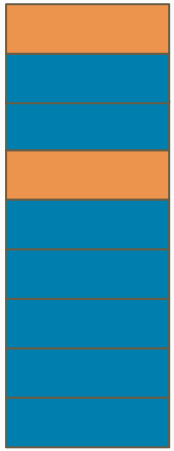
Machine learning with imbalanced clinical data sets

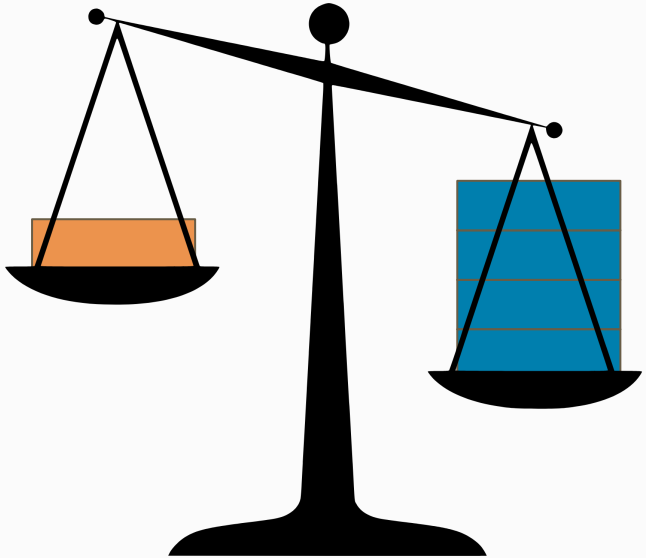
21st January, 2019

Speaker: Víctor Vicente Palacios
Clinical Data Scientist

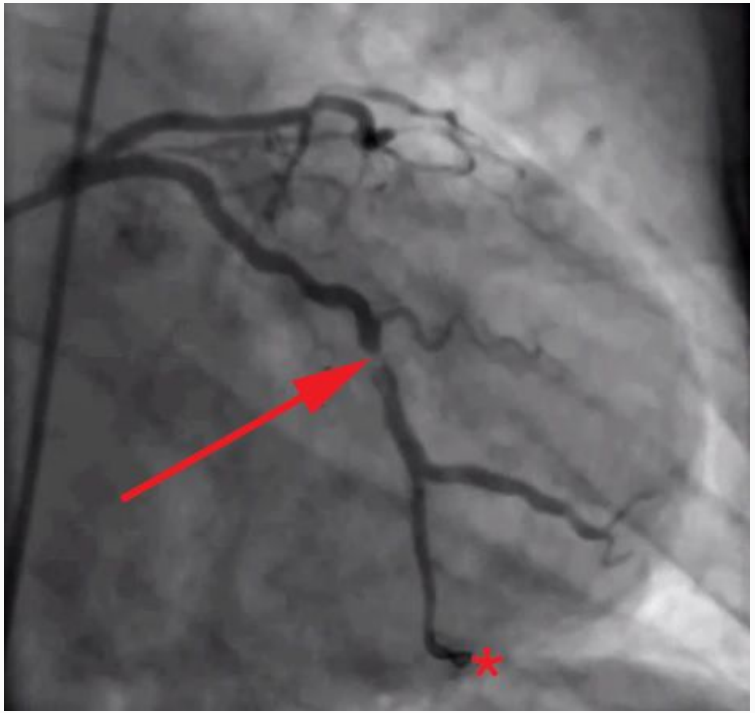
PyData Salamanca meetup

Imbalanced?

