Classification Algorithms

TERRORISM PREDICTION in Middle East and North Africa

29/11/2018

V. Fokker, T.J.C. Meulenbroek,
K. Raijmann, R. Warmels

JADS Jheronimus Academy of Data Science

Terrorism

- Highly relevant topic
- Interest in applicability of the research
- Data from the Global Terrorism Database

Iran military parade attacked by gunmen in Ahvaz

22 September 2018

Share

The moment gunmen open fire on an Iranian military parade

Gunmen have opened fire on an Iranian military parade in the south-western city of Ahvaz, killing at least 25 people, including civilians, and injuring 60, state media say.

The attackers shot from a park near the parade and were wearing military uniforms, reports say.

An anti-government Arab group, Ahvaz National Resistance, and Islamic State (IS) have both claimed the attack.

President Hassan Rouhani has vowed a "harsh response".

"The response of the Islamic Republic of Iran to the smallest threat will be harsh, but those who sponsor the terrorists must be held accountable," he said in a statement.

Original Study


Internatio... nal of Emerging Trends & Technology in Compute... (IJETTCS)
...Site: www.ijettcs.org Email: editor@ijett...
Volume ... 2015 ... -6856

H... ICATIO...
F... e

Motaz M. H. ... A. Soliman

Faculty of Computers ... Giza 12613, Egypt

Abstract
Machine lear... support a... impli... accur... knowled... s. Hybrid machine ... gths of both knowledge ... n this research we compare pre... comprehensibility of explicit, implici... ine learning models. This research based o... terrorist groups responsible of attacks in Middle E... North Africa from year 2009 up to

...pirically shown to ... n single weak ... ensional, ... [2]. Adap... al in computati... a main reason for be... itary of its components.
The integration of th... s into hybrid machine learning solutions ... more intelligent search and reasoning methods ... ch various domain

Research Question

What standard supervised machine learning techniques aid in classifying terrorist groups responsible for terrorist attacks in the Middle East and North Africa based on open source data?

Our approach

- Focus on the middle east
- Use open data (GTD)
- Use ‘standard’ supervised methods:
  - Support Vector Machine
  - Naive Bayes
  - Decision Trees
  - K-nearest neighbour
  - New: Random Forests

Preprocessing Variables

- iyear
- imonth
- **latitude**
- **longitude**
- attacktype1
- weaptype1
- targtype1
- **nkill**
- **nwound**

- **population**
- **Environment**

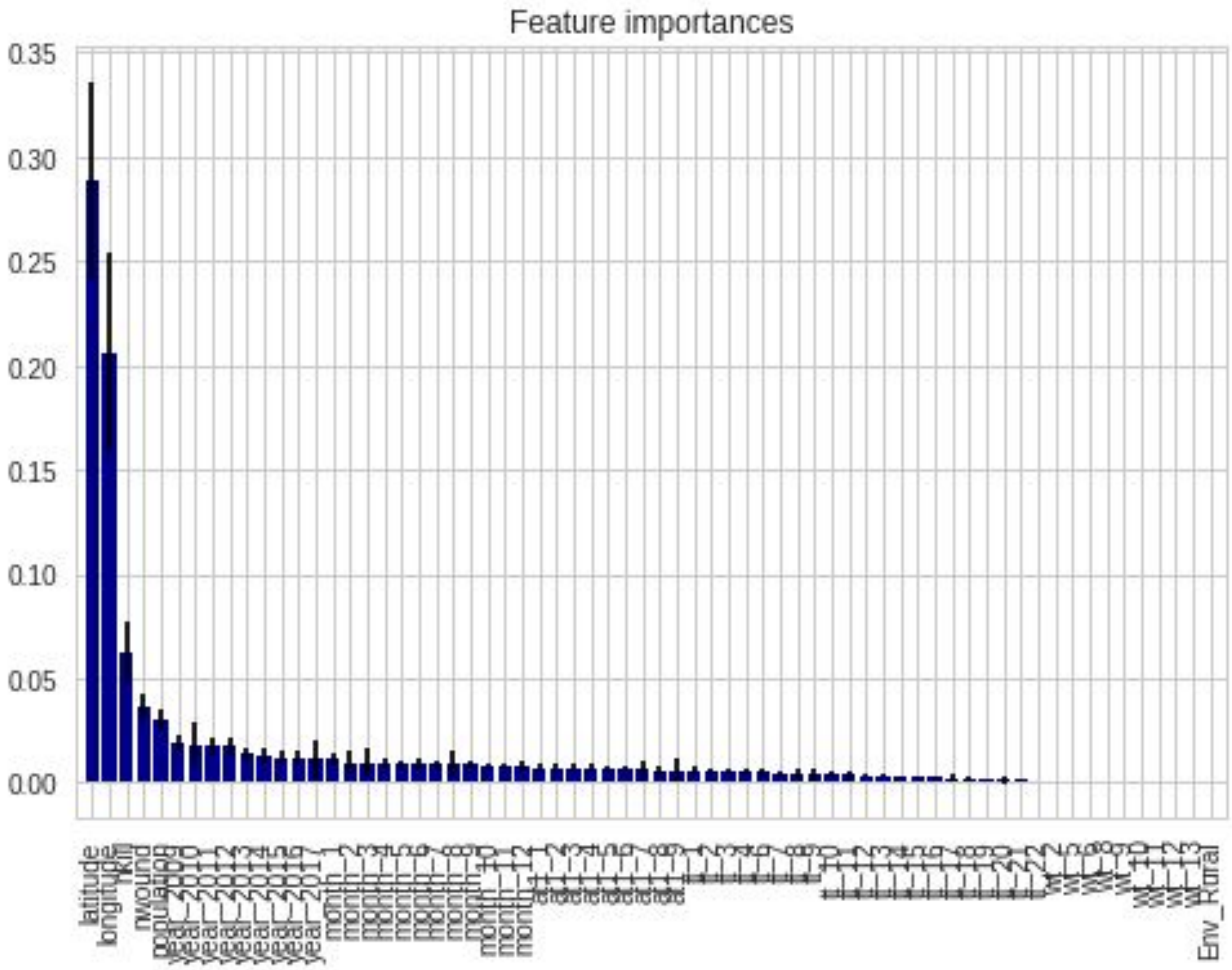
```
Int64Index: 11607 entries, 4 to 34325
Data columns (total 12 columns):
iyear          11607 non-null int64
imonth         11607 non-null int64
gname          11607 non-null object
latitude       11607 non-null float64
longitude      11607 non-null float64
attacktype1    11607 non-null int64
weaptype1      11607 non-null int64
targtype1      11607 non-null int64
nkill          11607 non-null float64
nwound         11607 non-null float64
population     11607 non-null float64
Environment    11607 non-null object
dtypes: float64(5), int64(5), object(2)
```

Preprocessing Locations

Jheronimus Academy of Data Science

Preprocessing Feature importance

- Latitude / Longitude
- NKill
- Nwound
- Population
- All other ‘binarized’

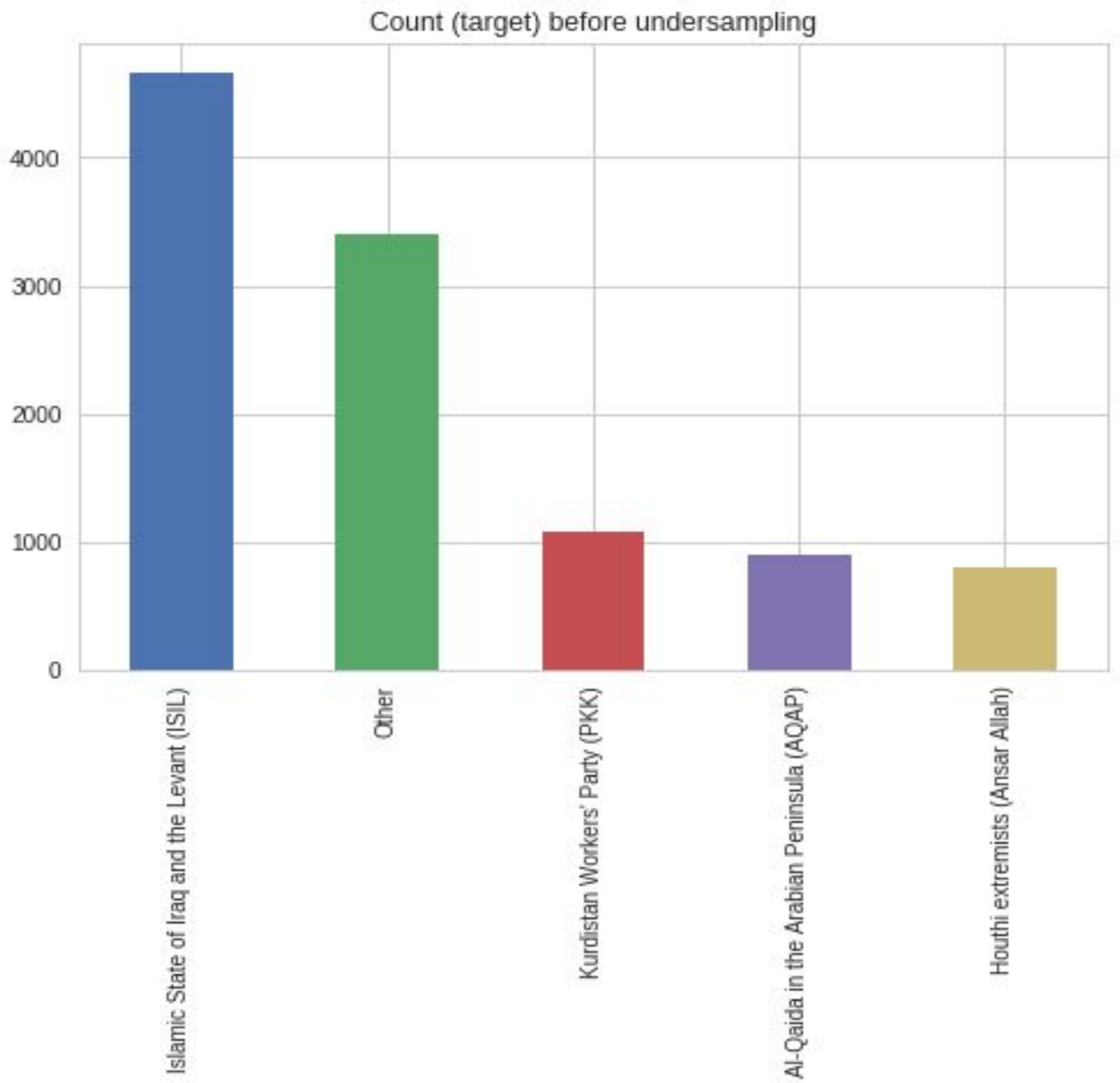


Preprocessing Target Variable

- GName
- 463 groups
- 193 groups after filtering for inactive groups and dropping missing values

- Only use 4 largest groups (64%)
- Other group for other terrorist groups
- Total of 5 groups

- Sampling using pipeline



Train / Test split

- Random State = 42 used for replication purposes
- Train / Test split with RepeatedKFold with 6 splits and 10 repeats.

