PREDICTING HARD DISK FAILURE AT THE BACKBLAZE DATA CENTRE

Each day in the Backblaze data centre, a snapshot of each operational hard drive is taken. This snapshot includes basic drive information along with the S.M.A.R.T. 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, 2013-04-10.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 drive was operational before failing.

2013-2014 SMART Stats – 80 columns of data, that are the Raw and Normalized values for 40 different SMART stats as reported by the given drive. Each value is the number reported by the drive.

I will analyse the data corresponding to January of 2016 (already downloaded). I will also restrict the analysis to six S.M.A.R.T failures, namely:

- **SMART_1** : Read Error Rate
- **SMART_5** : Reallocated Sectors Count
- **SMART_9** : Power-On Hours
- **SMART_194** : Temperature Celsius
- **SMART_197** : Current Pending Sector Count
- **SMART_211**: Vibration During Write

The aim of this project is predict when a hard disk breaks down, based on the SMART failures listed above.