Machine Learning: Regression

In this lab exercise you will perform supervised machine learning with Spark. The exercises are split into the following parts:

- Exercise 1: Logistic Regression
- Exercise 2: Verifying Alarms

Exercise 1: Logistic Regression

In this exercise, you perform a machine learning analysis using logistic regression. Given a training set and a test set, you will evaluate the performance using MLlib's provided functionality (data set: Iris_modified6_5mod.csv). The exercise demonstrates that MLlib enables a whole machine learning pipeline with only a few lines of code.

Note: The first column is the label, the columns a0 and a1 are IDs, the remaining columns are features

The instructions for this exercise can be found directly in the notebook:
LogisticRegression.ipynb

Exercise 2: Verifying Alarms

In this exercise, you experiment with logistic regression to verify alarms provided by the London Firebrigade (data set: LFB Incident data from January 2017.csv). You can approach the problem as follows:

- Analyze the data set.
- Load it into a DataFrame.
- Use logistic regression to predict (verify) the outcome of the information stored in the **column "Incident Group".**
- Apply logistic regression to see if you can verify true/false alarms.
- Experiment with different parameters.
- Analyze the performance of the models.

The instructions for this exercise can be found directly in the notebook:
LogisticRegressionLFD_1File.ipynb

Note: **Logistic Regression can only be used with numerical features, therefore we need to first index columns that are of String type in order to be able to use them**. For this purpose, use a StringIndexer for every String feature. Study the example below and apply it accordingly to your data set.

This line is an example of how to construct a string indexer on the Postcode_district column:

```
indexer_ZipCode = StringIndexer(inputCol="Postcode_district",
                                outputCol="Postcode_district_indexed")
```

To apply the transformation, you can use the following code:

```
df = indexer_ZipCode.fit(df).transform(df).drop("Postcode_district")
```