Challenge Report

Angel Salazar - 513236

Master in Applied Artificial Intelligence

TC-4029 - Analytics and Data Science

Data Cleansing, Visualization, Clustering, Classification

Intro

The challenges were designed to put into practice the topics covered in the subject during this academic period.

They are a great example of what a data scientist must face when working with actual data.

This report aims to provide a summary of the challenges and outcomes after applying data scientist and machine learning techniques to a given dataset.

Motivation

Learn and practice the following data scientist and machine learning techniques:

- Data cleansing and data normalization
- Unsupervised learning using a clustering technique (KMeans)
- Supervised learning using multiclass classification (Random Forest / Decision trees)

Research Statements

Find out whether the water quality can be related to its source location

Evaluate and select a multiclass classification model

Data

- Mexico
- Records related to underground water quality
- 1,068 water sources

Methodology

1

Data cleansing and analytics

Deal with missing data, normalize values, visualize outliers and mean

2

Clustering

Compute optimal number of clusters, compare clustering result to sempharo coloring result

3

Model Training

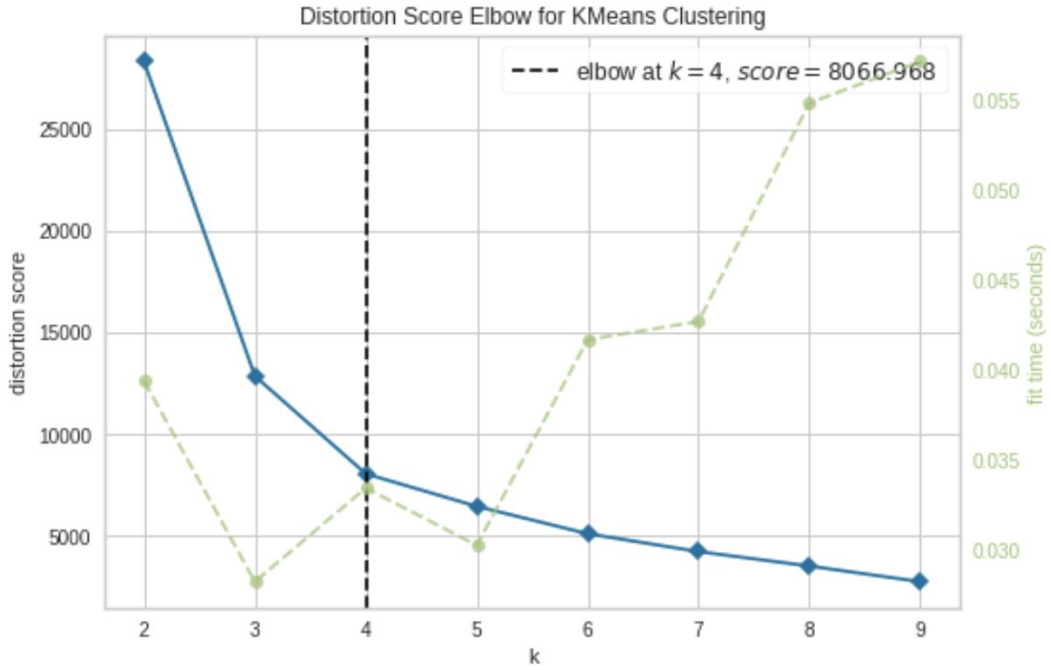
Compute feature importance, match up classification models, select optimal model

