“Predicting the age of Abalone”

First stage of Project

Danil Kirilenko
Ilda Alushaj

- Датасет (https://archive.ics.uci.edu/ml/datasets/Abalone) :
- Состоит из описания физических характеристик 4177 морских ушек (8 описательных признаков и 1 целевой – количество колец/возраст).
- Есть как категориальные, так и числовые признаки.

- Мы занимались решением двух задач:
- 1- предсказание самых часто встречающихся классов (7, 8, 9)
- 2- сравнение классификаторов и регрессоров для большего числа классов (19).

Первая часть
Решение задачи мультиклассовой (3) классификации

- Normalization of data
  1. Convert categorical feature into numerical one
  2. Garbage in/ Garbage out (Check if we have any missing values)
  3.Check if all values are int/float

- Selecting the target feature (number of rings) and classes for prediction (7, 8 and 9).
- Visualizing the dependence of specific features on the target feature using graphics
- Removing outliers that have no useful information for our data prediction
- Splitting our data into Train and Test data
- Creating classification models :
  1. Support Vector Machine (SVM)
  2. Random Forest Classifier
  3. Gradient Boosting
  4. KNN

" Support Vector Machine - SVM "

- Using Grid Search Algorithm (Hyperparameter tuning) to get the best parameters for our model
- In SVM - Grid Search gave us this best parameters : ({'C': 1.0, 'gamma': 10}, 0.5337704836165706)

```
Classification Report:
              precision    recall  f1-score   support

           7       0.56      0.51      0.53       140
           8       0.44      0.38      0.41       185
           9       0.63      0.74      0.68       242

    accuracy                           0.56       567
   macro avg       0.54      0.54      0.54       567
weighted avg       0.55      0.56      0.55       567

Training Score:  0.5741444866920152
Testing Score:  0.562610229276896
MSE: 0.6649029982363316
Wall time: 241 ms
```

"Random Forest Classifier"

- Grid search parameters : ({'max_depth': 3, 'n_estimators': 10}, 0.5386909059783622)
-

```
Classification Report:
              precision    recall  f1-score   support

           7       0.57      0.49      0.53       140
           8       0.42      0.35      0.38       185
           9       0.61      0.74      0.67       242

    accuracy                           0.55       567
   macro avg       0.53      0.53      0.53       567
weighted avg       0.54      0.55      0.54       567

Training Score:  0.5884030418250951
Testing Score:  0.5502645502645502
MSE: 0.6931216931216931
Wall time: 75 ms
```

Gradient Boosting

- Grid search hyperparameter tunning :

({'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 400}, 0.5320544160621445)

```
Classification Report:
              precision    recall  f1-score   support

           7       0.59      0.53      0.56       140
           8       0.36      0.34      0.35       185
           9       0.60      0.67      0.63       242

    accuracy                           0.53       567
   macro avg       0.52      0.51      0.51       567
weighted avg       0.52      0.53      0.52       567

Training Score:  0.6977186311787072
Testing Score:  0.5255731922398589
MSE: 0.6649029982363316
Wall time: 3.72 s
```

Knearest Neighbours -KNN

- Grid search best parameters :
- ({'metric': 'minkowski', 'n_neighbors': 8, 'weights': 'uniform'}, 0.5145910538232238)

```
Classification Report:
              precision    recall  f1-score   support

           7       0.51      0.61      0.56       140
           8       0.39      0.38      0.39       185
           9       0.61      0.55      0.58       242

    accuracy                           0.51       567
   macro avg       0.50      0.51      0.51       567
weighted avg       0.51      0.51      0.51       567

Training Score:  0.6169201520912547
Testing Score:  0.5114638447971781
MSE: 0.7689594356261023
Wall time: 165 ms
```

Results :

Model	Training_Score	Testing_Score	MSE	Wall_time(Lenovo_Thinkpad(i5-5300U))
SVM	0.57	0.56	0.66	241 ms
RandomForestClassifier	0.58	0.55	0.69	75 ms
GradientBoosting	0.69	0.52	0.66	3.72 s
KNN	0.61	0.51	0.76	165 ms

From the table, we can see that the best result was in SVM model with an accuracy 0.56 and a MSE(mean square error) 0.66. Wall time is 241 ms.

Вторая часть:
сравнение классификационного и регрессионного подходов для предсказания возраста морских ушек от 3 лет до 21 года (19 классов)

- Data preprocessing is done in the same way as in the first part.
- We build classification and regression models: LogReg, LinReg, RandomForest, GradientBoosting.
- To calculate the percentage of correct responses in regression models and then compare them with their classification counterparts, the regressor predictions are rounded to the nearest integer.

Random Forest Classifier

```
Accuracy:  0.28060522696011003
Wall time: 448 ms
```

Random Forest Regressor

```
MSE: 4.305987310766452
Wall time: 546 ms
```

Gradient Boosting Classifier

```
Accuracy:  0.26478679504814306
Wall time: 23.3 s
```

Gradient Boosting Regressor

```
MSE: 3.9578240946919103
Wall time: 1.05 s
```

Results :

Модель	Доля верных ответов	Скорость обучения и тестирования (Ryzen2600)	Среднеквадратичная ошибка (для регрессоров)
LogReg	0.268	93.2 ms	-
LinReg	0.242	2 ms	4.204
RanForClass	0.28	448 ms	-
RanForReg	0.233	546 ms	4.305
GradBoostClass	0.264	23.3 s	-
GradBoostReg	0.257	1.05 s	3.957

All pairs of classifiers coped with the task better than the regressors.However, the best classifier was Random Forest, when Gradient Bossting was the best among regressors