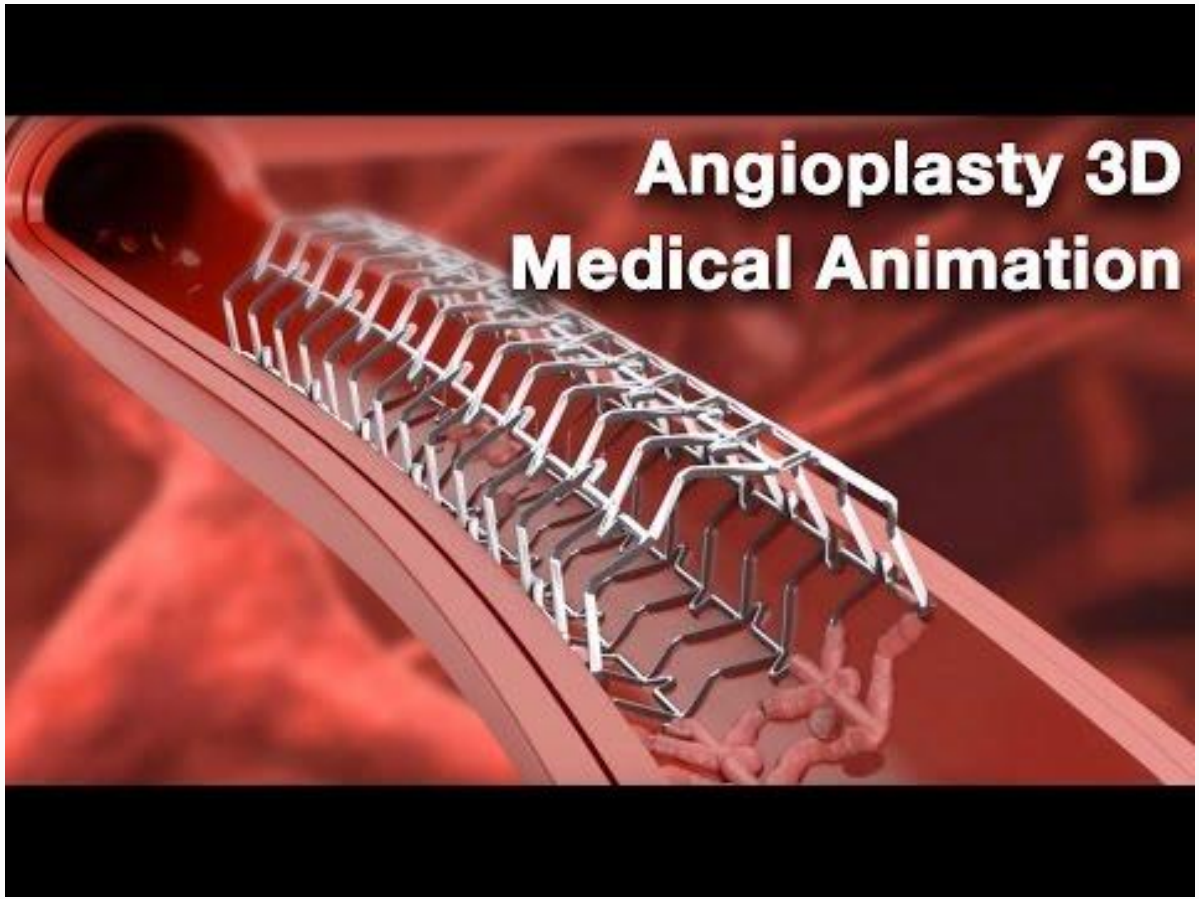




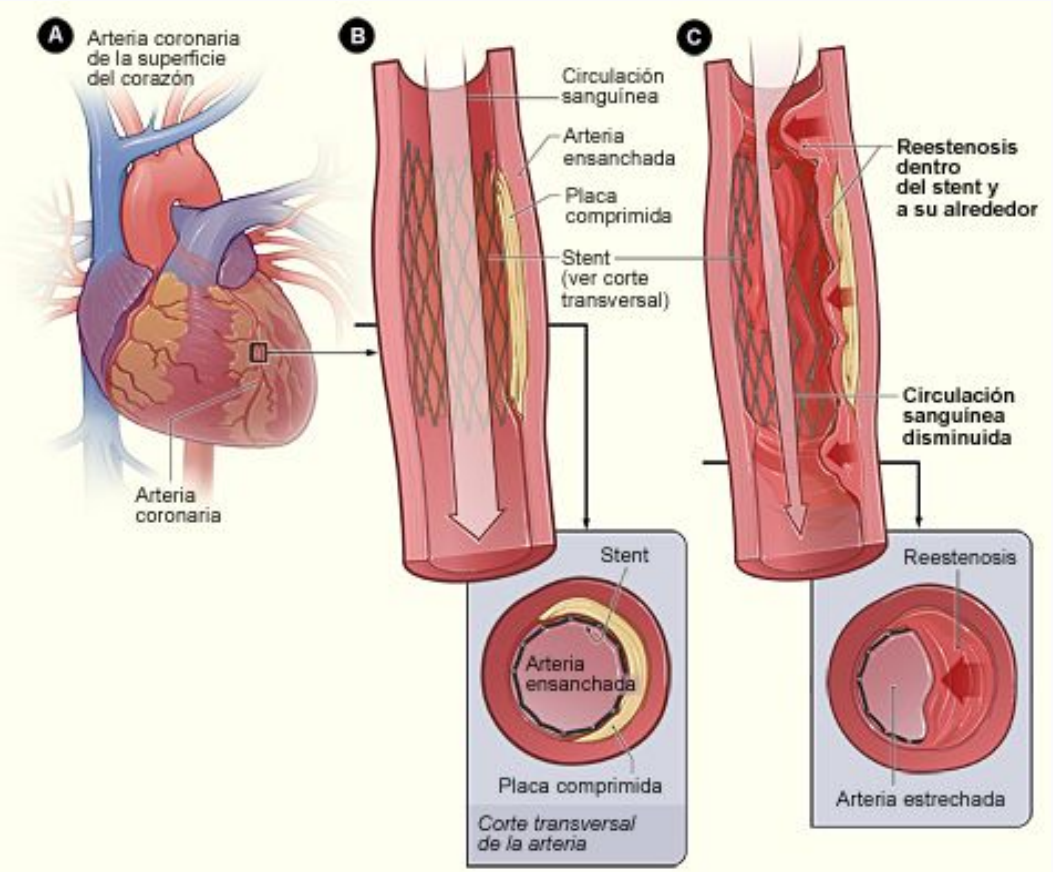
Caso clínico



Implantación stent



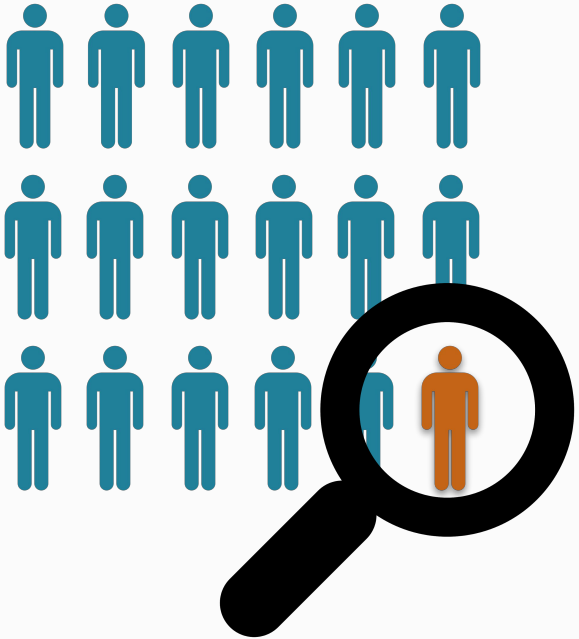
Reestenosis



Casos clínicos

Coronografía

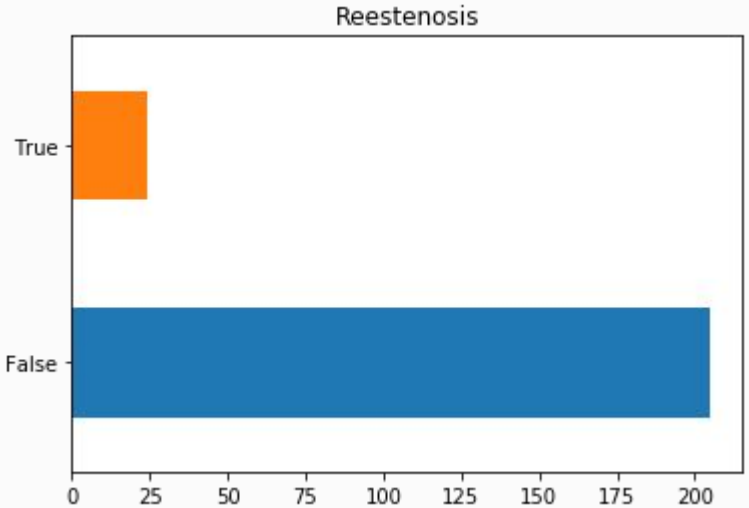
Historia clínica



Ensayo Clínico

	stent	tirofiban	sexo	diabetes	hipertension	his_familiar	edad	tabaquismo	localiz	seg_me_pre	...	tmpg_post	timi_12m
0	1	2	1	1	2.0	2	63	2	RCA	RCADistal	...	3.0	3.0
1	2	2	1	1	1.0	2	78	1	RCA	RCADistal	...	0.0	3.0
2	2	1	1	2	2.0	2	51	1	LAD	LADMid	...	3.0	3.0
3	2	1	2	2	1.0	2	70	1	LAD	LADMid	...	0.0	3.0
4	1	2	1	1	2.0	2	64	2	LAD	LADProximal	...	NaN	NaN

- Pacientes con reestenosis **10%**
- Medidas angiográficas
- Historial clínico



imbalanced-learn

imbalanced-learn

stable

Search docs

Install and contribution

Prerequisites

The imbalanced-learn package requires the following dependencies:

- numpy (>=1.8.2)
- scipy (>=0.13.3)
- scikit-learn (>=0.20)
- keras 2 (optional)
- tensorflow (optional)

Our release policy is to follow the scikit-learn releases in order to synchronize the new feature. **imbalanced-learn 0.4 is the last version to support Python 2.7**

Install

imbalanced-learn is currently available on the PyPi's reporitories and you can install it via *pip*:

