



INTRODUCTION TO MACHINE LEARNING

Midterm Project

Topic

Diagnosis of Cancer Using Blood Microbiome Data

Yusuf Seyitoğlu
1030516739

AMAÇ

Projedeki amaç kan mikrobiyom verilerini kullanarak kanser teşhisi yapabilmektir. En yaygın 4 kanser türüne ait toplamda 355 veri ile bu analizi gerçekleştirip kullanılan algoritmaların doğruluk oranları karşılaştırılacaktır.

YÖNTEM VE METOTLAR

Projede veri analizi için WEKA yazılımı python kullanılmıştır. Random Forest ve Gradient Boosted Tree algoritmaları kullanılarak veri setinin başarı oranları test edilmiştir. Algoritmaların başarı oranlarına etki eden parametrelerde değişiklikler yapılarak, bu değişikliklerin başarı oranındaki etkisi gözlemlenmiştir.

- **RANDOM FOREST**

- **GRADIENT BOOSTED TREE**

Yapılacak olan başarı karşılaştırmaları, yukarıda belirtilen algoritmalar ile yapılacaktır.

ALGORİTMALARIN DOĞRULUK KARŞILAŞTIRILMASI

RANDOM FOREST

Başarı Oranı: %95.7746

The screenshot displays the Weka Explorer interface with the Random Forest classifier selected. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Classifier output' section provides a detailed summary of the model's performance.

Classifier output

==== Described cross-validation ====

==== Summary ====

Correctly Classified Instances	340	95.7746 %
Kappa statistic	0.9393	
Mean absolute error	0.0944	
Root mean squared error	0.1595	
Relative absolute error	27.08 %	
Root relative squared error	38.2354 %	
Total Number of Instances	355	

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.962	0.012	0.973	0.962	0.977	0.967	1.000	0.999	colon cancer
	0.944	0.006	0.995	0.944	0.919	0.915	0.999	0.973	lung cancer
	0.925	0.004	0.990	0.925	0.957	0.940	0.999	0.997	breast cancer
	0.967	0.038	0.929	0.967	0.947	0.920	0.995	0.992	prostate cancer
Weighted Avg.	0.958	0.018	0.959	0.958	0.958	0.940	0.998	0.994	

==== Confusion Matrix ====

a	b	c	d	<-- Classified as
107	0	0	2	a = colon cancer
0	17	0	1	b = lung cancer
2	0	99	6	c = breast cancer
1	2	1	117	d = prostate cancer

Status

OK

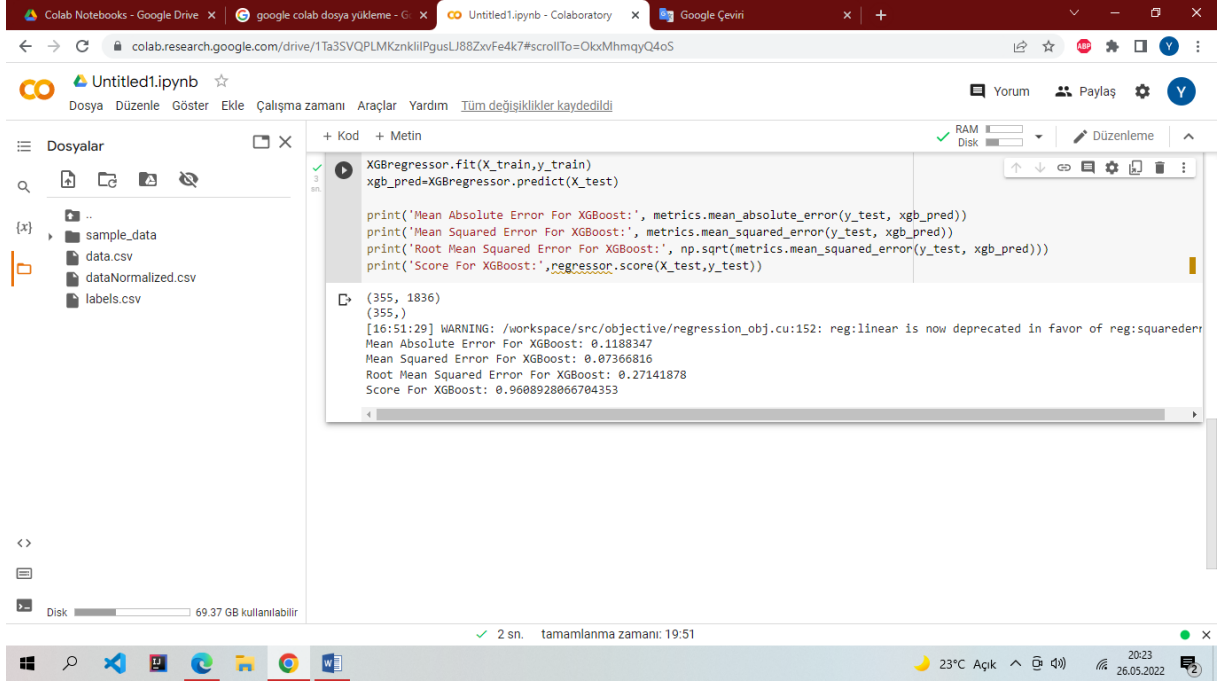
Log x 0

20:17 25.05.2022

GRADIENT BOOSTED TREE

Başarı Oranı: %96.08928066704353

Bu algoritma için Google Colab üzerinden XGBoost python uygulamasını ve birkaç kütüphaneyi kullandım. Csv dosyalarını okumak için pandas, lineer cebir işlemleri için numpy ve sklearn kütüphanelerini ekledim.



The screenshot shows a Google Colab notebook titled 'Untitled1.ipynb'. The left sidebar displays a file explorer with a folder named 'sample_data' containing files 'data.csv', 'dataNormalized.csv', and 'labels.csv'. The main area shows a code cell with the following Python code:

```
XGBRegressor.fit(X_train,y_train)
xgb_pred=XGBRegressor.predict(X_test)

print('Mean Absolute Error For XGBoost:', metrics.mean_absolute_error(y_test, xgb_pred))
print('Mean Squared Error For XGBoost:', metrics.mean_squared_error(y_test, xgb_pred))
print('Root Mean Squared Error For XGBoost:', np.sqrt(metrics.mean_squared_error(y_test, xgb_pred)))
print('Score For XGBoost:', regressor.score(X_test,y_test))
```

The output of the code cell is as follows:

```
(355, 1836)
(355,)
[16:51:29] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror
Mean Absolute Error For XGBoost: 0.1188347
Mean Squared Error For XGBoost: 0.07366816
Root Mean Squared Error For XGBoost: 0.27141878
Score For XGBoost: 0.9608928066704353
```

The bottom status bar indicates that the code was executed 2 seconds ago and the notebook is still running.