**Problem statements**

**TensorFlow – Classification – Image identification**

1. Classify the test images of architectural portions of historical monuments. 10 categories: Altar: 829 images; Apse: 514 images; Bell tower: 1059 images; Column: 1919 images; Dome (inner): 616 images; Dome (outer): 1177 images; Flying buttress: 407 images; Gargoyle (and Chimera): 1571 images; Stained glass: 1033 images; Vault: 1110 images.

[Dataset.](Image%20classification/Intel%20contest/Sea,%20forest,%20archive.zip)

1. Classify the test images of all the six categories, namely sea, mountain, forest, glacier, street, buildings in the same order.

[Dataset](Image%20classification/Intel%20contest/Architectural_Heritage_Elements_Dataset_128(creative_commons).zip).

**Logistic regression – Scikit – Sentiment Analysis**

1. A sentiment analysis model to classify movie reviews as positive or negative (binary classifier), based on the text of the review. This is an example of binary—or two-class—classification, an important and widely applicable kind of machine learning problem.

These are split into 25,000 reviews for training and 25,000 reviews for testing. The training and testing sets are balanced, meaning they contain an equal number of positive and negative reviews.

[Dataset.](Sentiment%20analysis/aclImdb_v1.tar.gz)

**Linear regression – Supervised - Multi-class classification on Stack Overflow questions – Sentiment analysis**

1. A multi-class classifier to predict the tag of a programming question on Stack Overflow.

A dataset has been prepared for you to use containing the body of several thousand programming questions (for example, "How can I sort a dictionary by value in Python?") posted to Stack Overflow. Each of these is labeled with exactly one tag (either Python, CSharp, JavaScript, or Java). Your task is to take a question as input, and predict the appropriate tag, in this case, Python.

The dataset you will work with contains several thousand questions extracted from the much larger public Stack Overflow dataset on BigQuery, which contains more than 17 million posts.

Given a new complaint comes in, we want to assign it to one of 12 categories. The classifier makes the assumption that each new complaint is assigned to one and only one category. This is multi-class text classification problem.

[Dataset.](Sentiment%20analysis/complaints.csv.zip)

1. **Classifying Consumer Finance Complaints into 12 pre-defined classes.**

The problem is supervised text classification problem. Given a new complaint comes in, we want to assign it to one of 12 categories. The classifier makes the assumption that each new complaint is assigned to one and only one category. This is multi-class text classification problem.

[Dataset.](Sentiment%20analysis/complaints.csv.zip)

**Naïve Bayes classifier – Supervised – Classifying consumer complaints – Sentiment analysis**

1. **Classifying Consumer Finance Complaints into 12 pre-defined classes.**

The problem is supervised text classification problem. Given a new complaint comes in, we want to assign it to one of 12 categories. The classifier makes the assumption that each new complaint is assigned to one and only one category. This is multi-class text classification problem.

[Dataset.](Sentiment%20analysis/complaints.csv.zip)

**K-Means clustering – Customer segmentation – Unsupervised – Grouping customers based on age, products purchased etc.**

1. **Grouping customers into ideal number of clusters.**

Develop a model to visualize data from mall customers and cluster them into ages and potential products purchased.

[Dataset](Customer%20segmentation/Mall_Customers.csv).

Anomaly detection – Outlier detection – Novelty detection – Unsupervised learning

1. **Classification – Unsupervised – KNN – Fraudulent credit card transactions**

The dataset we will use contains transactions made by credit cards in September 2013 by European cardholders. The dataset has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

The dataset contains 31 columns, only 3 columns make sense which are Time, Amount and Class (fraud or not fraud). If required use PCA to reduce unnecessary dimensions.