Train	Test
9027	1805

Pipeline

Undersampling

- Random Undersampling
- Undersampling with Instance hardness threshold
- Undersampling with Condensed Nearest Neighbour

Kits used

- IMbalanced-learn 0.4.3
- Scikit-learn 0.20.0
- scipy 0.13.3
- numpy 01.8.2

Models - overview

Naive Bayes

K Nearest Neighbors

Support Vector Machine

Decision Trees

Random Forest

Naive Bayes

	undersampling method	precision	recall	F-measure
GaussianNB	**Random Undersampling**	**0.87**	**0.83**	**0.84**
	Instance Hardness	0.85	0.80	0.81
	Condensed Nearest Neighbors	0.37	0.48	0.37
MultinominalNB	Random Undersampling	0.37	0.37	0.41
	Instance Hardness	0.64	0.35	0.39
	Condensed Nearest Neighbors	0.39	0.27	0.27
BernoulliNB	Random Undersampling	0.56	0.46	0.48
	Instance Hardness	0.55	0.39	0.41
	Condensed Nearest Neighbors	0.25	0.16	0.11

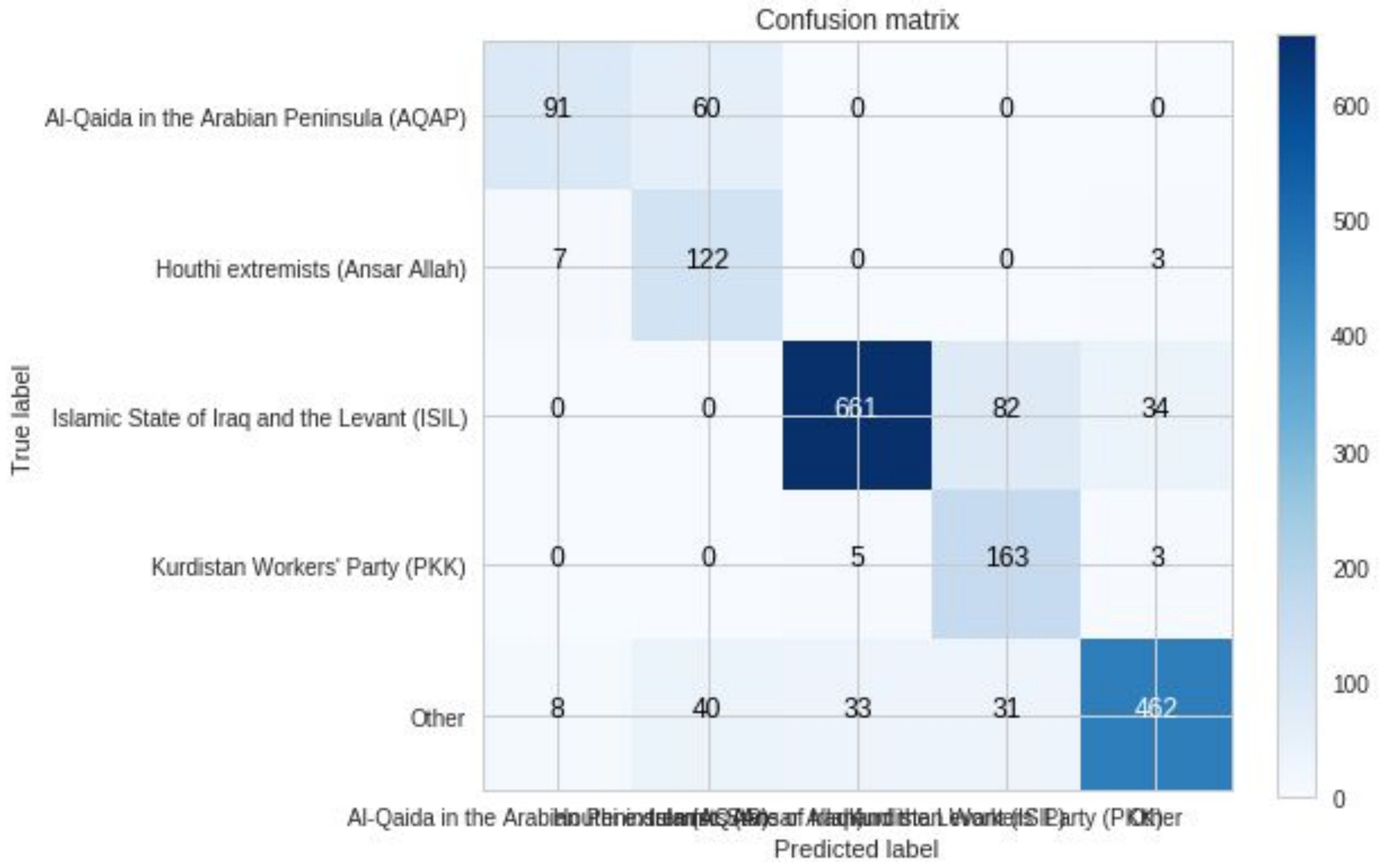


Figure: GaussianNB with random undersampling

Naive Bayes - assumption

- With GaussianNB, the **assumption** is that the features need to be normally distributed.
- Check for normality!

- Normality Rejected for for top 5 of the most important features.

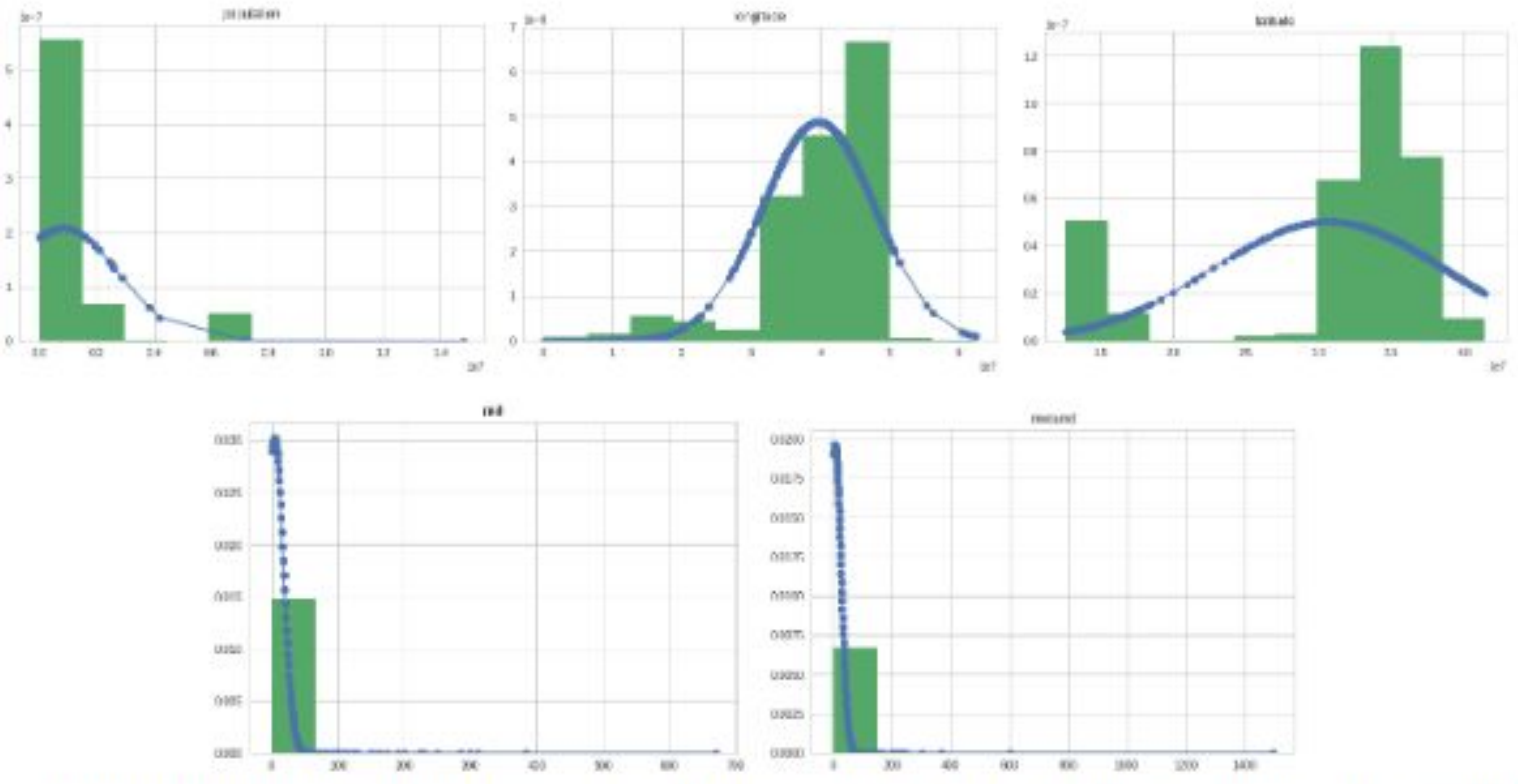


Figure: Distributions of the features “population”, “longitude”, “latitude” , “nkill” and “nwound”. These features are most important according to the feature imporantance.