4

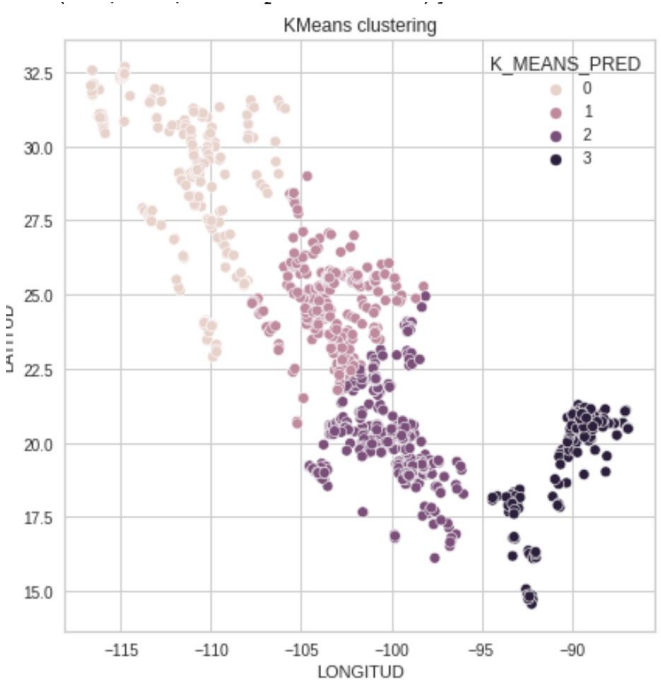
Classify data

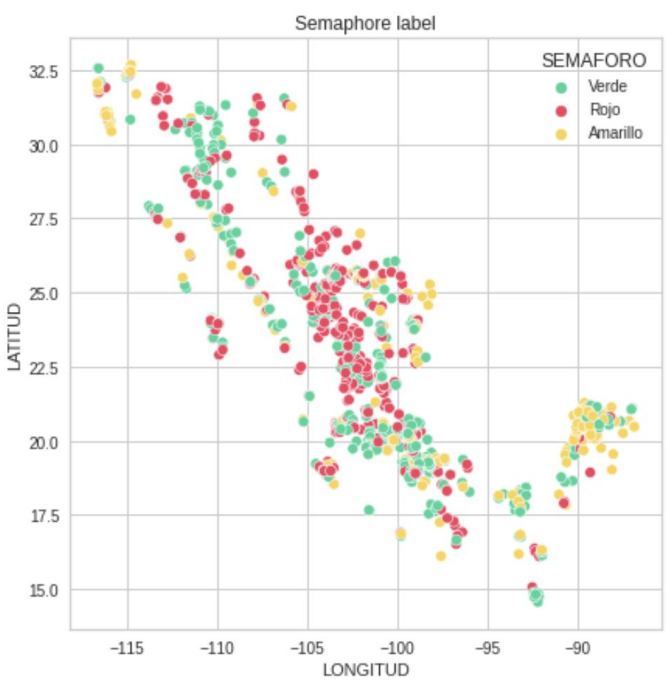
Predict water quality based on selected features, compute models statistics

