Process and Visualize heterogeneous data

Elasticsearch, Kibana, Sagemake

Project

The goal of this project is to pre-process heterogeneous data and visualizes the information using Elasticsearch, Kibana and Amazon SageMaker Studio.
Prepare a report for the Project.

Start by analysing the datasets given on and choose a dataset:
https://archive.ics.uci.edu/ml/datasets.php

The project should contains at least the following 5 parts:

1. Load the CSV file(s) in Elasicsearch
2. Visualize the information using Kibana
3. Dimensionality reduction and data projection with Sagemaker (Python)
4. Results analysis and Conclusions

1. Elasticsearch is an open-source, distributed search and analytics engine built on Apache Lucene. Elasticsearch is the most popular search engine, and is commonly used for log analytics, full-text search, security intelligence, business analytics, and operational intelligence use cases.
Load the data from csv format in Elasticsearch by using LogStash.
What are the advantages of having data in Elasticsearch?
What are the advantages of using LogStash?

2. Kibana is an open-source data visualization and exploration tool used for log and time-series analytics, application monitoring, and operational intelligence use cases. It offers powerful and easy-to-use features such as histograms, line graphs, pie charts, heat maps, and built-in geospatial support.
Use different visualization functions in order to plot some statistical information from the data stored in Elasticsearch i.e. piecharts, histograms, … Explain the obtained visualizations.

3. Amazon SageMaker Studio is an integrated machine learning environment where you can build, train, deploy, and analyze your models all in the same application.
Import the data into the SageMaker Studio. You can use the Python in local if you have problems with AWS.
Use Python with scikit-learn and matplotlib to Project the data in 2 dimensions and plot it using scatter. You should use at least 3 projections methods (PCA, LDA, SNE, MDS, LLE,…) and compute the execution (learning) time. Compare the obtained results. Change the used parameters and find the best ones.

4. Make a general conclusion about the use of these techniques and the advantage of using them on heterogeneous data. Conclude the Project.

Note: If Amazon will charge you for the manipulation you are doing for the Project or Practical Lecture, please send quickly an email at Amazon support and email M. Frédéric Sananes indicating the charging reason/error.