

Project Title

**P04 Binary Classification Model development for Sonar
measurement (mines vs rocks detection) data
(Python 3.x and Ubuntu 16.04)**

Developed by: MSc. Shiva Agrawal

Place: Germany

Date: September 2018

Content

Project Title	1
Content	2
1. Introduction	3
1.1. Project description	3
1.2. Outline.....	3
2. Python packages and datasets.....	4
2.1. Python Packages	4
2.1.1. Scikit-learn.....	4
2.1.2. Pandas.....	4
2.1.3. Numpy.....	4
2.1.4. Matplotlib.....	4
2.2. Sonar Mines vs Rocks Dataset	4
3. ML - Predictive Model Development (Binary Classification)	5
4. Results.....	5
5. Conclusion.....	5
6. References	6

1. Introduction

1.1. Project description

The aim of the project is to develop the predictive model for the Sonar sensor's mines vs rocks detection. This dataset can be downloaded from [1]. This is binary classification type of Machine learning model. In this project, **python scikit-learn package** is used for the machine learning algorithms and data preprocessing and **Pandas** to import the dataset into python environment in form of **dataframe**.

The project is developed by referring the book Mastering Machine Learning using Python by Jason Brownlee [2], scikit-learn official website, Prof. Andrew Ng videos of machine learning and other online references.

1.2. Outline

- Chapter 1: Project description and outline of the project report
- Chapter 2: Python packages information and Sonar Mines vs Rocks Dataset
- Chapter 3: Predictive model development
- Chapter 4: Results
- Chapter 5: Conclusion
- Chapter 6: References

2. Python packages and datasets

2.1. Python Packages

2.1.1. Scikit-learn

[3] This is open source package available for Machine Learning in Python. It is built on python other packages like numpy, scipy and matplotlib. It contains most of the required functions and tools to preprocess, analyze and develop the ML models.

2.1.2. Pandas

[4] It is an open source, BSD License library providing high performance, easy to use data structure and data analysis tool for the python programming language.

2.1.3. Numpy

[5] It is the fundamental package for scientific computing in python. It has powerful N dimension array object, sophisticated broadcasting functions and useful linear algebra, Fourier transform, and random number capabilities.

2.1.4. Matplotlib

[6] Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. It can generate plots, histograms, power spectra, bar charts, error charts, scatterplots, etc.

2.2. Sonar Mines vs Rocks Dataset

Dataset Information:

Name: Sonar metal vs rocks

Samples: 208

Features: 60 (all numeric)

output: 1 (M or R) - (Metal or Rock)

In this dataset, as all the input features are just signal measurements, they do not have some specific name. So, there is no header used specifically. Hence for the column, only numbers define them.

Hence from column 1 to 60 are input features and column 61 is output class .

3. ML - Predictive Model Development (Binary Classification)

The source code of the model and dataset are available in src folder of the repository. The source file contains well commented steps of the model development process. The model is saved using the pickle package of python.

During the model development, at first the raw features were used and 6 algorithms were implemented and compared.

After comparison it was founded that the distribution of the output is not uniform due to different scales in the features. Hence at first rescaling of the data is done and then checked which was also not appropriate. Later Standardization of the data is done and it provided proper scaling and distribution of the features.

Hence in the source code only standardization is implemented, but the resultant comparison of graphs using box and whisker plots are generated for

1. Without preprocessing
2. With Rescaling
3. With Standardization

All these plots are available in results folder.

The complete model is developed as one function and hence a separate test.py is used to call the function and to generate the results. It is also available in src folder.

4. Results

The developed model is saved as SVM_model.sav inside results folder of the project repository. The same folder also contains the generated plots and result.txt file which contains the output from all the steps of the model development process.

5. Conclusion

Binary classification / Logistic regression model for the Sonar measurement for rocks and mines detection dataset is developed and tested using different machine learning algorithms. After comparison and validation, it is found that SVM (Support Vector Machine Classifier) algorithm fits best for the problem. Hence the model is finally trained with SVM and saved for future use.

6. References

- [1] "Sonar Mines vs rocks dataset UCI," [Online]. Available: [http://archive.ics.uci.edu/ml/datasets/connectionist+bench+\(sonar,+mines+vs.+rocks\)](http://archive.ics.uci.edu/ml/datasets/connectionist+bench+(sonar,+mines+vs.+rocks)).
- [2] J. Brownlee, Mastering Machine Learning using Python.
- [3] "Home page," [Online]. Available: <http://scikit-learn.org/stable/>.
- [4] "pandas," [Online]. Available: <https://pandas.pydata.org/>.
- [5] "Numpy," [Online]. Available: <http://www.numpy.org/>.
- [6] "Matplotlib," [Online]. Available: <https://matplotlib.org/>.
- [7] "scipy," [Online]. Available: <https://www.scipy.org/>.