Outcomes

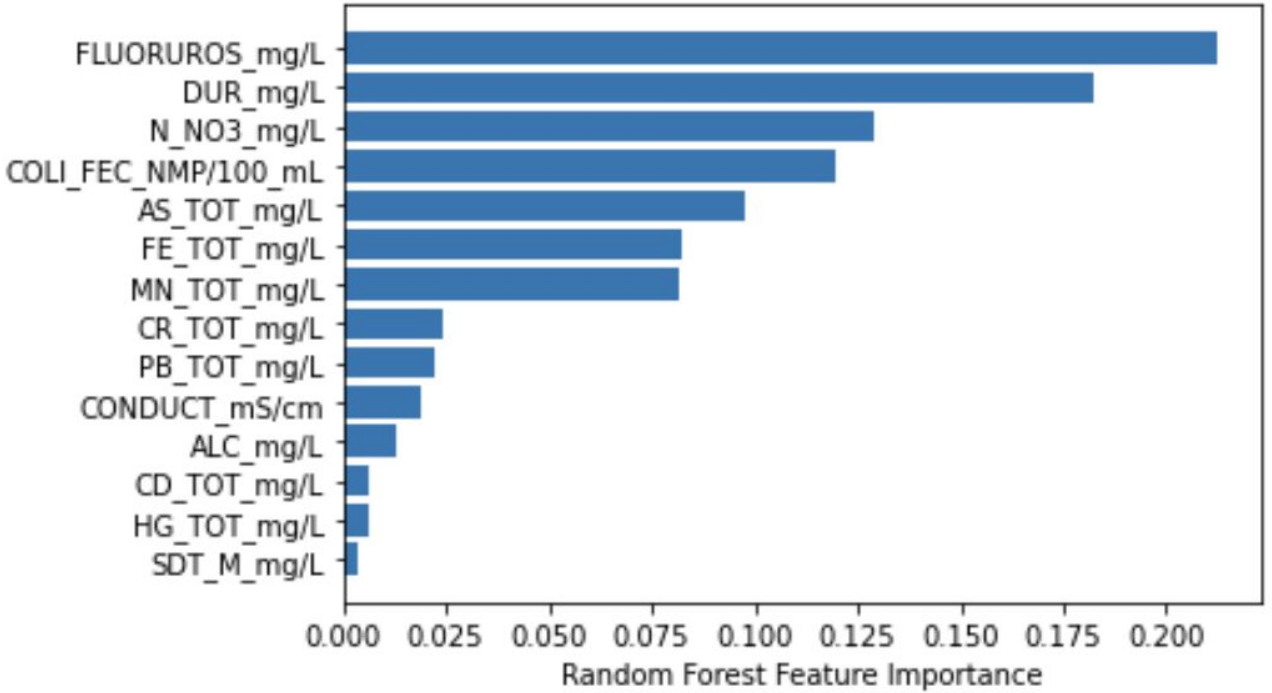

Distortion Score Elbow for KMeans Clustering
elbow at $k = 4$, $score = 8066.968$
distortion score
fit time (seconds)
k
25000
20000
15000
10000
5000
0.055
0.050
0.045
0.040
0.035
0.030
2
3
4
5
6
7
8
9

Outcomes


KMeans clustering
K_MEANS_PRED
0
1
2
3
32.5
30.0
27.5
25.0
22.5
20.0
17.5
15.0
−115
−110
−105
−100
−95
−90
LONGITUD


Semaphore label
SEMAFORO
Verde
Rojo
Amarillo
32.5
30.0
27.5
25.0
22.5
20.0
17.5
15.0
LATITUD
−115
−110
−105
−100
−95
−90
LONGITUD

Outcomes


FLUORUROS_mg/L
DUR_mg/L
N_NO3_mg/L
COLI_FEC_NMP/100_mL
AS_TOT_mg/L
FE_TOT_mg/L
MN_TOT_mg/L
CR_TOT_mg/L
PB_TOT_mg/L
CONDUCT_mS/cm
ALC_mg/L
CD_TOT_mg/L
HG_TOT_mg/L
SDT_M_mg/L
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Random Forest Feature Importance

Outcomes

RandomForestClassifier
Regular training

	precision	recall	f1-score	support
Verde	0.78	0.92	0.85	87
Amarillo	0.91	0.58	0.71	55
Rojo	0.98	0.94	0.96	69
micro avg	0.87	0.84	0.86	211
macro avg	0.89	0.81	0.84	211
weighted avg	0.88	0.84	0.85	211
samples avg	0.88	0.84	0.84	211

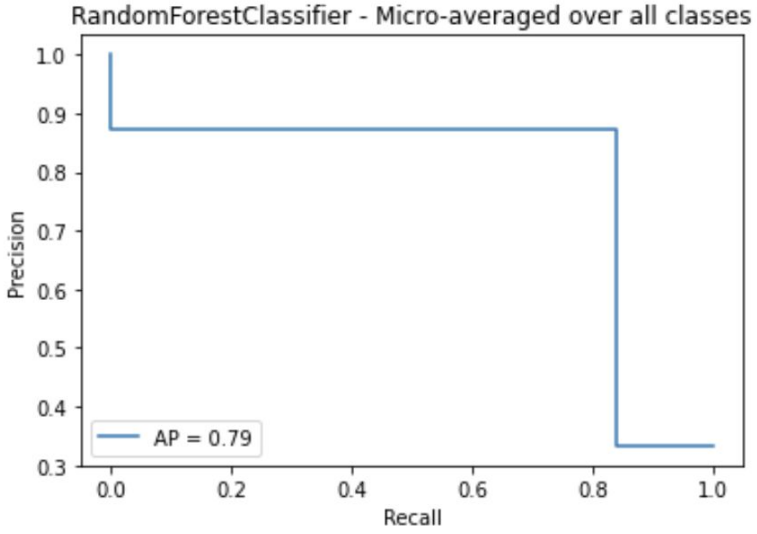
Cross validation training
accuracy 0.863582443653618