Models - overview

Naive Bayes

K Nearest Neighbors

Support Vector Machine

Decision Trees

Random Forest

K-Nearest Neigborgs

		precision	recall	F-measure
KNN	**Random Undersampling**	**0.92**	**0.91**	**0.91**
	Instance Hardness	0.90	0.89	0.89
	Condensed Nearest Neighbors	0.62	0.67	0.63

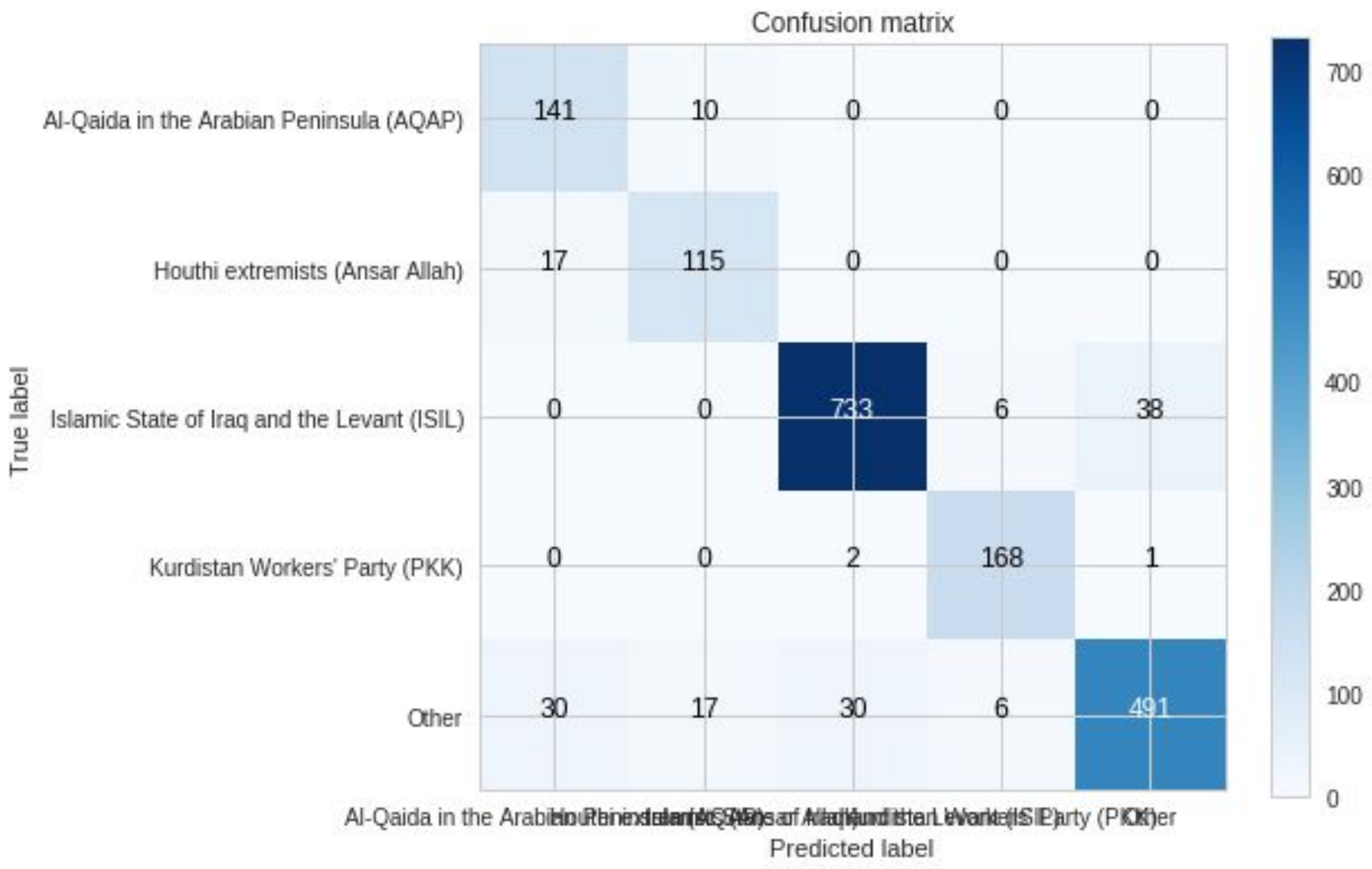


Figure: KNN with random undersampling

K-NN - k value tuning

- K- value is tuned to **k= 3**

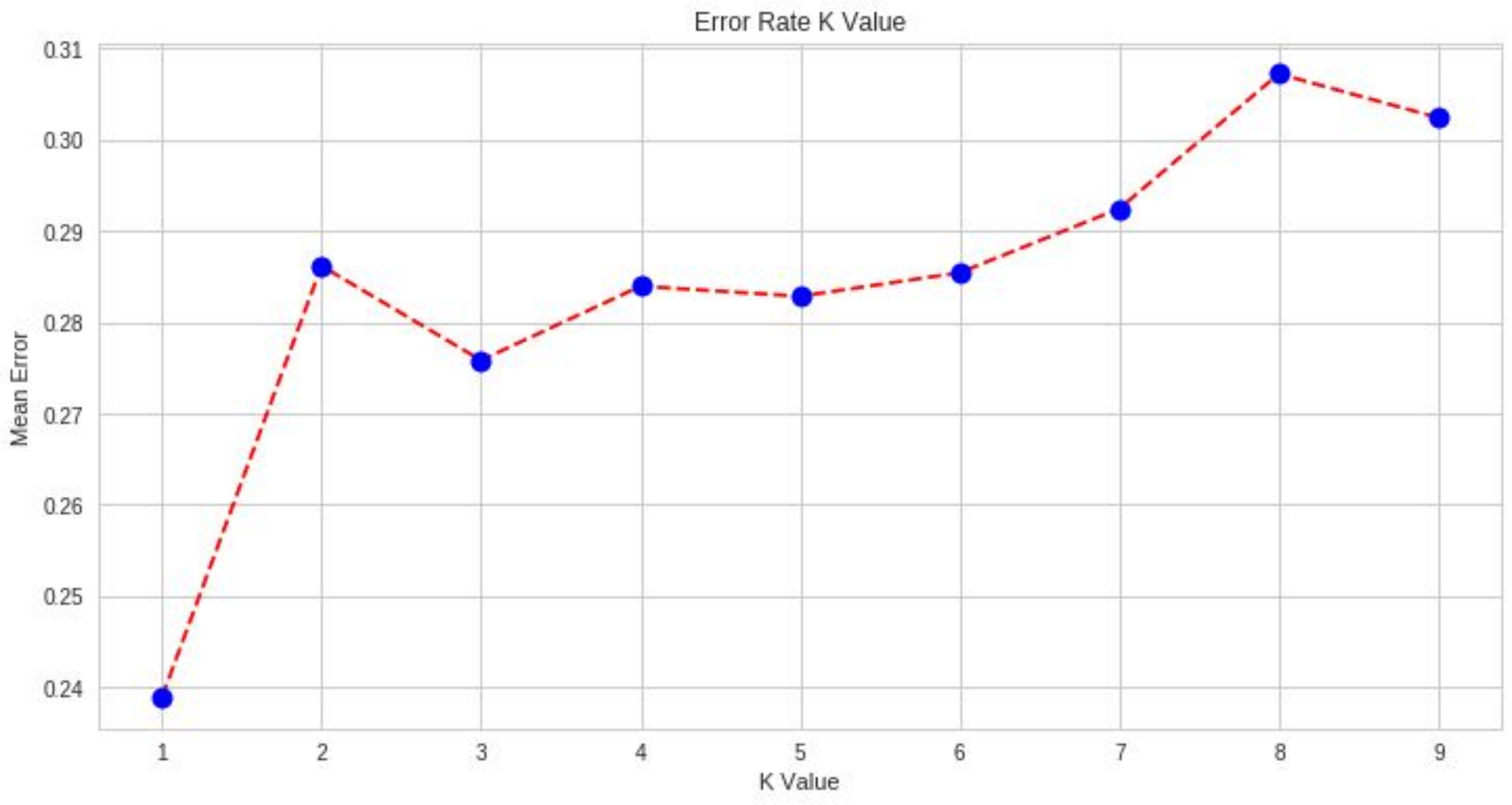


Figure: Error rate versus the k-value

Models - overview

Naive Bayes

K Nearest Neighbors

Support Vector Machine

Decision Trees

Random Forest

Support Vector Machine

		precision	recall	F-measure
SVM	**Random Undersampling**	**0.86**	**0.58**	**0.58**
	Instance Hardness	0.88	0.52	0.52
	Condensed Nearest Neighbors	0.07	0.07	0.07

