



STAATLICH  
ANERKANNTE  
HOCHSCHULE

# Big Data Programming-2 Project

## Adding new functionality to Python Package: Scikit-Learn

Presented on 01.04.2020 by:

Shekhar Singh (11011694)

Sanika Medankar (1011861)

Rishabh Garg (11011875)

Rohit Keshav Bewoor (11011831)

Students of Big Data and Business Analytics 2018-20 batch  
SRH Hochschule Heidelberg

# Content

- Introduction and Motivation
- Logic
  - Explanation
  - Modules impacted
- Issues and Workaround
- Pull Request
- Environment setup package uploaded to Test Pypi
- Demo
- Q&A

# Introduction

- Popular Machine Learning packages: SCIKIT-LEARN and STATSMODELS
- About Scikit-learn package
  - Source Code (PyPI): <https://pypi.org/project/scikit-learn/#files>
  - For Machine Learning, built on top of Scipy
  - Website: <http://scikit-learn.org>
- In Big Data 1 project, we choose to dissect Scikit-learn's ***train\_test\_split*** function
- Big Data 2 project:
  - Currently Functionality:
    - `train_test_split` is used to split the input arrays into two subsets, usually called "Train" and "Test".
  - New functionality implemented:
    - Allow splitting the input arrays into three subsets: "Train", "Test" and ***an additional "Validation"***.
- Issue created on Github of Scikit-learn: <https://github.com/scikit-learn/scikit-learn/issues/13990>

Source: <https://pypi.org/project/scikit-learn/>

### Statistics

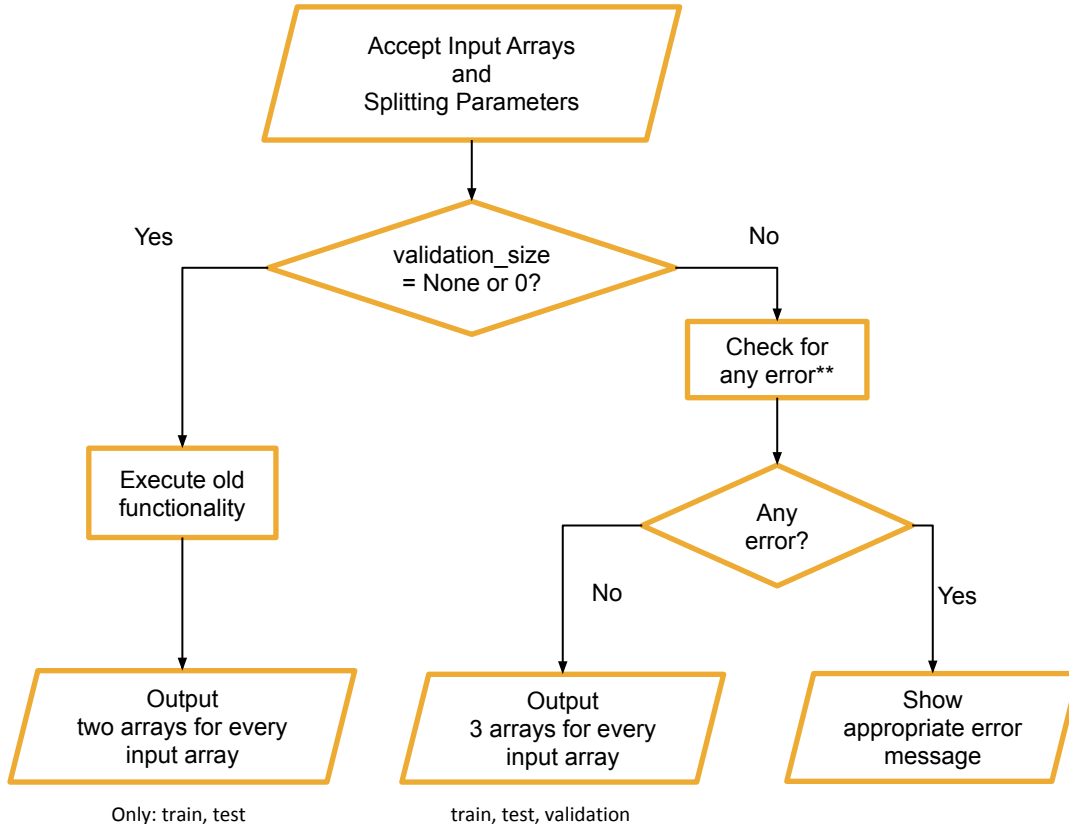
GitHub statistics:

- ★ Stars: 39,875
- 🔗 Forks: 19,393
- 📢 Open issues/PRs: 2,182

# Motivation

- Machine learning models are trained on “Train” set. Then, the “Test” set was used for model accuracy.
- Now, during training, a “Validation” set (aka. “Dev” set) is used at end of each epoch to track model accuracy during training.
- *Only after training*, Test set (unseen data by model) used for final model accuracy.
- Currently, Scikit-learn allows only two-way splitting. ***User needs to call function twice*** to:
  - Split Full data into Train and Intermediate sets
  - Split Intermediate set into Validation and Test sets
  - User needs to track the number of samples (or percentage of split) manually and this is prone to errors
- New approach:
  - ***Existing function updated to split three ways***
  - No chance of manual calculation errors
  - Maintained Backward compatibility to allow two-way split

# Logic - Flowchart



## \*\*Error Checks:

1.  $\text{validation\_size} + \text{train\_size} + \text{test\_size} \leq \text{n\_samples}$  ( or  $\leq 1.0$  if proportion)
2.  $\text{validation\_size}$  must be an Integer or Float
3. If integer; ensure  $0 \leq \text{validation\_size} < \text{n\_samples}$
4. If float; ensure  $0.0 \leq \text{validation\_size} < 1.0$

## Default Value of *validation\_size*= 0.0

- Allows existing functionality for two-way split

# Logic - Truth Table

Optional parameters during the function call to `train_test_split()`:  
`train_size`, `test_size`, `validation_size`, `shuffle`, `stratify`

Acceptance for processing and the actions taken in each scenario.

SN#	Input Parameters			Expected outcome and Actions				
	train_size	test_size	validation_size	Error?	train_size	test_size	validation_size	Comments
1	Not Specified	Not Specified	Not Specified	No	0.75	0.25	0	Defaults per existing functionality and will return only Train and Test sets.
2	0.35	Not Specified	Not Specified	No	0.35	0.65	0	
3	Not Specified	0.35	Not Specified	No	0.65	0.35	0	
4	0.1	0.2	Not Specified	No	0.1	0.2	0	Proportion total < 1.0
5	0.1	0.2	0.3	No	0.1	0.2	0.3	Proportion total < 1.0
6	Not Specified	Not Specified	0.2	No	0.55	0.25	0.2	Test set to default 0.25. Train = complement(Validation + Test).
7	Not Specified	Not Specified	0.8	Yes	NA	NA	NA	Will first attempt to set the Test Size as 0.25 by default and then fails with error message as the total proportion has crossed 1.0
8	0.6	Not Specified	0.1	No	0.6	0.3	0.1	Test = complement(Train + Validation).
9	Not Specified	0.6	0.1	No	0.3	0.6	0.1	Train = complement(Test + Validation).

Default values: Test = 0.25,  
 Validation = 0.0,  
 Train = complement(Test + Validation).

=> **If no values specified:** then original functionality where `train_size` = 0.75 and `test_size` = 0.25 and no validation set is created.

