456	0.001869	0.001628	0.198356
std	0.066605	0.043194	0.040318	0.398762
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000

	Soil_Type30	Soil_Type31	Soil_Type32	Soil_Type33 \
count	581002.000000	581002.000000	581002.000000	581002.000000
mean	0.051928	0.044175	0.090394	0.077716
std	0.221881	0.205485	0.286745	0.267724
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000

	Soil_Type34	Soil_Type35	Soil_Type36	Soil_Type37 \
count	581002.000000	581002.000000	581002.000000	581002.000000
mean	0.002773	0.003255	0.000205	0.000513
std	0.052584	0.056957	0.014310	0.022642
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000

	Soil_Type38	Soil_Type39	Soil_Type40	Cover_Type
count	581002.000000	581002.000000	581002.000000	581002.000000
mean	0.026802	0.023762	0.015060	2.051444
std	0.161504	0.152308	0.121792	1.396483
min	0.000000	0.000000	0.000000	1.000000
25%	0.000000	0.000000	0.000000	1.000000
50%	0.000000	0.000000	0.000000	2.000000
75%	0.000000	0.000000	0.000000	2.000000
max	1.000000	1.000000	1.000000	7.000000

Temos uma amostra de 581.002 linhas. Pelo somatório acima, podemos observar que não há nenhuma feature com valores faltantes.

1.2.3 Distribuição de classes

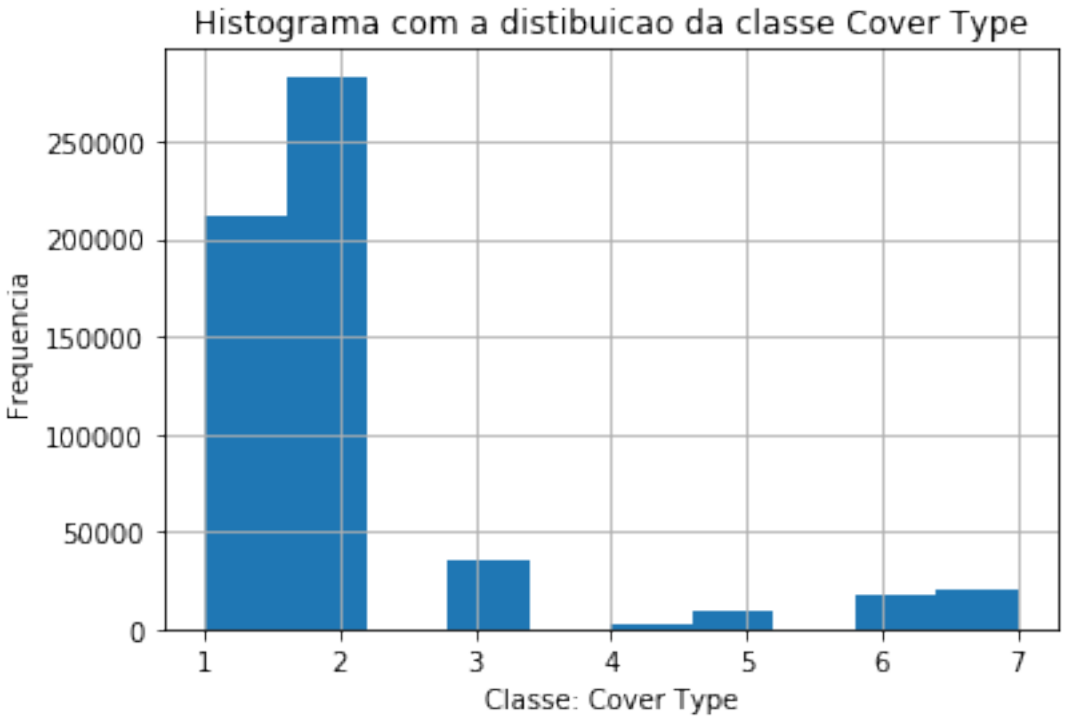
Distribuição da classe **Cover Type** Existem sete tipos classes diferentes de Cover Type (área de cobertura):

1. Spruce/Fir
2. Lodgepole Pine
3. Ponderosa Pine
4. Cottonwood/Willow
5. Aspen
6. Douglas-fir
7. Krummholz

```
In [4]: print("Histograma com a distibuicao da classe Cover Type")
        dataset.Cover_Type.hist()
        plt.title('Histograma com a distibuicao da classe Cover Type')
        plt.xlabel('Classe: Cover Type')
        plt.ylabel('Frequencia')
```

```
Histograma com a distibuicao da classe Cover Type
```

```
Out[4]: Text(0, 0.5, 'Frequencia')
```



Pelo histograma acima, percebemos que no dataset utilizado para treino, há uma concentração muito maior de instâncias nas duas primeiras classes de Cover Type. Isto significa que a amostra não está balanceada.

1.3 Algorítmos de classificação

Comparamos o desempenho de três algoritmos diferentes: - Knn - Naive Bayes - Decision Trees

Para cada um deles, avaliamos a acurácia com através de Cross Validation. Avaliamos também, quando possível, a busca dos melhores hiper parâmetros através do Grid Search.