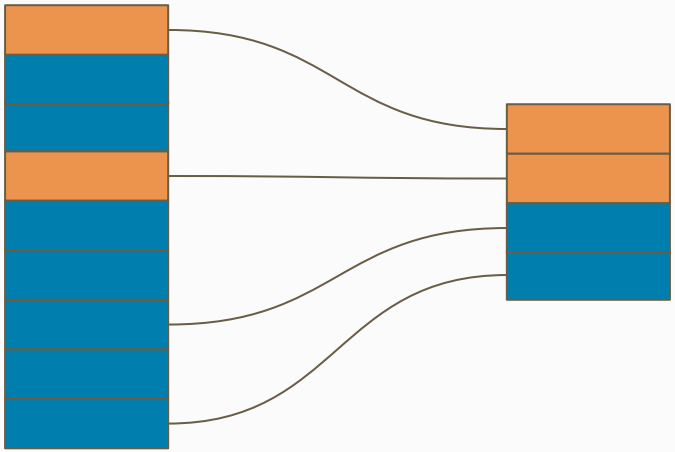
```
pip install -U imbalanced-learn
```

imbalanced-learn.org

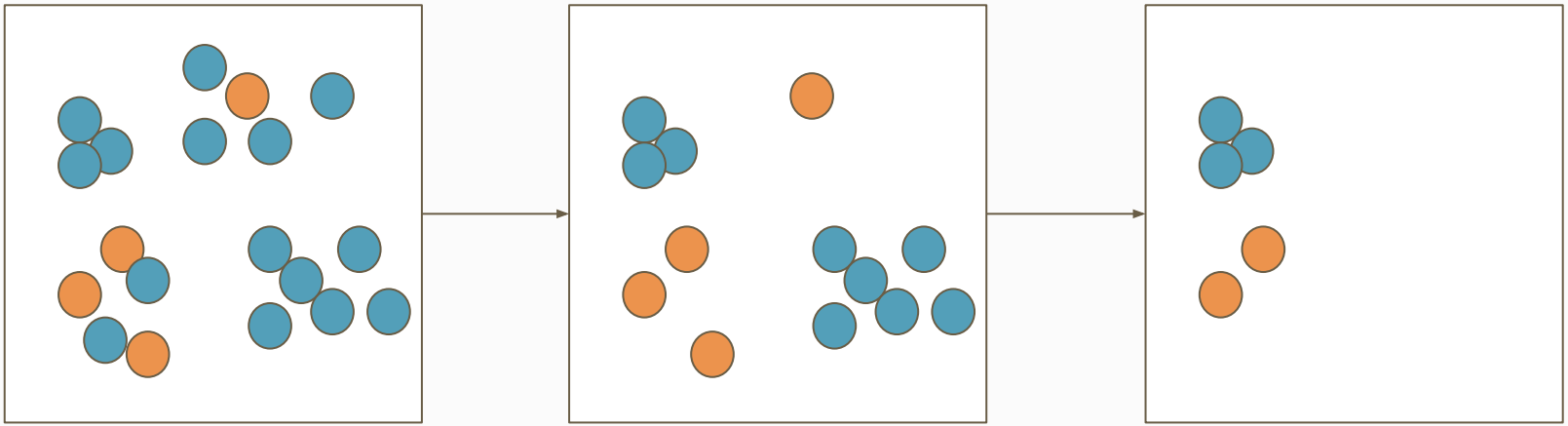
Técnicas de remuestreo

- Undersampling
- Oversampling
- Combinada
- Ensemble

Undersampling



Undersampling (Edited Nearest Neighbours)



imbalanced-learn.org

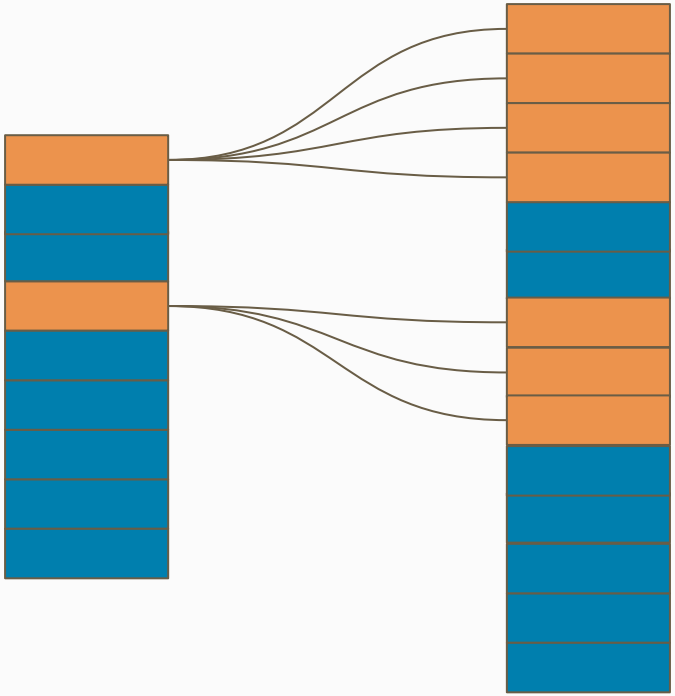
Undersampling (Edited Nearest Neighbours)

```
from sklearn.ensemble import RandomForestClassifier
from imblearn.under_sampling import EditedNearestNeighbours
from imblearn.pipeline import make_pipeline

rf = RandomForestClassifier(random_state=42)

enn = EditedNearestNeighbours(random_state=42)
pipe_enn = make_pipeline(enn, rf)

pipe_enn.fit(X_train, y_train)
y_pred = pipe_enn.predict(X_test)
```

Resultados

	ACCURACY	SPECIFICITY	SENSITIVITY
RF	0.9	0.99	0.07
RF + ENN	0.9	0.98	0.14

Oversampling



Oversampling (Synthetic Minority Oversampling Technique)

imbalanced-learn.org

Oversampling (Synthetic Minority Oversampling Technique)