Same results can be achieved with or without stratify

# Impacted modules

- The following modules are changed for this project:
  - sklearn / model\_selection / \_split.py
    - Logic changes here
  - sklearn / model\_selection / tests / test\_split.py
    - Changes to test new functionality
  - sklearn / utils / fixes.py
    - Small change to address import error for “comb”.
  - sklearn / base.py
    - Small change to handle longer messages (100 instead of 75 characters)


# Issues Faced and Workaround


- Extensive testing for all scenarios done using our own test script:
  - Script location:  
[https://github.com/rbewoor/BigData2\\_Project\\_Bkup/blob/master/Dedicated\\_Test\\_Cases/Test\\_Cases\\_SciKit-Learn\\_fromTestPypi.ipynb](https://github.com/rbewoor/BigData2_Project_Bkup/blob/master/Dedicated_Test_Cases/Test_Cases_SciKit-Learn_fromTestPypi.ipynb)
- Test Pypi does not allow a simple linux wheel to be uploaded. Possibly existing issue:
  - Binary wheels for linux are not supported #120 ( <https://github.com/pypa/pypi-legacy/issues/120> )
  - Recommend manylinux wheels in the Error 400 response for "linux" package uploads #6545 ( <https://github.com/pypa/warehouse/issues/6545> )
- Supposedly, a many-linux wheel version can be uploaded. But we are unable to create it.
- **Therefore, only uploaded the source and not a wheel (i.e. output of sdist and not bdist\_wheel).**




# Pull Request






- Faced many issues with creating a pull request (indentation, circleCI checks, azure pipeline checks). Fixed as much as possible.
- Final pull request: Updating to allow 3-way split using `train_test_split` function.


 **Updating to allow 3-way split using `train_test_split` function #16781**  
shekharsingh8811 wants to merge 50 commits into `scikit-learn:master` from `rbewoor:master`



 **Some checks were not successful** [Hide all checks](#)

10 failing, 2 neutral, and 7 successful checks

✓	 <b>LGTM analysis: Python</b> Successful in 17m — 2 new alerts <a href="#">Details</a>
✓	 <b>ci/circleci: deploy</b> — Your tests passed on CircleCI! <a href="#">Details</a>
✓	 <b>ci/circleci: doc</b> — Your tests passed on CircleCI! <a href="#">Details</a>
✓	 <b>ci/circleci: doc artifact</b> — Link to 0/doc/_changed.html <a href="#">Details</a>
✓	 <b>ci/circleci: doc-min-dependencies</b> — Your tests passed on CircleCI! <a href="#">Details</a>

 **This branch has no conflicts with the base branch**  
Only those with [write access](#) to this repository can merge pull requests.

Pull request link: <https://github.com/scikit-learn/scikit-learn/pull/16781>

# Environment Setup from Test PyPi

- Test PyPi details:
  - Project: Scikit-learn-VAL-TestPyPi, Latest Version: 0.0.2
  - Link: <https://test.pypi.org/project/scikit-learn-VAL-TestPyPi/>
- Automatic install of dependencies not working, so manual install required
- Environment from Test PyPi - build with source option only
  - Automatic install of dependencies not working; so **manual install required**:
    - > Cython>=0.28.5, setuptools, wheel, numpy>=1.14.0, scipy>=1.1.0, joblib>=2.0.0, threadpoolctl>=2.0.0
  - Now install from Test PyPi:
    - > pip3 install -i https://test.pypi.org/simple/ scikit-learn-VAL-TestPyPi
  - Additional packages for our test script:  
[https://github.com/rbewoor/BigData2\\_Project\\_Bkup/blob/master/Dedicated\\_Test\\_Cases/Test\\_Cases\\_SciKit-Learn\\_fromTestPyPi.ipynb](https://github.com/rbewoor/BigData2_Project_Bkup/blob/master/Dedicated_Test_Cases/Test_Cases_SciKit-Learn_fromTestPyPi.ipynb)
    - > Jupyter, pandas

# Demo

- Various test cases for combinations of splitting variable values:
  - Expecting only Train+Test vs. Train+Test+Validation
  - With and without use Shuffle
  - With and without Stratification

# Q&A

- Thank you. Open to questions!