NASA Kepler Exoplanet Search

Data Mining

Aleix Boné Eduard Bosch David Gili Albert Mercadé

December 23, 2020

UPC

Data source

Our Dataset

NASA's Exoplanet Search Results [1]

Original Dataset

9.564 data records
8.48% of missing data
50 variables

Target column

`koi_disposition`

- CONFIRMED
- FALSE POSITIVE
- (CANDIDATE)
- (NOT DISPOSED)

[1] Dataset URL: https://www.kaggle.com/nasa/kepler-exoplanet-search-results

Goal

The focus of the project is to correctly classify potential exoplanets based on the raw measurements taken by NASA's Kepler Space Observatory.

Pre-processing

Pre-processing steps

- Feature removal
- Example removal
- Value transformation
- Error and missing data treatment
- Feature selection

Feature selection

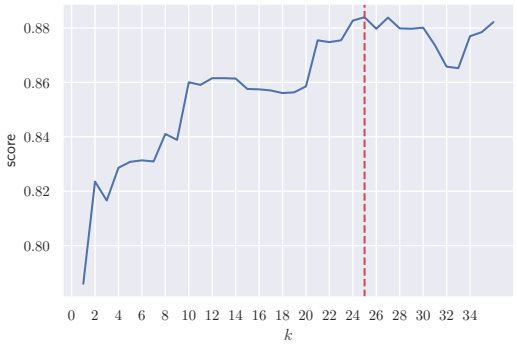


Figure 1: Cross validation score for different k values

Evaluation criteria

Validation Dataset

- 30(test) : 70(training) split
- Target category proportions maintained

Parameter Optimization

5-fold cross validation

Metrics

F-measure

Execution of Machine Learning methods

Naïve Bayes - Normalization

	No normalization	Standardization	Power Transform
Confusion matrix	$\begin{bmatrix} 208 & 200 \\ 3 & 189 \end{bmatrix}$	$\begin{bmatrix} 287 & 121 \\ 2 & 190 \end{bmatrix}$	$\begin{bmatrix} 357 & 51 \\ 19 & 173 \end{bmatrix}$
Accuracy:	0.661	0.795	0.883
F1 score:	0.65	0.755	0.831

Naïve Bayes - Parameter tuning

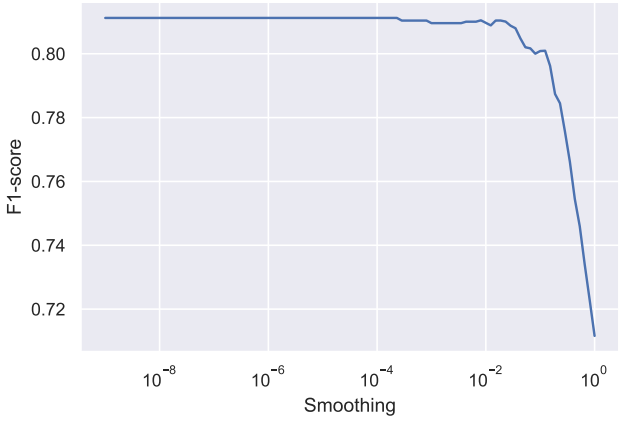


Figure 2: Naïve Bayes smoothing

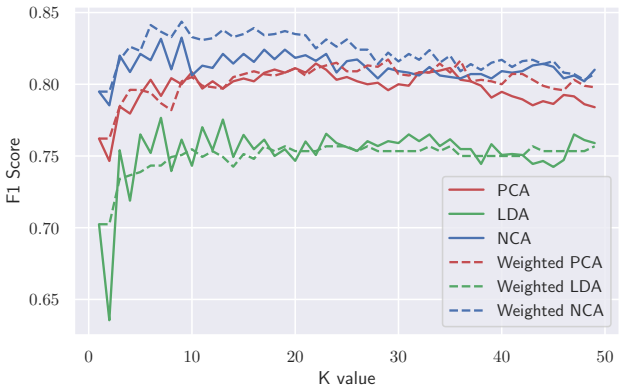


Figure 3: weighted and unweighted knn with PCA, LDA and NCA

K-NN

Confusion matrix on test set: $\begin{bmatrix} 367 & 41 \\ 22 & 170 \end{bmatrix}$

