E3.1: Research report and conceptualization of automatic learning models for data patterns energy consumption, cataloging and recommendation to stakeholders.

This deliverable is divided into 2 distinct parts: 1) incremental selection of new variables and data for model training, and 2) machine learning models of app energy consumption patterns for cataloging and recommendation to stakeholders.

Extracting features or variables and consequently creating a dataset for model training in the scope of Android cellphones is a challenging task since the energy consumption can not be calculated in a straightforward manner. In this task, first, a dataset including different variables has been developed comprised of eighty-eight different applications found in the Aptoide app store within the category of games and across eighteen different subcategories. All these apps have been downloaded, at least, half a million times.

To calculate the energy consumption for each app, an Asus Zenfone 4 Max is used and it is connected to an Arduino Nano that will register the energy being consumed by the device through a program developed in python that registers the values from each test in a csv file. The python code also outputs a graph containing the various energy changes gathered during the testing phase of the application. This allows for a better visualization of how energy consumption shifts whilst the app is being executed by the device.

Two different csv files have been created from the information available on each app with the intent of gathering the necessary features that will be supplied to the machine learning model for the model training. The first dataset is comprised of both categorical and numerical values, with the exception of the last two columns. These serve the purpose of ease of identification of the apps within the app store in case it is needed for future research. The second one is comprised of a table relating each application to the respective permissions that they require. One hundred and twelve different permissions where identified in total across all the apps of the dataset.

Through the first dataset it will be possible to search for correlations between the size, number of permissions and minimum android version of the app and its total energy consumption. The second dataset will be used to correlate the energy consumption to the various permissions of each app.

The categorical feature values were also changed according to the ordinal encoding method that changes the categories with arranged integers expressing the original hierarchy. Through the testing device it was possible to gather the energy consumption in Watts of these apps over a period of 5 minutes, from this data, the average of energy consumption of each app was calculated. These averages were then split into 3 categories these were:

• Low: For apps whose average consumption is below 1.207 Watts;

• Medium: Apps with a consumption above or equal to 1.207 but below 2.228 Watts fit in this category ;

• High: For all other apps with a consumption above or equal to 2.228 Watts.

In the task of finding the best variables for data training in a dataset, the combination of machine learning and evolutionary computation techniques was also applied. Specifically, the feature selection step focuses on identifying the most suitable variables that can enhance prediction performance and reduce the number of variables required to represent the state of a smartphone accurately. To accomplish this objective, 12 algorithms have been selected to determine the optimal set of variables. These algorithms employ evolutionary computation methods to explore different combinations of variables and evaluate their impact on prediction accuracy. The evolutionary computation process involves the generation of multiple candidate feature subsets, which are then evaluated using machine learning models to measure their performance in energy consumption prediction.

This approach has several benefits. First, by reducing the number of variables, it helps in minimizing computational requirements, memory usage, and training time, which can be crucial in resource-constrained environments such as smartphones. Second, it enhances the interpretability of the prediction models by focusing on the most relevant variables. Simpler models with fewer features are often easier to understand and explain to stakeholders. Furthermore, the selection of these optimal variables does not negatively impact the performance of energy consumption prediction. This implies that the chosen variables subset effectively captures the necessary information required for accurate prediction, while eliminating redundant or irrelevant features. The algorithms ensure that the final feature set maintains or even improves the overall prediction performance, thus maximizing the efficiency of the energy consumption prediction system.

The mobile market has seen tremendous development throughout the past few years both in terms of hardware and the software that is available for the devices. Despite this, the batteries that power these devices have not seen major improvements and have been unable to accompany the progress seen in this field. From the literature, there have been few to no studies regarding the development of user-side solutions to help solve this problem. In order to fill this gap this task also focused on providing a machine learning solution with the intent of identifying links between the information available in the store page of an application and its energy consumption to develop an a priori method for the classification and certification of mobile applications. Hence the main contribution of this research resides on the machine learning model, adapted to the Aptoide Appstore and mainly targeting applications that belong to the games category, given that these have the highest volume of downloads and interest by the users of the appstore.

Two general approaches are used for the modeling process, classification, and regression. Both models were tested using a base feature set, which is comprised of only the permissions each app, totaling 89 features, and an extra feature set, which adds the minimum android version, total number of permissions and size of the app as features and adds up to 92 features in total.

In summary, the main contributions of the this section for cataloging and recommendation to stakeholders are

1) Developed a database containing a set of the most downloaded applications and the permissions for each app,
2) Designed a machine learning model capable of automatically prediction the energy consumption from Appstore data for an application (in the game category),
3) Tested the model with the generated dataset belonging to the game's category under different circumstances,
4) Evaluated the results of the model with different machine learning algorithms.
5) Investigated a hardware-based method to calculate the energy consumption for an App
6) Developed a database containing a set of the most downloaded applications and the permissions for each app.

This task is a part of M.Sc. Thesis Dissertation:

Daniel Afonso Valente

Machine Learning for the Prediction of App Energy Consumption from App store Data

Master in Engenharia Informática,

University of Beira Interior

Defended October 2022