1.4 Knn

1.4.1 Treino

O treino foi realizado com os parâmetros default.

```
In [5]: from sklearn.neighbors import KNeighborsClassifier

        KNN_classifier = KNeighborsClassifier()
        KNN_classifier.fit(X, y)
```

```
Out[5]: KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                             metric_params=None, n_jobs=None, n_neighbors=5, p=2,
                             weights='uniform')
```

1.4.2 Cross Validation

```
In [6]: knn_cv_score_AC = cross_val_score(KNN_classifier, X, y, cv=10, scoring='accuracy')
        print("Acuracia KNN Com CV:", knn_cv_score_AC.mean())
```

```
Acuracia KNN Com CV: 0.5414039569515248
```

1.4.3 Grid Search

```
In [7]: from sklearn.model_selection import GridSearchCV

        grid_params = {'n_neighbors': [3,5,11,19],
                       'weights': ['uniform','distance'],
                       'metric': ['euclidean', 'manhattan']
                       }
        gs = GridSearchCV(KNeighborsClassifier(),
                          grid_params,
                          verbose=1,
                          cv=10,
                          n_jobs=-1
                          )
```

```
        gs_results = gs.fit(X,y)
        #gs_results_estimator = gs_results.best_estimator_
        gs_results_params = gs_results.best_params_


        #print("Best KNN Estimator: {}".format(gs_results_estimator))
        print("Melhores parâmetros para o Knn: {}".format(gs_results_params))
        print("Melhor score do Knn {}".format(gs_results.best_score_))
```

```
Fitting 10 folds for each of 16 candidates, totalling 160 fits


[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=-1)]: Done  42 tasks      | elapsed: 14.7min
[Parallel(n_jobs=-1)]: Done 160 out of 160 | elapsed: 99.6min finished


Melhores parâmetros para o Knn: {'metric': 'euclidean', 'n_neighbors': 3, 'weights': 'distance
Melhor score do Knn 0.5723043982636893
```

1.5 Naive Bayes

Não encontramos hiper parâmetros relevantes para tunagem, por isto, optamos por não realizar o Grid Search.

1.5.1 Treino

```
In [9]: from sklearn.naive_bayes import GaussianNB

        naive = GaussianNB()
        naive.fit(X, y)
```

```
Out[9]: GaussianNB(priors=None, var_smoothing=1e-09)
```

1.5.2 Cross Validation

```
In [11]: naive_cv_score_AC = cross_val_score(naive, X, y, cv=10, scoring='accuracy')
         print("Acuracia Naive Com CV:", naive_cv_score_AC.mean())
```

```
Acuracia Naive Com CV: 0.4420708082441541
```

1.6 Decision Trees

1.6.1 Treino

```
In [12]: from sklearn.tree import DecisionTreeClassifier
```

```
Tree_classifier = DecisionTreeClassifier()
Tree_classifier.fit(X, y)
```

```
Out[12]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                     max_features=None, max_leaf_nodes=None,
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=1, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                     splitter='best')
```

1.6.2 Cross Validation

```
In [13]: scores = cross_val_score(Tree_classifier, X, y, cv=10, scoring='accuracy')
         print("Acuracia Decision Com CV:",scores.mean())
```

```
Acuracia Decision Com CV: 0.5909442027365903
```

1.6.3 Grid Search

```
In [14]: # GRID SEARCH PARA O DECISION TREE

         grid_params_tree = {"max_depth": [3, 5, 10],
                             "max_features": [ 3, 5, 10, 15, 20],
                             "min_samples_leaf": [1, 3],
                             "criterion": ["gini", "entropy"]
                             }

         # Instantiating Decision Tree classifier
         tree = DecisionTreeClassifier()

         tree_cv = GridSearchCV(tree, grid_params_tree, cv = 10, n_jobs=-1)

         tree_cv.fit(X, y)

         # Print the tuned parameters and score
         print("Melhores parâmetros para a Decision Tree: {}".format(tree_cv.best_params_))
         print("Melhor score para a Decision Tree {}".format(tree_cv.best_score_))
```

```
Melhores parâmetros para a Decision Tree: {'criterion': 'gini', 'max_depth': 5, 'max_features'
Melhor score para a Decision Tree 0.6597516015435403
```

1.7 Algorítmo com melhor acurácia

Para os três algorítmos, levando em conta os resultados do **Cross Validation** e também da tunagem utilizando **Grid Search**, os melhores scores de acurácia são os seguintes: - Knn: 57% - Naive Bayes: 44% - Decision Trees: 65%

O melhor algorítmo é o **Decision Trees**

1.8 Aplicando os melhores parâmetros

```
In [12]: best_tree_classifier = DecisionTreeClassifier(criterion='gini', max_depth=5, max_featu
         best_tree_classifier.fit(X, y)
```

```
Out[12]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=5,
                     max_features=20, max_leaf_nodes=None,
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=3, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                     splitter='best')
```

1.8.1 Cross Validation dos melhores parâmetros

```
In [13]: best_tree_scores = cross_val_score(best_tree_classifier, X, y, cv=10, scoring='accura
         print("Acuracia Decision Com CV:",best_tree_scores.mean())
```

```
Acuracia Decision Com CV: 0.65114109489697
```

1.8.2 Exportando o modelo

```
In [14]: from joblib import dump

         dump(best_tree_classifier, 'model.joblib')
```

```
Out[14]: ['model.joblib']
```

1.9 Predição de instâncias

```
In [21]: to_predict = pd.read_csv('C:/dev/PosUniritter/MachineLearning/Uniritter-ML_1_Trabalho
         pred_cols = list(to_predict.columns.values)

         pred = pd.Series(best_tree_classifier.predict(to_predict[pred_cols]))
         print(pred)
```

```
0    2
1    2
2    1
3    2
4    2
5    2
6    2
7    3
8    4
9    3
dtype: int64
```