Accuracy on test set: 0.895

F1 score on test set: 0.844

Decision Trees - Parameter tuning

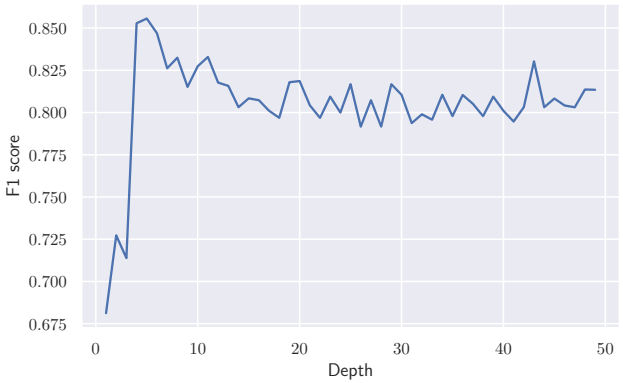


Figure 4: Decision trees F1 Score depending on the maximum depth of the decision tree.

Decision Trees

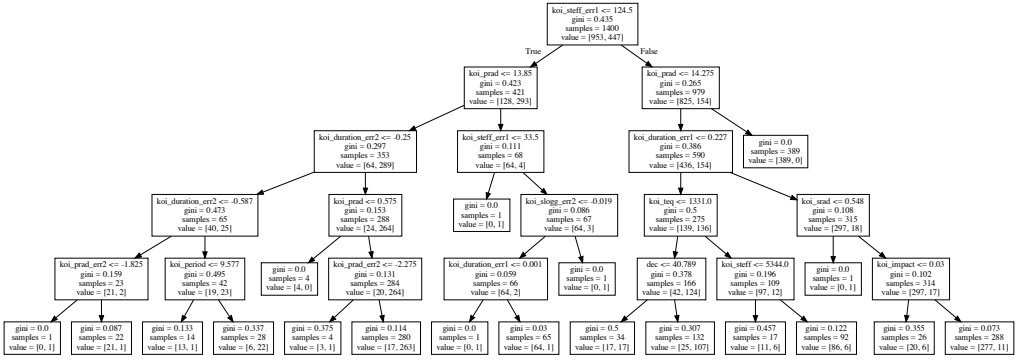


Figure 5: Our best-performing decision tree.

Support Vector Machines - Lineal kernel

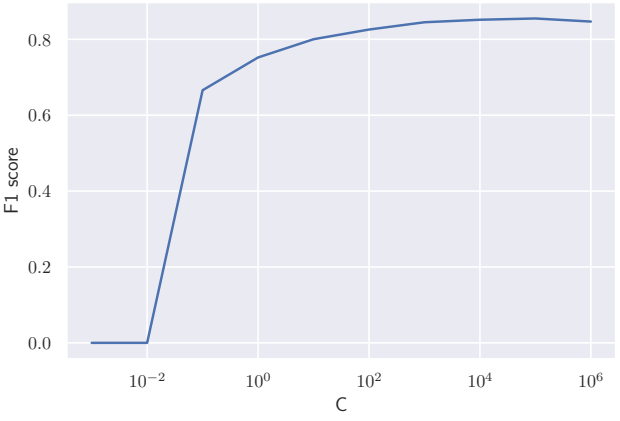


Figure 6: linear SVM C parameter search

Support Vector Machines - Lineal kernel

Best results ($C = 10^5$)	
Confusion matrix on test set:	$\begin{bmatrix} 383 & 25 \\ 26 & 166 \end{bmatrix}$
Accuracy on test set:	0.915
F1 on test set:	0.8668
Number of supports:	288 (269 of them have slacks)
Proportion of supports:	0.2057

Support Vector Machines - Polynomial and RBF kernels

Table 1: Comparison of different SVM kernels

Kernel	Accuracy	F1	Supports	Proportion	C	γ
Linear	0.9150	0.8668	288(269)	0.2057	10^5	
Polynomial 2	0.9067	0.8542	332(283)	0.2371	10^4	
Polynomial 3	0.9050	0.8503	356(317)	0.2543	10^3	
RBF	0.9083	0.8564	330(302)	0.2357	10^6	0.001

Performance majority voting i

Table 2: Majority voting results

Method	Accuracy
Naïve Bayes	0.884
K-NN	0.857
Decision Tree	0.877
Majority voting	0.914
Majority voting (weighted)	0.914

Performance majority voting ii

With hard voting:

Confusion matrix on test set:	$\begin{bmatrix} 375 & 33 \\ 20 & 172 \end{bmatrix}$
Accuracy on test set:	0.9117
F1 score on test set:	0.8622

With weighted voting (2 1 2):

Confusion matrix on test set:	$\begin{bmatrix} 373 & 35 \\ 19 & 173 \end{bmatrix}$
Accuracy on test set:	0.9100
F1 score on test set:	0.8492

Bagging

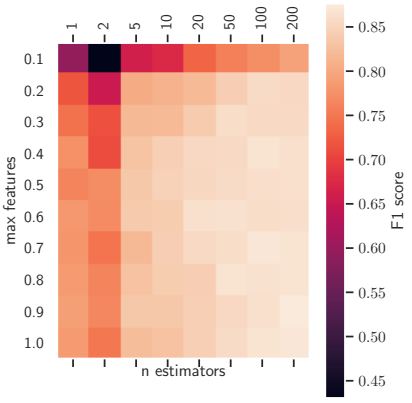


Figure 7: Bagging parameter search

Bagging

Best results: $\texttt{n\_est} = 200$, $\texttt{max\_features} = 0.9$

Confusion matrix on test set:	$\begin{bmatrix} 388 & 20 \\ 24 & 168 \end{bmatrix}$
Accuracy on test set:	0.9267
F1 score on test set:	0.8825

Table 3: RandomForest best parameters

Parameter	Value
bootstrap	True
max_depth	150
max_features	10
min_samples_leaf	5
min_samples_split	5
n_estimators	500

Random Forest

Confusion matrix on test set: $\begin{bmatrix} 384 & 24 \\ 21 & 171 \end{bmatrix}$

Accuracy on test set: 0.925

F1 score on test set: 0.883

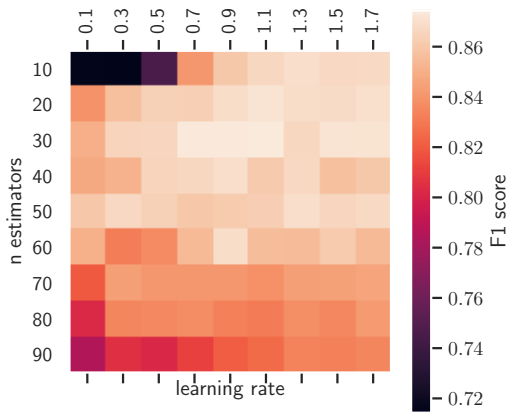


Figure 8: AdaBoost parameter search

AdaBoost

When executing Adaboost with the best parameters found we obtain the following results:

Confusion matrix on test set:	$\begin{bmatrix} 383 & 25 \\ 21 & 171 \end{bmatrix}$
Accuracy on test set:	0.923
F1 score on test set:	0.881

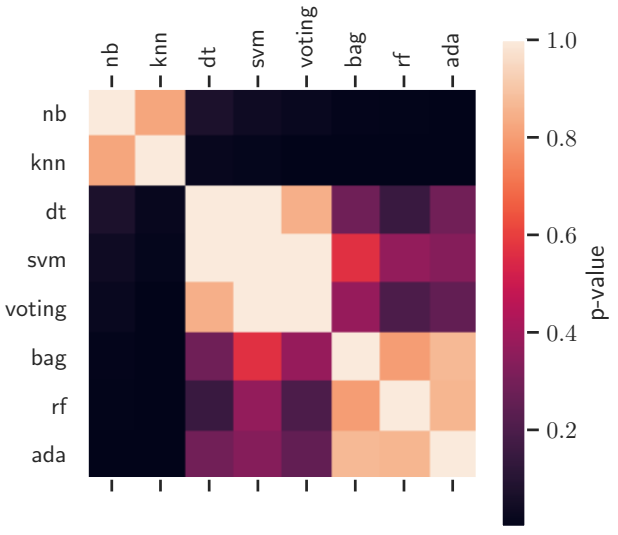
Comparison

Comparison - Accuracy and F1

Table 4: Comparison of metrics

	Accuracy	F1
nb	0.883333	0.831731
knn	0.878333	0.814249
dt	0.910000	0.858639
svm	0.911667	0.865140
voting	0.910000	0.860825
bag	0.920000	0.874346
rf	0.923333	0.878947
ada	0.923333	0.881443

Comparison - mcNemar



Thank you for your attention, any questions?

Thank you for your attention, any questions? I guess not :/