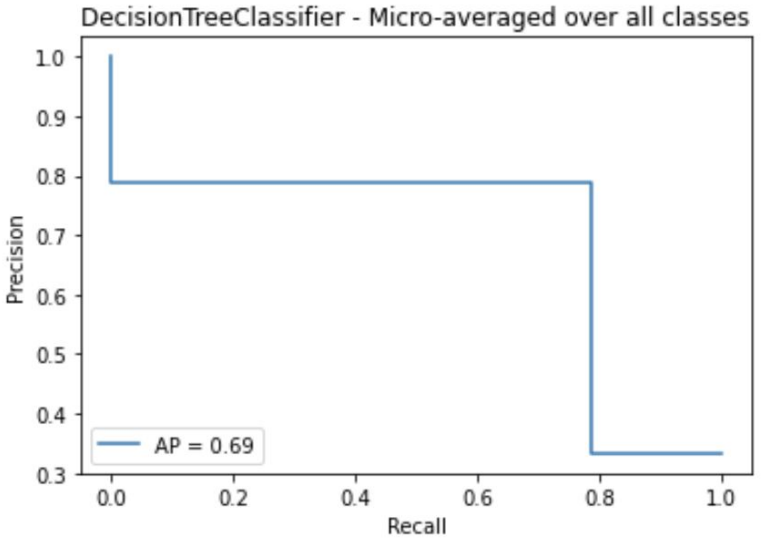
DecisionTreeClassifier
Regular training

	precision	recall	f1-score	support
Verde	0.77	0.77	0.77	87
Amarillo	0.69	0.64	0.66	55
Rojo	0.88	0.93	0.90	69
micro avg	0.79	0.79	0.79	211
macro avg	0.78	0.78	0.78	211
weighted avg	0.78	0.79	0.78	211
samples avg	0.79	0.79	0.79	211

Cross validation training
accuracy 0.8315539739027283

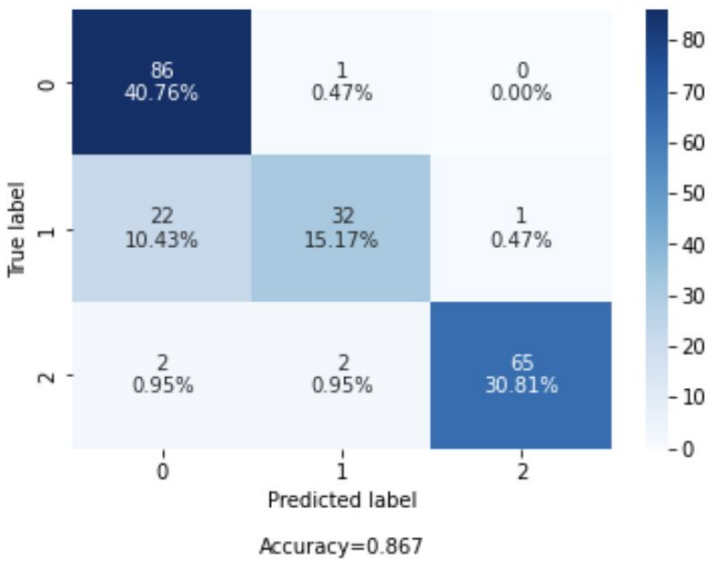
Outcomes


RandomForestClassifier - Micro-averaged over all classes
Precision
Recall
AP = 0.79
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

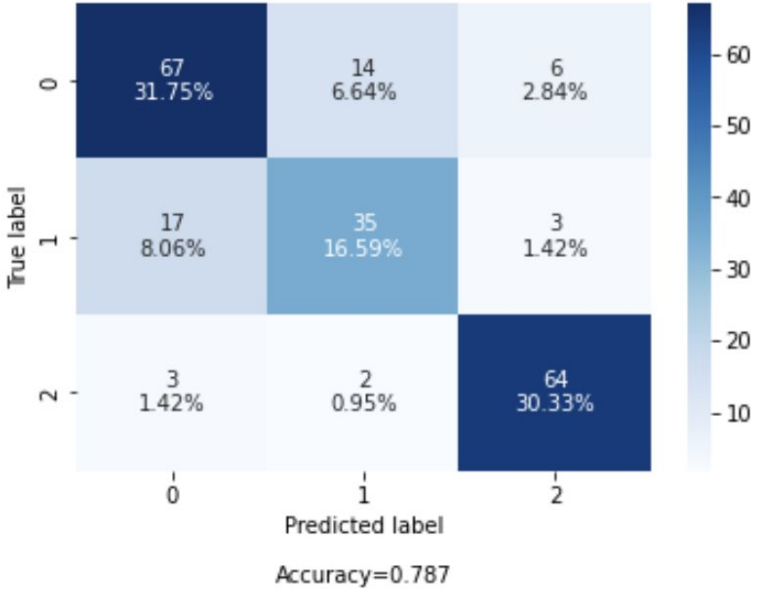

DecisionTreeClassifier - Micro-averaged over all classes
Precision
Recall
AP = 0.69
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

Outcomes

Random Forest Classifier



Decision Tree Classifier



Challenges

- Data cleansing and normalization
- Plotting illustration

Conclusion

Find out whether the water quality is related to its source location

There is no relationship between the entries' coordinates and water quality

Evaluate and select a multiclass classification model

RandomForestClassifier which reported an accuracy reported of 86.3%