- **Radial basis function kernel used**
  - Gaussian

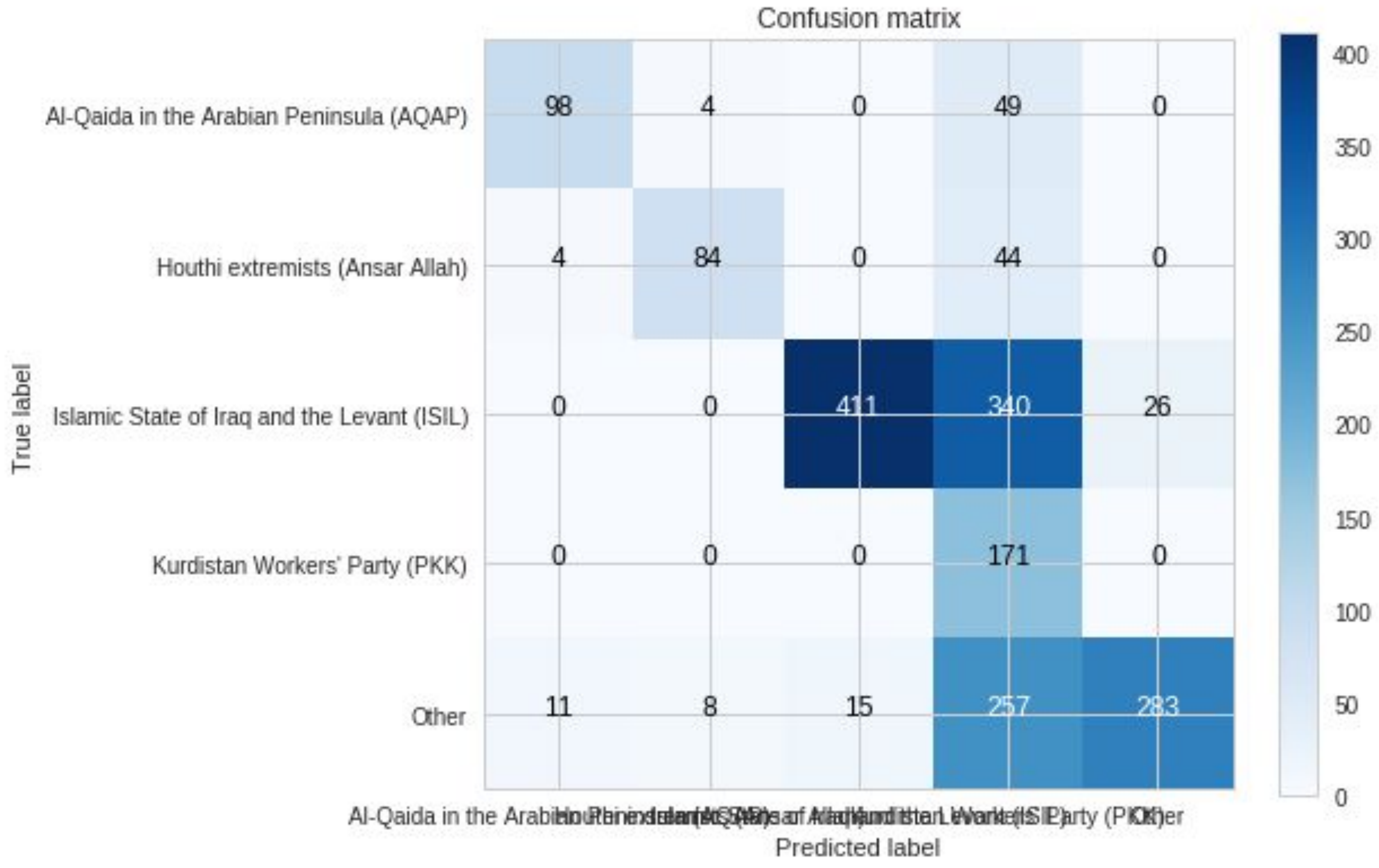


Figure: SVM with random undersampling

Models - overview

Naive Bayes

K Nearest Neighbors

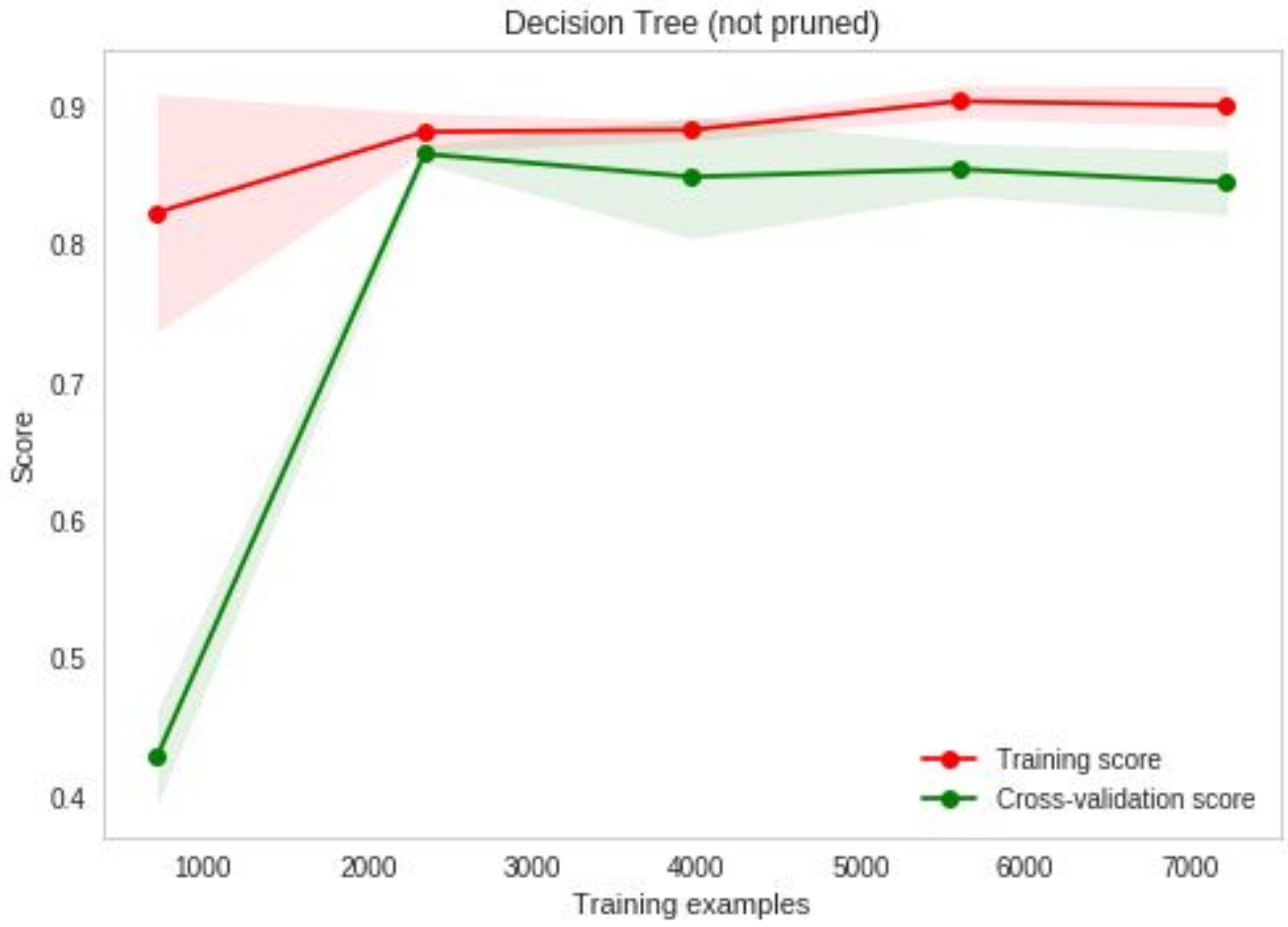
Support Vector Machine

Decision Trees

Random Forest

Decision Tree

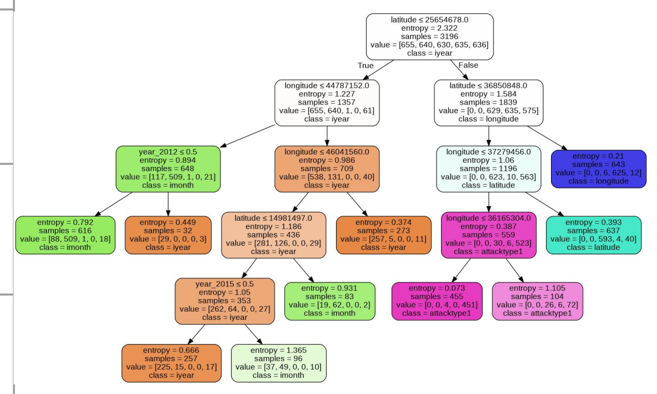
Decision Tree

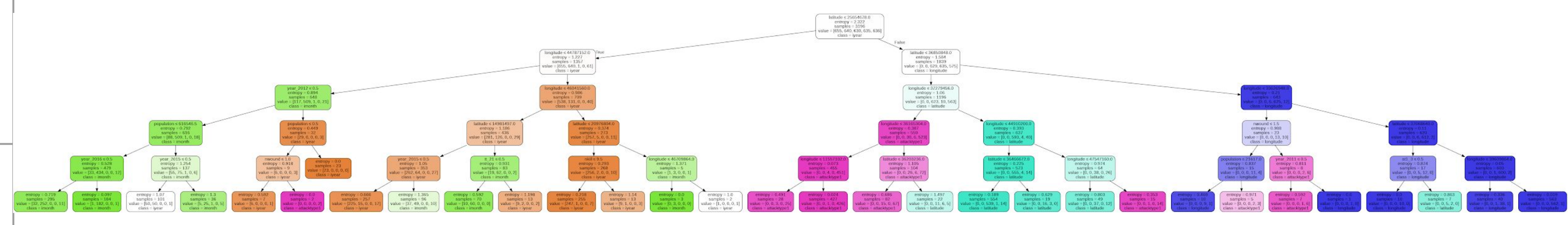


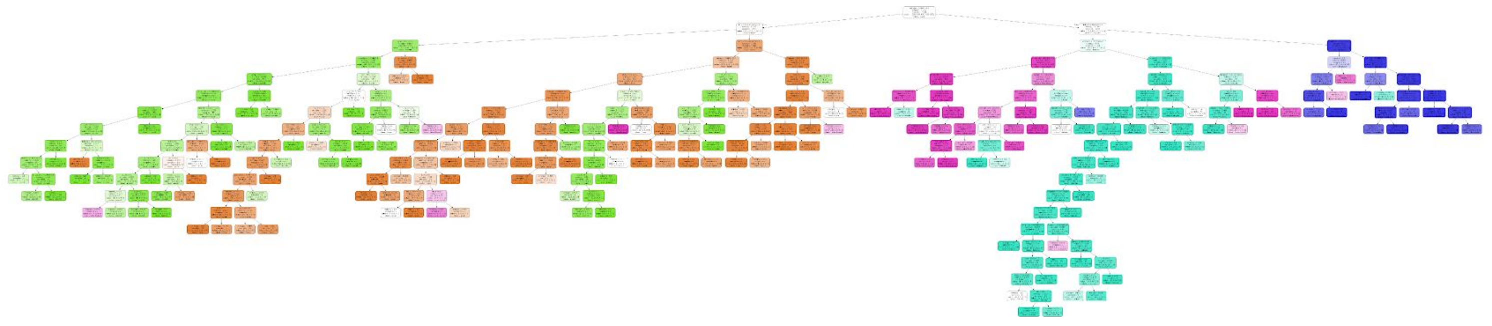
- Overfitting
  - Coping mechanisms

Decision Tree

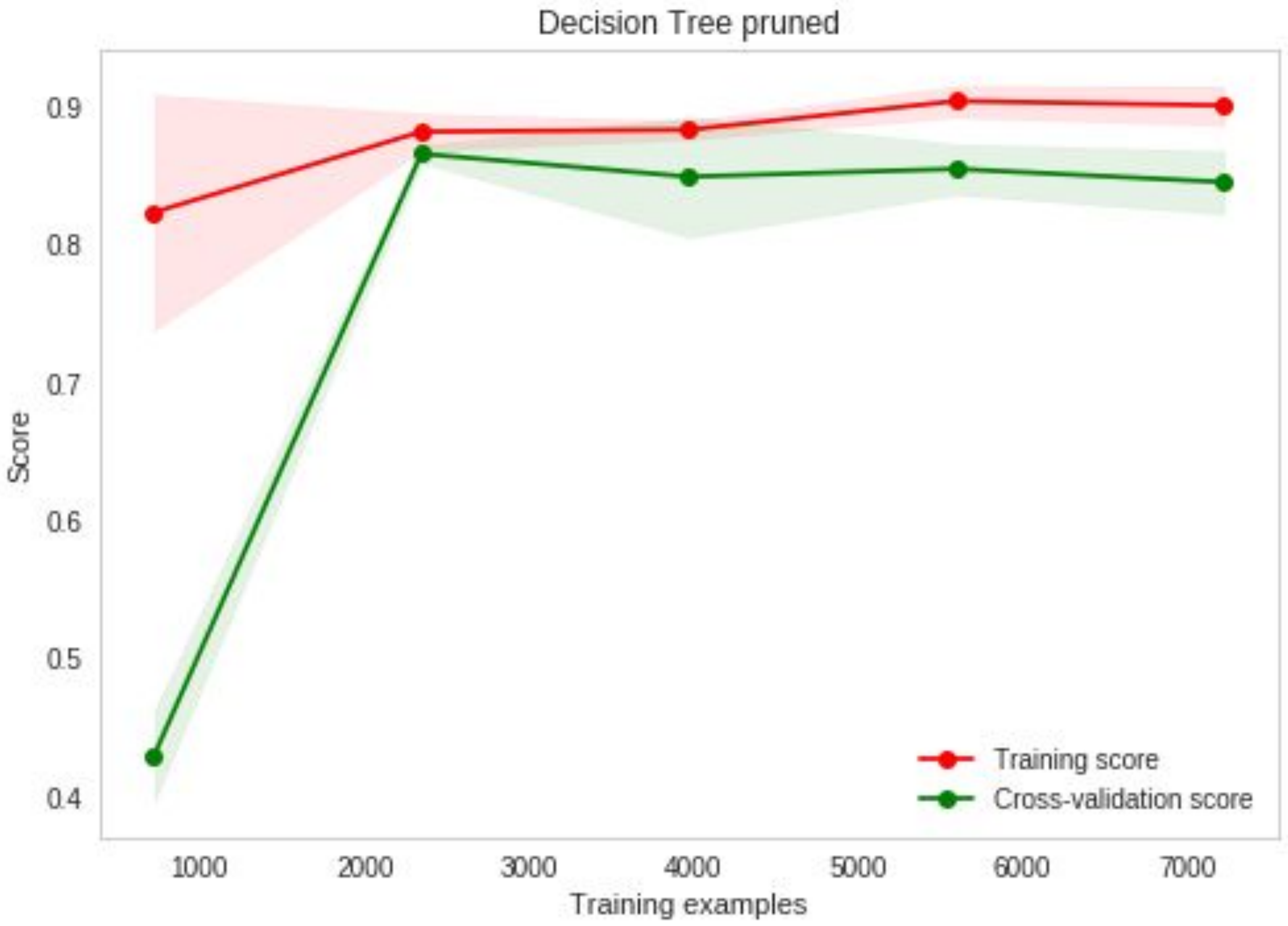
		precision	recall	F-measure
DT -unpruned	**Random Undersampling**	**0.92**	**0.91**	**0.91**
(underfitted)	Instance Hardness	0.90	0.88	0.89
	Condensed Nearest Neighbors	0.41	0.40	0.37
DT Pruned	**Random Undersampling**	**0.90**	**0.90**	**0.90**
10 leafs	Instance Hardness	0.89	0.87	0.88
	Condensed Nearest Neighbors	0.28	0.41	0.32
DT Max depth	**Random Undersampling**	**0.91**	**0.90**	**0.91**
Max Dept = 5	Instance Hardness	0.89	0.88	0.88
	Condensed Nearest Neighbors	0.35	0.36	0.33
DT	**Random Undersampling**	**0.92**	**0.91**	**0.91**
Min leafs 5	Instance Hardness	0.90	0.88	0.89
	Condensed Nearest Neighbors	0.36	0.38	0.35