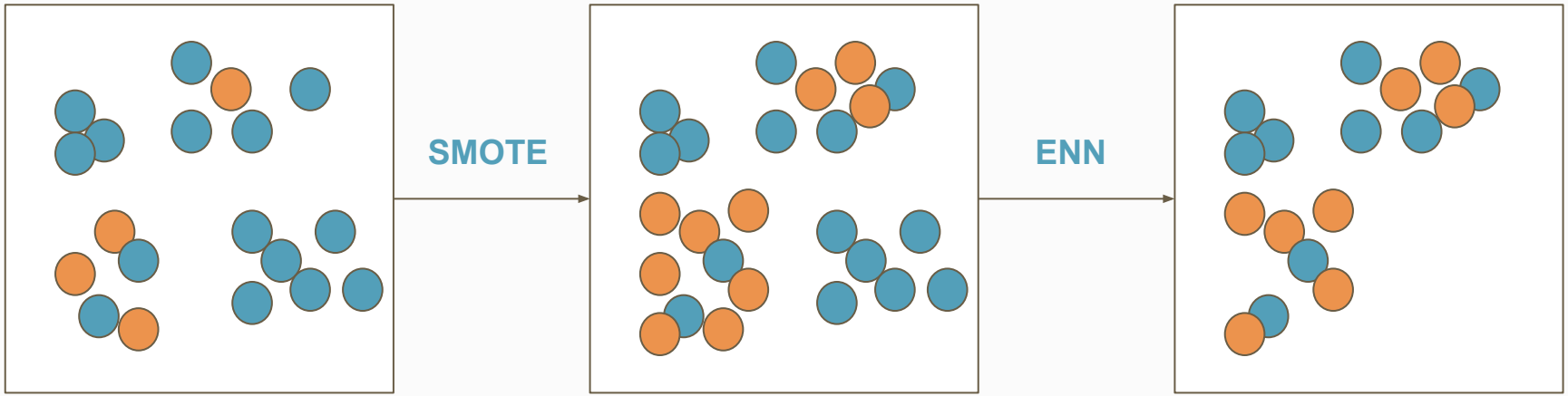
```
from sklearn.ensemble import RandomForestClassifier
from imblearn.over_sampling import SMOTE
from imblearn.pipeline import make_pipeline

rf = RandomForestClassifier(random_state=42)

smt = SMOTE(random_state=42)
pipe_smt = make_pipeline(smt, rf)
pipe_smt.fit(X_train, y_train)
y_pred = pipe_smt.predict(X_test)
```

imbalanced-learn.org

Resultados

	ACCURACY	SPECIFICITY	SENSITIVITY
RF	0.9	0.99	0.07
RF + ENN	0.9	0.98	0.14
RF + SMOTE	0.9	0.97	0.2

Over+Undersampling (SMOTE + ENN)



imbalanced-learn.org

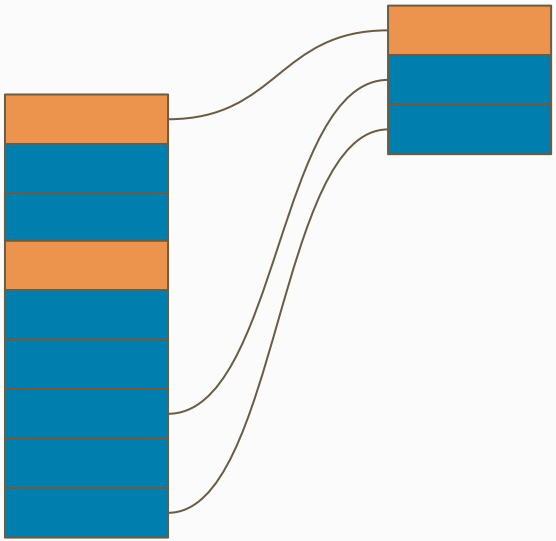
Over+Undersampling (SMOTE + ENN)

```
from sklearn.ensemble import RandomForestClassifier
from imblearn.combine import SMOTEENN
from imblearn.pipeline import make_pipeline

rf = RandomForestClassifier(random_state=42)

smtenn = SMOTEENN(random_state=42)
pipe_smtenn = make_pipeline(smtenn, rf)
pipe_smtenn.fit(X_train, y_train)
y_pred = pipe_smtenn.predict(X_test)
```

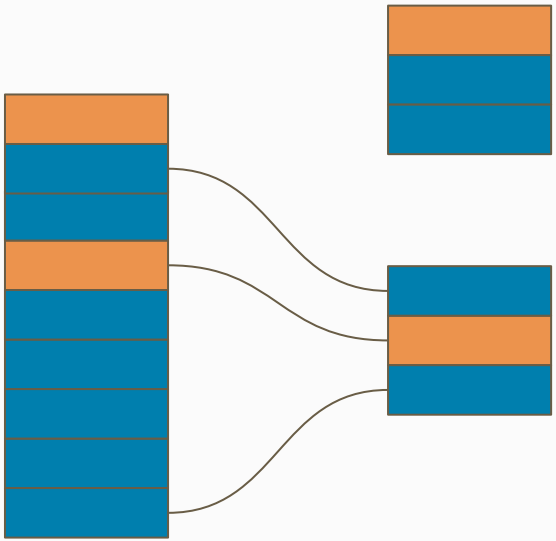
Resultados

	ACCURACY	SPECIFICITY	SENSITIVITY
RF	0.9	0.99	0.07
RF + ENN	0.9	0.98	0.14
RF + SMOTE	0.9	0.97	0.2
RF + SMOTEENN	0.87	0.94	0.3

