**Scikit-learn models with Flask**

[Scikit-learn](http://scikit-learn.org/) is a Python library which provides simple and efficient tools for data mining and data analysis. Scikit-learn has the following major modules:

* + Clustering
  + Regression
  + Classification
  + Dimensionality Reduction
  + Model selection
  + Preprocessing

(Be sure to check DataCamp's [Supervised Learning with scikit-learn](https://www.datacamp.com/courses/supervised-learning-with-scikit-learn?tap_a=5644-dce66f&tap_s=357540-5b28dd) course which is taught by the core developer of scikit-learn - Andreas Müller)

* Scikit-learn provides the support of serialization and de-serialization of the models that you train using scikit-learn. This saves you the time to retrain a model. With a serialized copy of your model made using scikit-learn you can write a Flask API.
* Scikit-learn models require the data to be in numerical format. That is why, if the dataset contains categorical features that are non-numeric, it is important to convert them into numeric ones. For this transformation, scikit-learn provides utilities like LabelEncoder, OneHotEncoder, etc. These can be found in sklearn.preprocessing module.
* Scikit-learn models cannot handle missing values implicitly. You need to handle missing values in your dataset by yourself, and then you can feed it to your model. For handling missing values, scikit-learn provides a wide range of utilities which can be found from sklearn.preprocessing module.

Label encoding and missing values are important data preprocessing steps which are very essential for building a good machine learning model. If you want to learn more on this, be sure to check the following course offered by DataCamp:

* [Preprocessing for Machine Learning in Python](https://www.datacamp.com/courses/preprocessing-for-machine-learning-in-python)

For this tutorial, you will use the Titanic dataset which is one of the most popular datasets for many reasons such as - the dataset contains a well great different types of variables, and the dataset contains missing values, etc. This [DataCamp tutorial](https://www.datacamp.com/community/tutorials/k-means-clustering-python) covers an excellent analysis of the dataset, and the dataset can be downloaded from [here](https://www.kaggle.com/c/titanic/data).

This dataset deals with a classification problem of predicting if a passenger would survive or not given some information about him/her.

**Note**: Variables and Features these terms are used interchangeably at many times in this tutorial.

To simplify things even further, you will only use four variables: age, sex, embarked, and survived where survived is the class label.