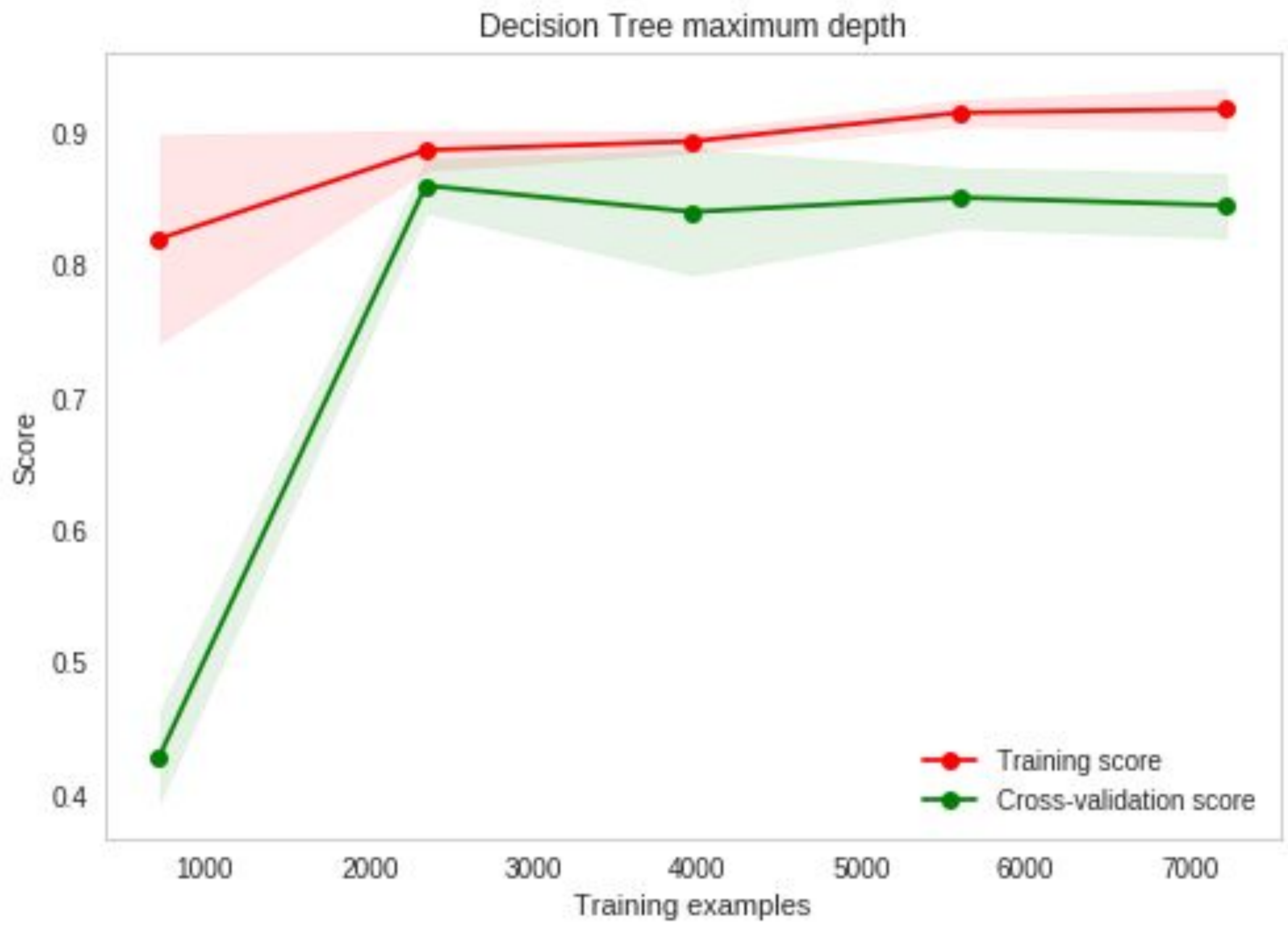




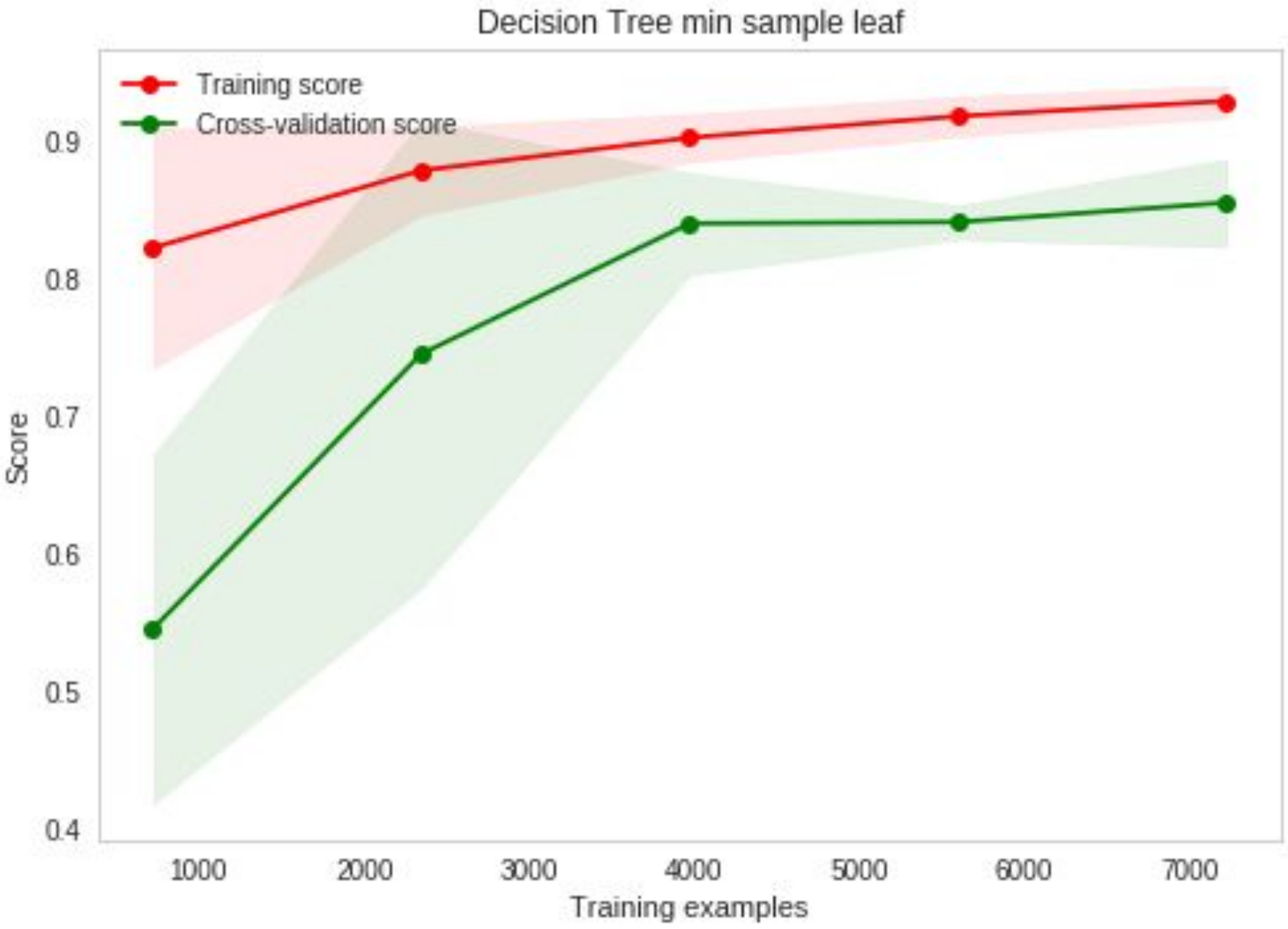
Decision Tree



Decision Tree



Decision Tree



Models - overview

Naive Bayes

K Nearest Neighbors

Support Vector Machine

Decision Trees

Random Forest

Random Forest

		precision	recall	F-measure
RF	**Random Undersampling**	**0.93**	**0.92**	**0.92**
	Instance Hardness	0.89	0.87	0.87
	Condensed Nearest Neighbors	0.31	0.19	0.16

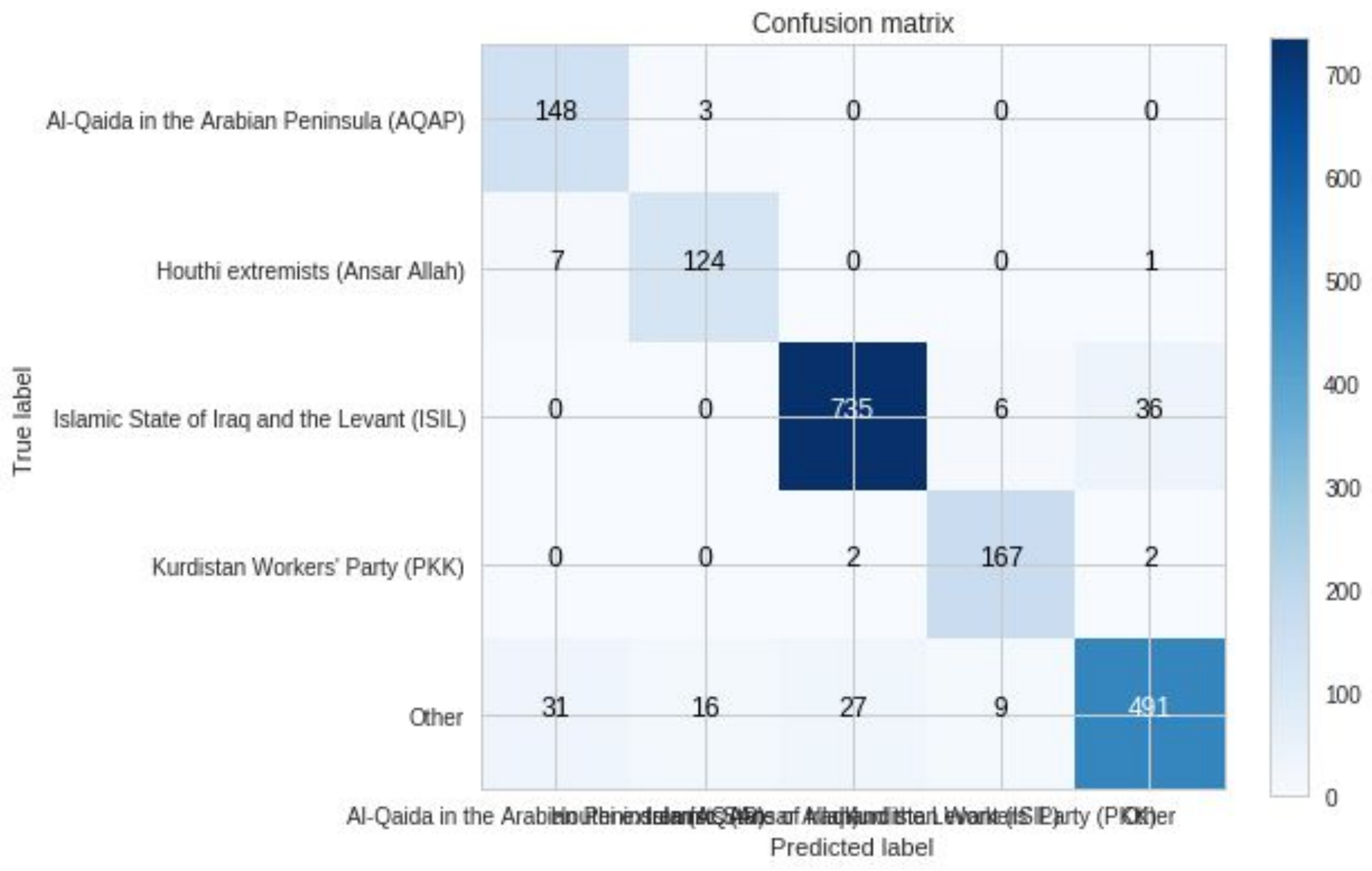
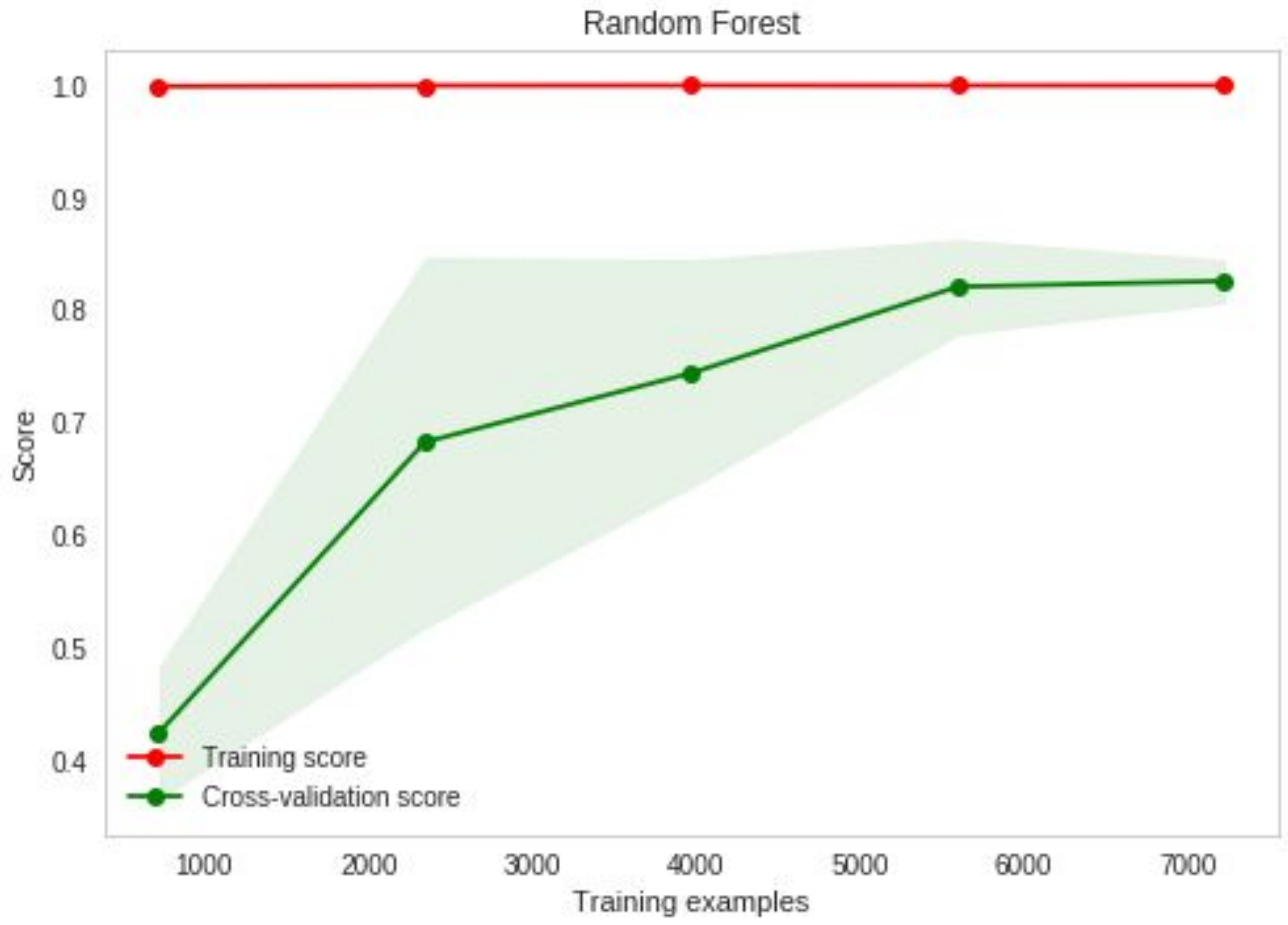


Figure: RF with random undersampling

Random Forest



- Overfitting
  - Remedy: insert more trees

Models - overview

- Random undersampling works best in all models
- All models show good F-measures, indicating decent models.
- GaussianNB performs well, but assumption not met.

