PREDICTING HARD DISK FAILURE AT THE DELTA CENTER IN THE NETHERLANDS

From centuries, Dutch people have pumped the water of the lakes and the sea in order to build big cities on the new dry land. That is why around of sixty percent of the surface area of the Netherlands is bellow the sea level, with a high risk of flooding. In order to prevent an overflow that could destroy the western part of the country, artificial beaches, sand dunes and dikes were built to absorb the forces of a rising sea. However, the Dutch hydraulic system was not built and maintained properly until the 50's. Proof of that were the effects of the most devastating flood in the Netherlands' history, where 1800 people and 200000 animals died as a result of the collapse of the dikes' structure.

The delta project started in 1953, twenty days after the flooding. The aim of the Delta project was to build a complex system of automatic dikes, barriers and dams that control the sea level and drain off the excess of water coming from the large rivers. Currently, the Netherlands has 700 km of dikes, which are divided in 53 dike areas.

The dikes and damns are controlled with supercomputers, which monitor the status of these structures 24 hours per day. A damage in the supercomputer; for instance, a failure in some of its hard disks, would produce devastating effects that would result in another flood.

The aim of this project is to predict the number of hard disks that fail during the first week of 2016 at the Delta center in the Netherlands. To make this prediction, I will use the data provided by the Backblaze center, which contains the information during the year of 2015 of the status of 62000 hard drive disks.

The data can be downloaded from this website:
https://www.backblaze.com/hard-drive-test-data.html

Every day, the Backblaze center takes a snapshot of each hard disk. This snapshot includes basic drive information along with the S.M.A.R.T. (Self Monitoring Analysis and Reporting Technology) statistics reported by that drive. The daily snapshot of one drive is one record or row of data. All of the drive snapshots for a given day are collected into a file consisting of a row for each active hard drive. The format of this file is a CSV (Comma Separated Values) file. Each day this file is named in the format YYYY-MM-DD.csv, for example, 2015-04-10.csv.

The data contains the following information (listed by columns):

- *Date* – The date of the file in yyyy-mm-dd format.
- *Serial Number* – The manufacturer-assigned serial number of the drive.
- *Model* – The manufacturer-assigned model number of the drive.
- *Capacity* – The drive capacity in bytes.
- *Failure* – Contains a "0" if the drive is OK. Contains a "1" if this is the last day the Hard disk was operational before failing.

- *2015 SMART Stats* – 90 columns of data, that are the Raw and Normalized values for 45 different SMART stats as reported by the given drive. Each value is the number reported by the drive.

From the 90 columns corresponding to different S.M.A.R.T. attributes of the disk, I will restrict my analysis to only six attributes namely:

- *Read Error Rate (S.M.A.R.T. 1)*: Stores data related to the rate of hardware read errors that occurred when reading data from a disk surface.
- *Spin-up time (S.M.A.R.T. 3)*: Time that spends the spin disk to be at the operational velocity (milliseconds).
- *Reallocated Sectors Count (S.M.A.R.T. 5)*: Count of the bad sectors that have been found and remapped. The higher the attribute value, the more sectors the drive has had to reallocate.
- *Running time (S.M.A.R.T. 9):* Number of hours a drive has been in service up that point. The raw value is measured in hours.
- *Power cycle count (S.M.A.R.T. 12)*: The count of full hard disk power on/off cycles.
- *Internal temperature (S.M.A.R.T. 194)*: Internal Temperature of the disk in Celsius

A preliminary analysis of the dataset shows that 2% of the disks fail while 97% are still working after one year of measurements. Therefore, it is necessary to choose an appropriate sample for the dataset in order to make successful predictions of the disks that stop working after the first week of 2016. This procedure is called resampling and it consists of choosing a random number of working disks from the whole dataset. The number of working disks that are chosen must be similar to the total number of damaged disks in the population. We plan to use inferential statistics to the new sample of data and we also will apply machine-learning techniques for the prediction.

For the final product I will provide the codes that produce the prediction and a paper explaining in a detailed manner all the procedures that were carried out during the analysis of the data.