The dataset consists of data collected from heavy Scania trucks in everyday usage. The system in focus is the Air Pressure system (APS) which generates pressurized air that are utilized in various functions in a truck, such as braking and gear changes. The dataset’s positive class consists of component failures for a specific component of the APS system. The negative class consists of trucks with failures for components not related to the APS. The data consists of a subset of all available data, selected by experts.

-- Challenge metric

Cost-metric of miss-classification:

Predicted class | True class |

| pos | neg |

-----------------------------------------

pos | - | Cost\_1 |

-----------------------------------------

neg | Cost\_2 | - |

-----------------------------------------

Cost\_1 = 10 and cost\_2 = 500

The total cost of a prediction model the sum of "Cost\_1" multiplied by the number of Instances with type 1 failure and "Cost\_2" with the number of instances with type 2 failure, resulting in a "Total\_cost".In this case Cost\_1 refers to the cost that an unnecessary check needs to be done by an mechanic at an workshop, while Cost\_2 refer to the cost of missing a faulty truck, which may cause a breakdown.

Total\_cost = Cost\_1\*No\_Instances + Cost\_2\*No\_Instances.

The training set contains 60000 examples in total in which 59000 belong to the negative class and 1000 positive class. The test set contains 16000 examples.

Number of Attributes: 171

Attribute Information:

The attribute names of the data have been anonymized for proprietary reasons. It consists of both single numerical counters and histograms consisting of bins with different conditions. Typically the histograms have open-ended conditions at each end. For example if we measuring the ambient temperature "T" then the histogram could be defined with 4 bins where:

bin 1 collect values for temperature T < -20

bin 2 collect values for temperature T >= -20 and T < 0

bin 3 collect values for temperature T >= 0 and T < 20

bin 4 collect values for temperature T > 20

| b1 | b2 | b3 | b4 |

-----------------------------

-20 0 20

The attributes are as follows: class, then anonymized operational data. The operational data have an identifier and a bin id, like "Identifier\_Bin". In total there are 171 attributes, of which 7 are histogram variables. Missing values are denoted by "na".