Conclusion

- Nice results (for only 4 and the ‘other’ group)
- The methods have no aid in real life

Discussion

- Feature selection and lots of categorical data
- Scope of variables
- Shortcomings API (Population & Environment)
- Small groups target variable
- Algorithms: more possibilities and algorithms possible
- Real life application

Recommendations

- More data necessary?
- Try more algorithms and parameters
- Include more variables
- Make use of closed source intelligence

Learning achievements

- Google Colab is not (yet) functioning correctly
- Touched upon multiple algorithms
- Preprocessing masters
- Improved our python skills
- More package knowledge for Data Science purposes
- Learned to make a trade-off between effort and result
- Got real enthousiastic for Machine Learning

Classification Algorithms

TERRORISM PREDICTION in Middle East and North Africa

29/11/2018

V. Fokker, T.J.C. Meulenbroek,
K. Raijmann, R. Warmels

JADS Jheronimus Academy of Data Science

Blank

Text

Body1

Comparison

Heading1

Body1

Heading2

Body2

Ordered List

1 | Item1

2 | Item2

3 | Item3

4 | Item4

5 | Item5

Unordered List

Heading1

- Body1
- Body2
- Body3

- Item1
- Item2
- Item3
- Item4
- Item5

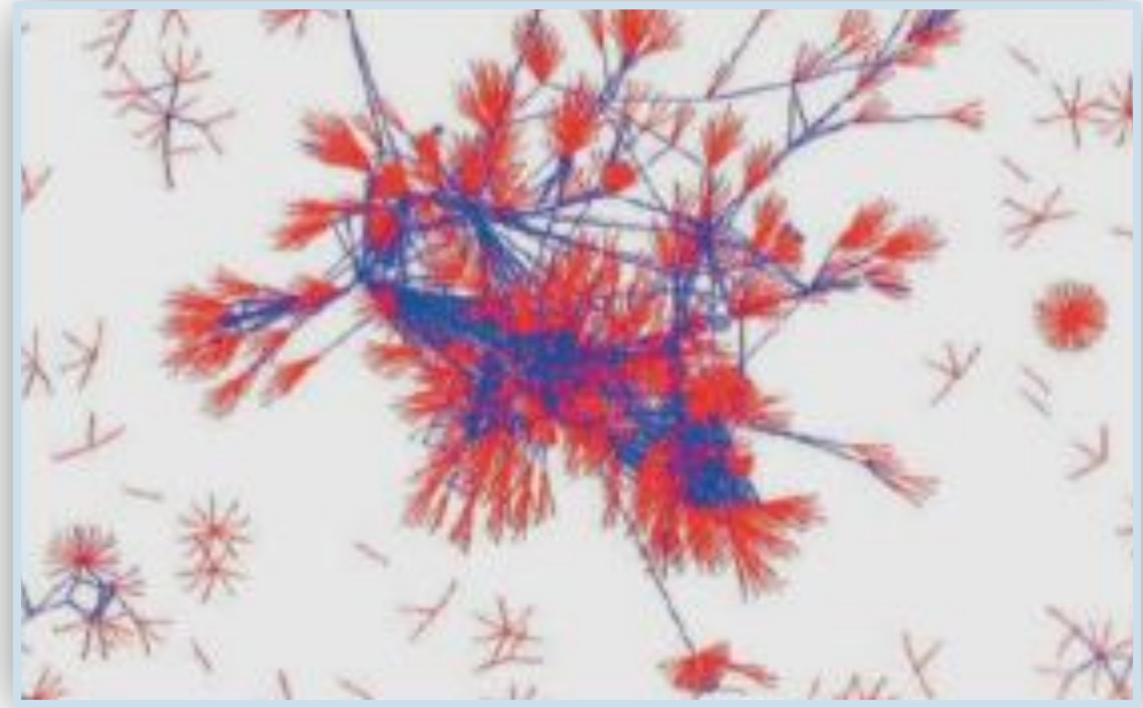
Single Image
Tagline

- Item1
- Item2
- Item3
- Item4
- Item5

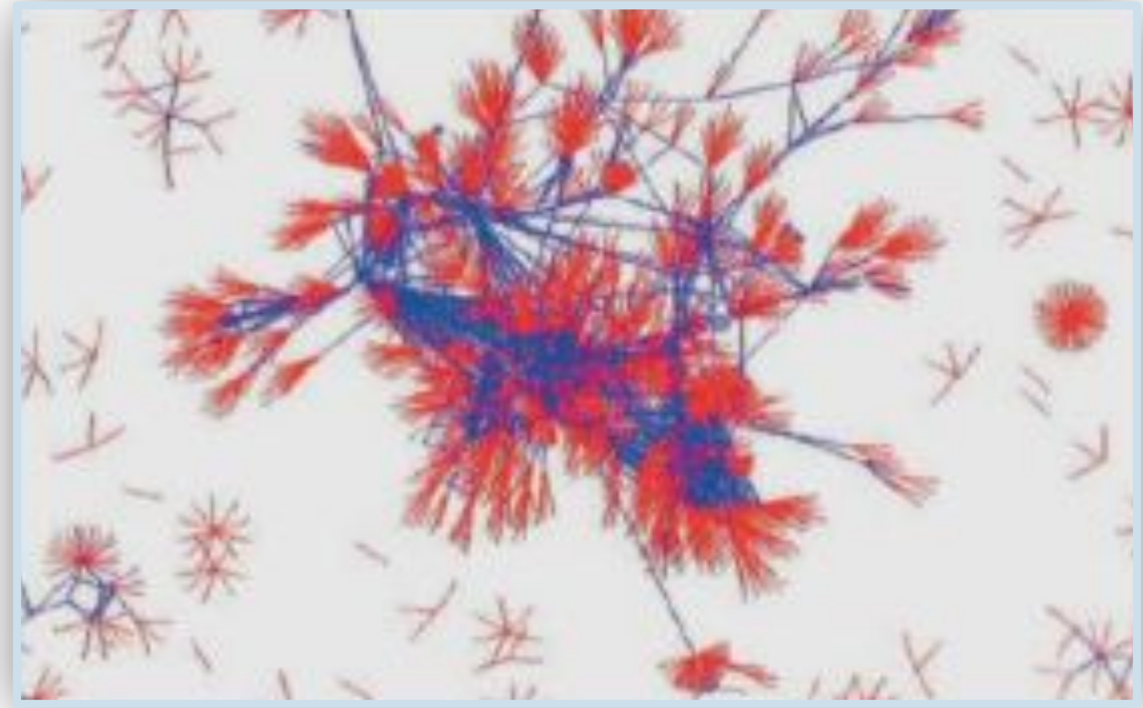
JADS Jheronimus Academy of Data Science

Multiple Images

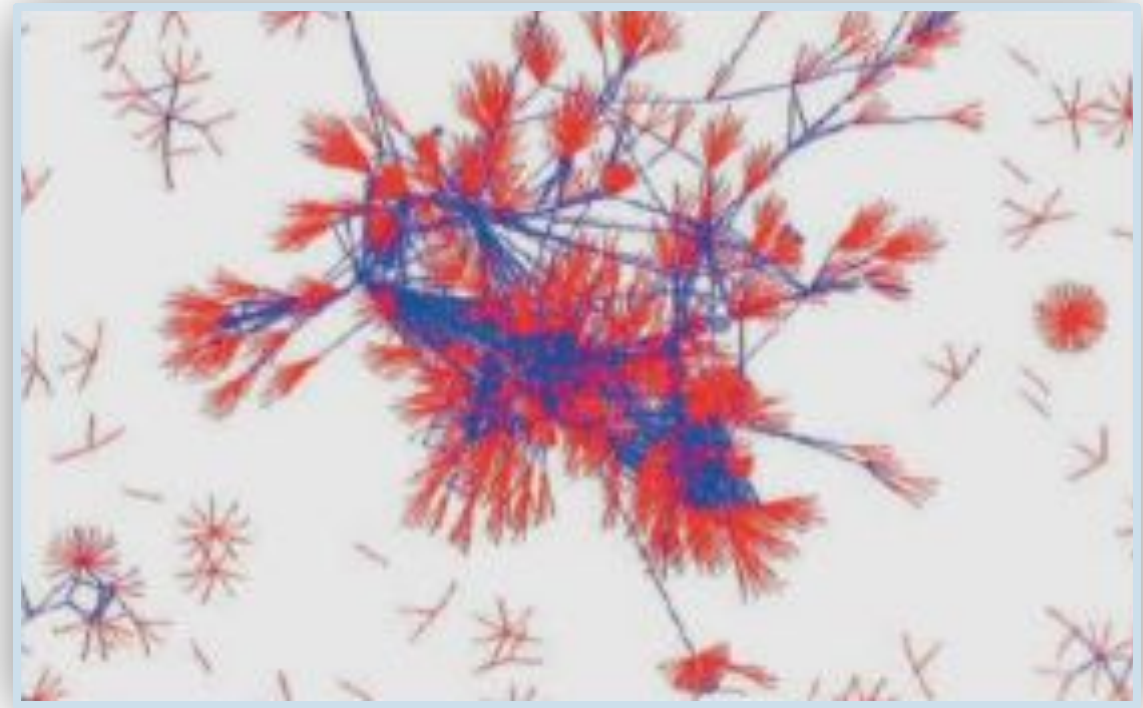
Taglin
e



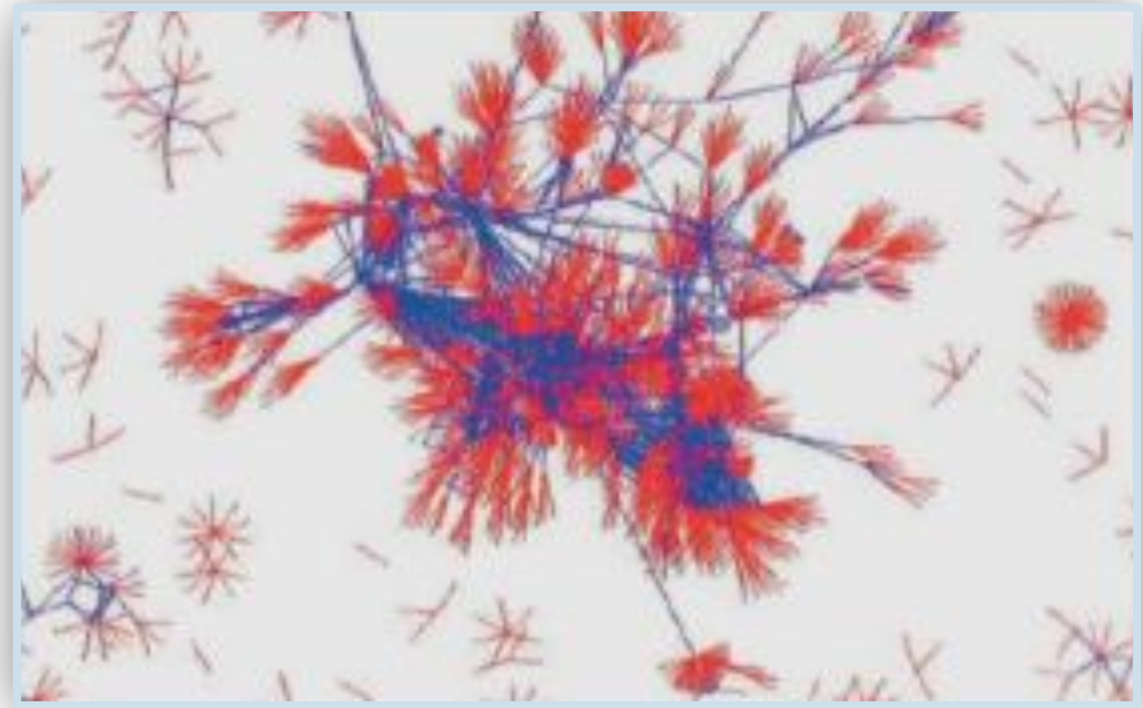
Caption1



Caption2



Caption3



Caption4

Heading1

Caption1

“A LONG QUOTE”

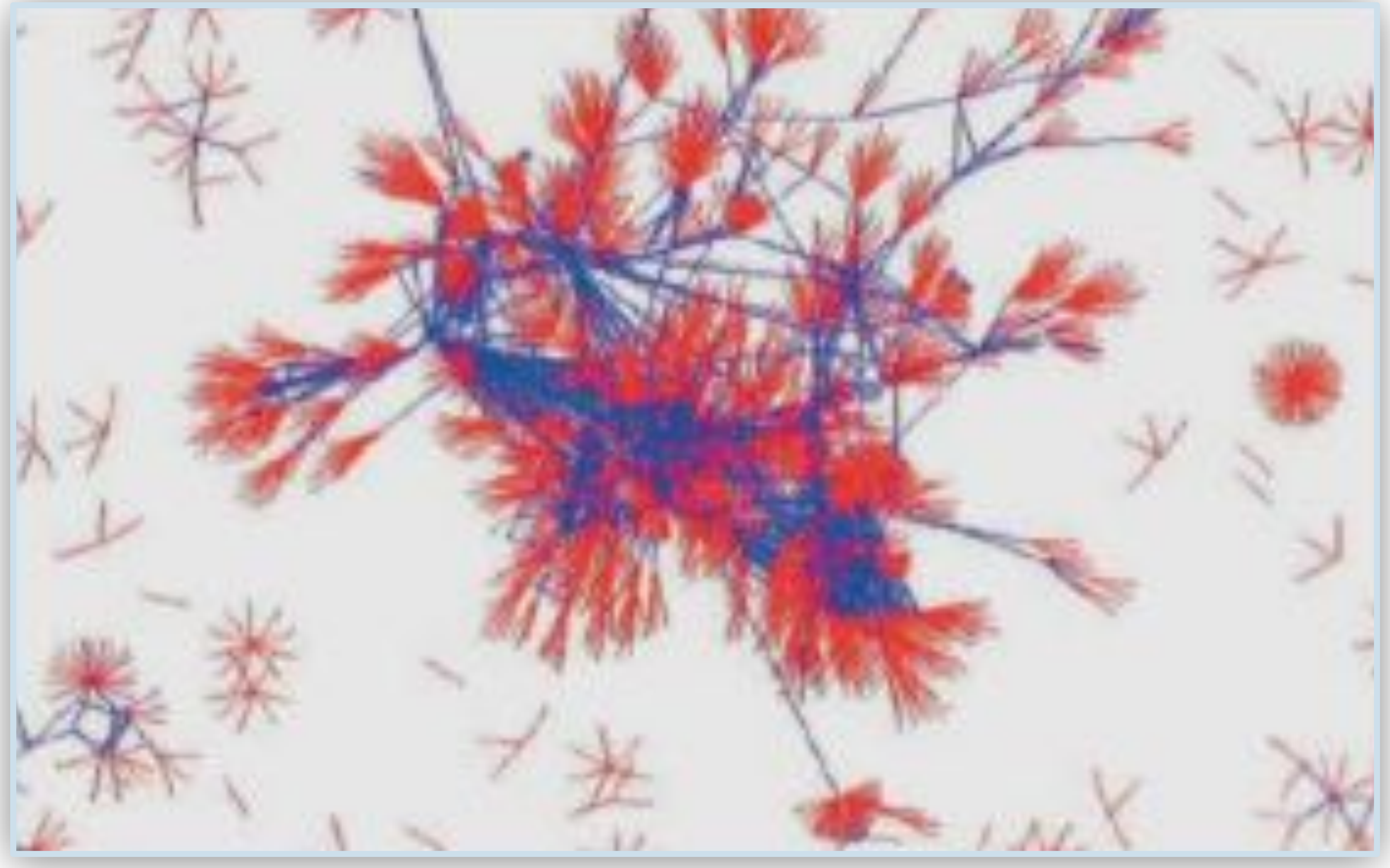
Source

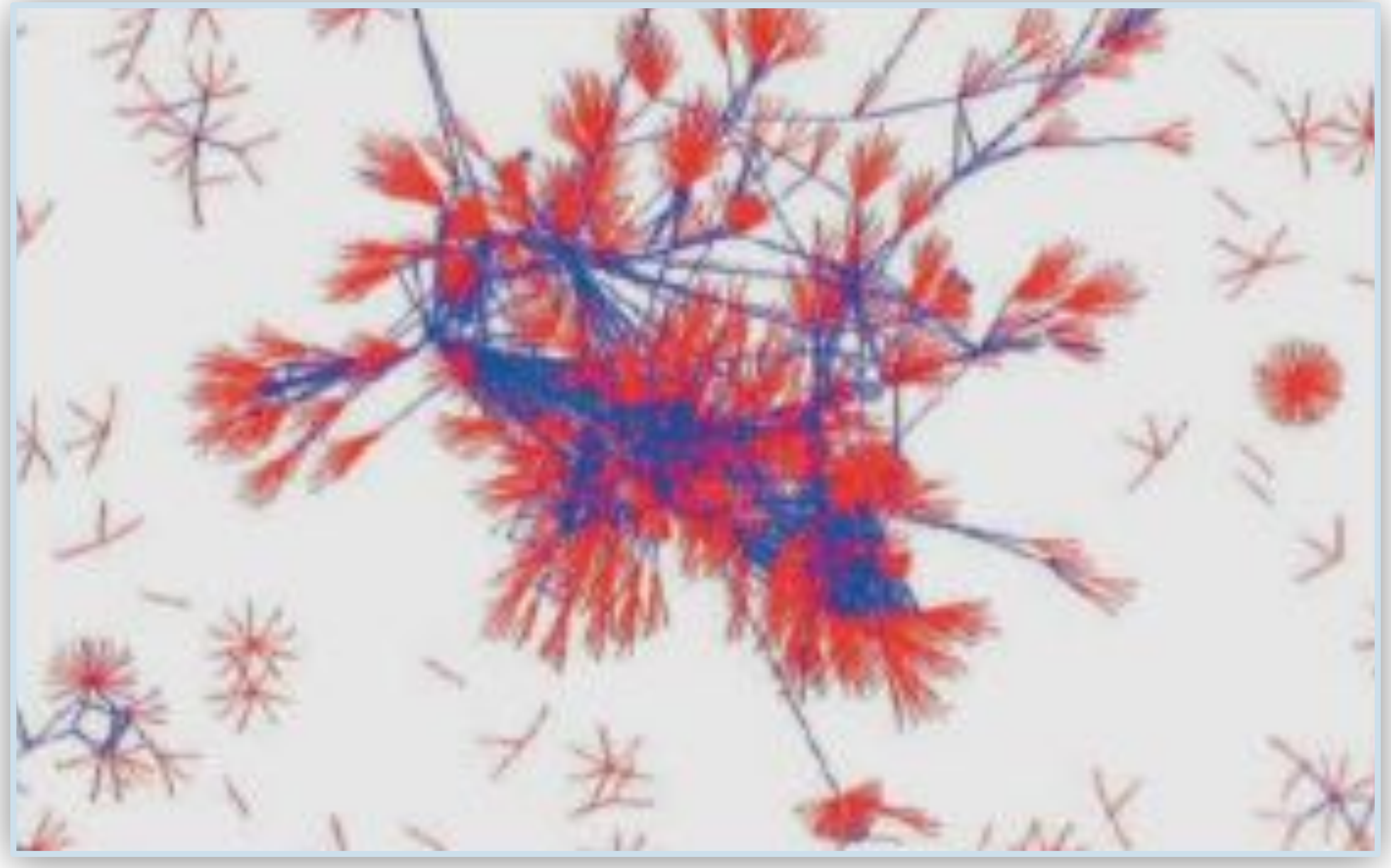
Heading1

Body1

Heading1

Body1





Heading1

- Item1
- Item2
- Item3
- Item4

Register now

URL