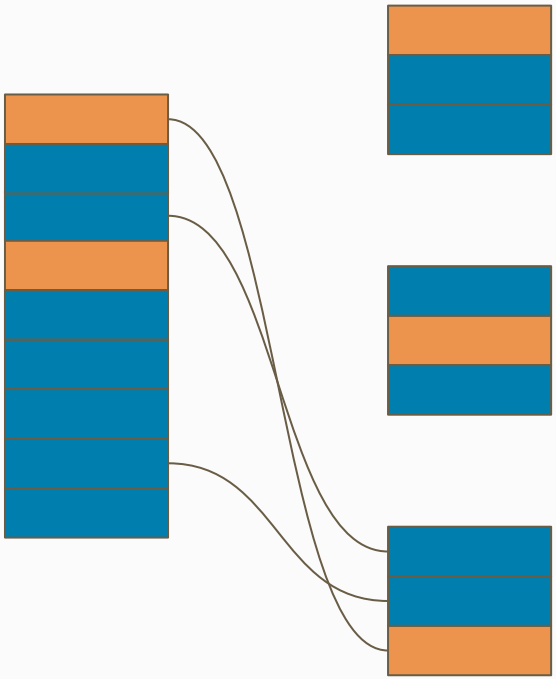
Balanced Bagging Classifier



Balanced Bagging Classifier

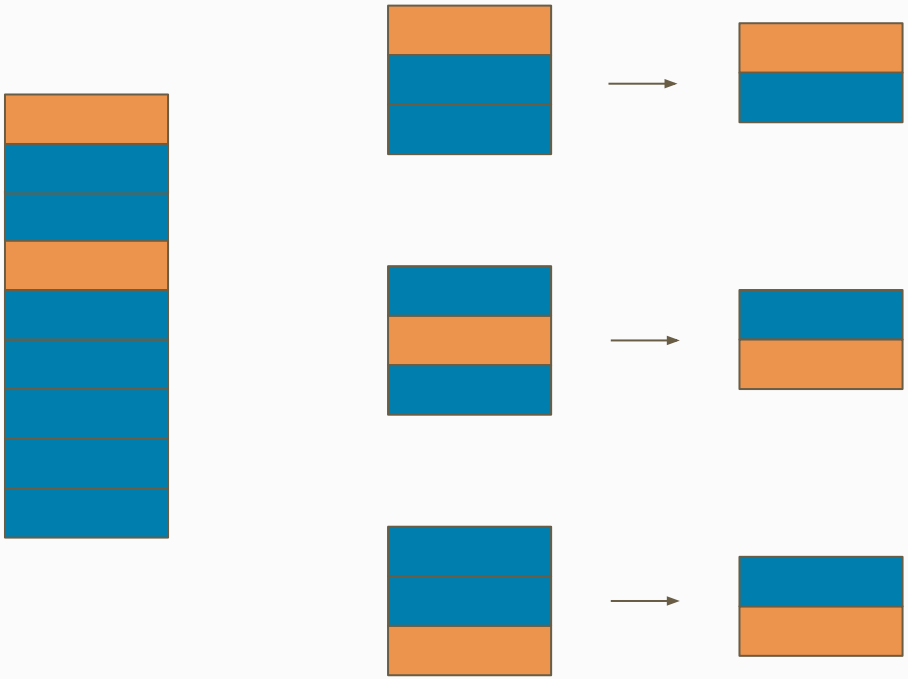


Balanced Bagging Classifier



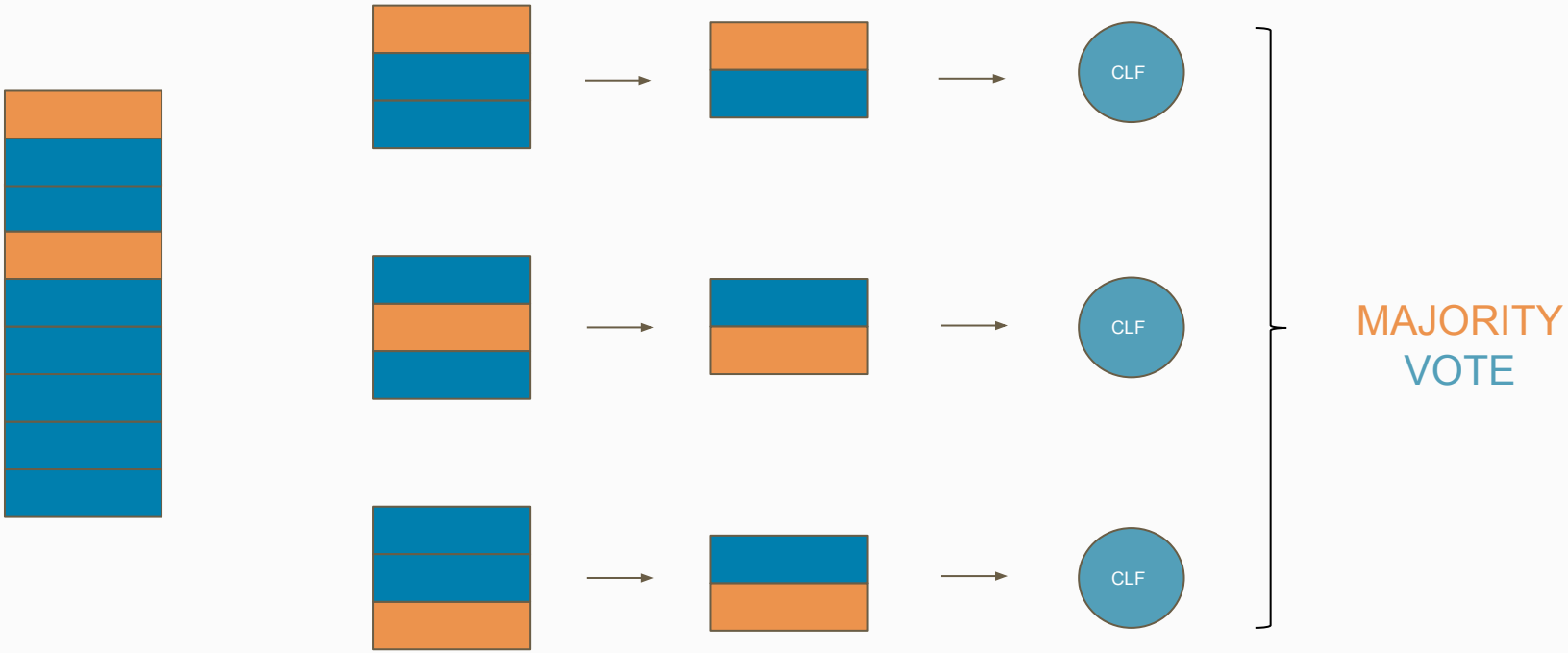
imbalanced-learn.org

Balanced Bagging Classifier



imbalanced-learn.org

Balanced Bagging Classifier



imbalanced-learn.org

Balanced Bagging Classifier

```
from sklearn.ensemble import RandomForestClassifier
from imblearn.ensemble import BalancedBaggingClassifier
from imblearn.pipeline import make_pipeline

rf = RandomForestClassifier(random_state=42)

bbc = BalancedBaggingClassifier(rf)
bbc.fit(X_train, y_train)
y_pred = bbc.predict(X_test)
```

Resultados

	ACCURACY	SPECIFICITY	SENSITIVITY
RF	0.9	0.99	0.07
RF + ENN	0.9	0.98	0.14
RF + SMOTE	0.9	0.97	0.2
RF + SMOTEENN	0.87	0.94	0.3
RF + BBC	0.7	0.7	**0.7**

Thanks to all attendees! and sponsors!

PyData Salamanca meetup will come back on 21st March 2019

Natural Language Processing

Martina Kienberger
(PhD Candidate University of Wien)

Ana Valdivia
(PhD University of Granada)

NUMFOCUS
OPEN CODE = BETTER SCIENCE

MEDIALAB USAL