# Itility Assignment – Yiannis Makantasis

For the purposes of this assignment the main task was to analyse the contents of the “Smart applicator” dataset and give advice regarding on which IPC’s could be considered to be downsized.

Regarding some initial details, the programming language chosen is python, while the packages used are pandas, numpy, matplotlib.plt and seaborn.

A black screen with white text

Description automatically generated

For the initial analysis, the dataset was checked for any duplicate entries such that they can be removed.

A screen shot of a computer

Description automatically generated

As there were not any duplicate rows that were found, no entry was removed.

Moving on, through investigating additional information of this dataset it was found that some of the columns were assigned incorrect datatypes (namely, “MemoryMB”, “CpuMHz”, “NumCpu”) and had to be changed, while also correcting some wrongly imported values that they had.

A screenshot of a computer

Description automatically generated

An example of the wrongly imported values can be seen in the following picture. Where for both columns “CpuMHz” and “NumCpu”, two values exist in the cell, when instead it should have been one.

A black background with white text

Description automatically generated

After changing the datatypes of the three columns and removing the wrong entries, the following picture corresponds to the new information of the dataset. The new datatypes for “MemoryMB”, “CpuMHz”, “NumCpu”is now int64, instead of object.

A screen shot of a computer

Description automatically generated

Some experiments were also performed, with the focus being on the two metrics found on the column “MetricID”, namely “MemActive”, “CpuUsageMHz”

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

As only the “CpuUsageMHz”, contains useful information it was the metric used for the later steps of the experimentation. It is interesting to see that the maximum value for the CPU usage is approximately 1.8 MHz which far outperforms the maximum capacity for the “CPuMHz”,

With those changes now complete, the first question can be answered namely determining how many IPC can be downsized if they exist.

To find out the number of the IPC’s that can potential be downsized, tests were performed on **4259** unique entries of the Smart application dataset. Their CPU metric was measured and from them it was found that **1756** of them are considered as oversized and can be downsized if need be.

To reach this result, the first step to be taken was to use the “CpuUsageMHz” as the metric from which a new dataset was created, having as entries only the IPCs with this value in the “MetricID” column.

A screenshot of a computer

Description automatically generated

As for the threshold used, according to research, a value of 30% was taken. This means that if the CPU of an IPC is using less than 30% of its capabilities for the period measured, then it is defined as oversized.

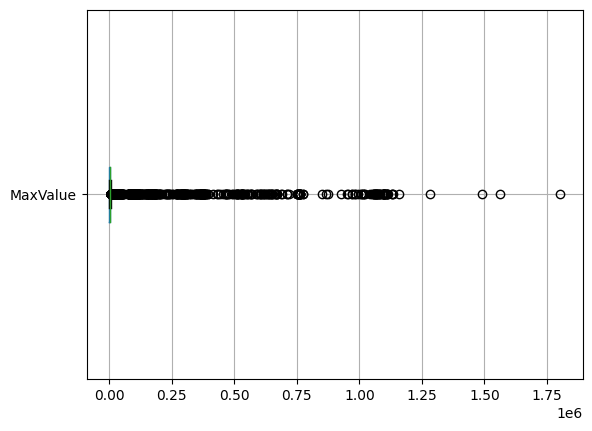
A black background with white text

Description automatically generated

Looking at the image above, a new column was added on the created DataFrame, called “Oversized” with its values initialized to 0. This column is intended to store whether a CPU is considered oversized or not. Then, to see if a CPU is actually oversized or not, the condition to be check is whether its maximum value is at most 30% of its actual capacity. If this condition is satisfied, it implied that the CPU is considered oversized, and the “Oversized” column is set to 1. This list can also be exported as a CSV file for any reference.

As for the peculiarities that can be found in the dataset, a first example tis the existence of cells containing two values instead of one. This was the case in the columns “MemoryMB”,”CpuMHz”, as well as, “NumCpu”. As no further explanation was found as to why those values appeared, possibly an oversight during the creation of the data, the entries containing those values were dropped.

Another strange activity can be found when looking at the boxplot below. Here the metric chosen to be analysed is the maximum value regarding the CPU utlilization. As it can be seen, there are several outliers where the usage of the CPU is much more when compared to the capacity of the largest CPU. This can possible be attributed to usage spikes in the CPU, causing the IPC to report wrong results due to errors that can exist.



Finally, with the help of machine learning, the process of continuously testing this analysis across all IPC’s can be achieved. As there are already IPC’s that are labelled with 0 and 1 based on the fact that they are considered oversized or not, the data can be split for training and testing purposes. As such, using binary classification the model can be trained and tested, and valuable information can be extracted.