ARISTOTLE UNIVERSITY OF THESSALONIKI
FACULTY OF SCIENCES
SCHOOL OF INFORMATICS
DEPARTMENT OF COMPUTER SCIENCE
«KNOWLEDGE, DATA AND SOFTWARE TECHNOLOGIES»

Master Thesis

Machine learning methods for the analysis of data of an Electricity Distribution Network Operator

by Ioannis Mamalikidis (UID: 633)
for the degree of Master of Science

Thesis Committee

Supervisor: Eleftherios Angelis
Members: Grigorios Tsoumakas
Ioannis Vlahavas

THESSALONIKI
MARCH 2017

Abstract

Once every few decades an invention changes the landscape of some aspects of our life. Industrial revolutions improved our everyday lives whilst medical revolutions expanded our lifespans. In the path we’re leading, most of sciences will be reduced to computer science, enabling faster and more accurate results. Machine learning is a vast field whose use spans over a plethora of tasks like optical character recognition, search engines and computer vision, to applications on other fields, such as the medical one. Here, two of the three main categories of machine learning are being used; namely unsupervised learning to cluster the data into geographical groups, and an array of different types of supervised learning to make predictions. The data which are subject to machine learning originate from the Hellenic Electricity Distribution Network Operator (HEDNO S.A.), the largest company for the operation, maintenance and development of the power distribution network in Greece. Working with real data is bound to come with solid hurdles, such as a substantial amount of noise in the data, erroneous entries, missing values and incomplete data. To this end, a considerable amount of time was devoted to pre-processing; cleaning up the data, retrieving or extrapolating from it and transforming it into a suitable form for the next steps. Since the dataset consists of projects dealing with construction or repair on actual locations, it has a geographical aspect to it. As the data themselves do not come with associated longitudes and latitudes, a method was devised to retrieve them and in turn use them to cluster the projects into geographical groups. The Last step was to apply 10 machine learning algorithms to predict which future projects are going to be approved and which are not, hence enabling the company to be better prepared in terms of needed items availability. Statistical analysis on the trained machine learning algorithms themselves was also important in order to identify the best model for this dataset.