## Team “123”

|  |
| --- |
|  |

Data Understanding

Contents

[1 Document summary 3](#_Toc469378620)

[1.1 Document purpose 3](#_Toc469378621)

[1.2 Document scope 3](#_Toc469378622)

[1.3 Versions 3](#_Toc469378623)

[1.4 Revisions 3](#_Toc469378624)

[2 Data Understanding report 4](#_Toc469378625)

[ *Initial data collection report* 4](#_Toc469378626)

[ *Data description report* 4](#_Toc469378627)

[ *Data exploration report* 4](#_Toc469378628)

[ *Data quality report* 5](#_Toc469378629)

1. Document summary

## Document purpose

Document is intended to reflect the results of data understanding process.

Document contains some data analysis in order to highlight interesting subsets of the data and forming hypotheses about hidden dependencies.

One of the purposes of research part is to compare measurements of two completely different sensors “GT”, “PT08”, and give a brief summary.

## Document scope

Document describes requirements and workaround of the study project Data Mining 2016.

Results of the documenting process are to be discussed at the lessons.

## Versions

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Editor | Description of version |
| 2016-12-13 | 0.1 | E.N. Pavlovskiy | Document template |
| 2016-12-13 | 1.0 | E.N. Pavlovskiy | Final revision |
| 2016-12-15 | 1.1 | Team “123” | Starting data understanding |
| 2016-12-16 | 1.2 | Team “123” | Report correcting |
| 2016-12-17 | 1.3 | Team “123” | Hypotheses are formulated |
| 2016-12-18 | 2.0 | Team “123” | Final revision |

## Revisions

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Approved by | Decision |
| 2016-12-13 | 1.0 | E.N. Pavlovskiy | Approved for the course use |
|  |  |  |  |

1. Data Understanding report

## *Initial data collection report*

These have a simple tabular view of the customer and are provided as a csv-file. Current amount of data allows them to send an e-mail. The dataset has been downloaded from <https://archive.ics.uci.edu/ml/datasets/Air+Quality>, the size is 1.47 Mb.

In the process, there has been committed to only one version of the data.

## *Data description report*

Data presented as a single table consisting of 15 columns and 9358 rows, the first row contains the names of the features, the others are values of 15 parameters.

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value. In this paper such values are called as “*failure*”.

Detailed description of the features shown in the Table 1:

Table 1. Description of data features.

|  |  |  |  |
| --- | --- | --- | --- |
| № | Name of the columns from (.csv) file | Feature description | Format |
| 0 | Date | Date | (DD/MM/YYYY) |
| 1 | Time | Time | (HH.MM.SS) |
| 2 | CO(GT) | True hourly averaged concentration CO (reference analyzer) | mg/m^3 |
| 3 | PT08.S1(CO) | PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted) | Non mentioned |
| 4 | NMHC(GT) | True hourly averaged overall Non Metanic HydroCarbons concentration(reference analyzer) | microg/m^3 |
| 5 | C6H6(GT) | True hourly averaged Benzene concentration (reference analyzer) | microg/m^3 |
| 6 | PT08.S2(NMHC) | PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted) | Non mentioned |
| 7 | NOx(GT) | True hourly averaged NOx concentration (reference analyzer) | ppb |
| 8 | PT08.S3(NOx) | (nominally NOx targeted) | Non mentioned |
| 9 | NO2(GT) | True hourly averaged NO2 concentration (reference analyzer) | microg/m^3 |
| 10 | PT08.S4(NO2) | PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted) | Non mentioned |
| 11 | PT08.S5(O3) | PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted) | Non mentioned |
| 12 | T | Temperature in Â | °C |
| 13 | RH | Relative Humidity | (%) |
| 14 | AH | AH Absolute Humidity | kg/m3 |

## *Data exploration report*

During date exploration there have been constructed several tables and graphics, which clearly interpreted the big massive of input data by each feature. Also percentages of data without errors (failure) have been calculated and are shown in the Table 2.

Unrepresentative data

A *representative sample* is a small quantity of something that accurately reflects the larger entity.

All data except of NMHC present by itself give quite a good quality correspondence to real information.

But given quantity of data is not in enough to make an adequate assessment of sensors working.

Also the given data of some parameters is presented in different formats depending on each sensor (for example CO(GT) is given in mg/m^3 but CO (PT08.S1) is given in *some average values*.

The Picture 2 clearly shows how much time it is appeared each parameter; histograms have been made for 6 parameters measured by 2 sensors. Also histograms for temperature, relative and absolute humidity have been made. The measurement ranges are various for each parameter, histograms confirm the existance of “failure” values -200.

|  |  |
| --- | --- |
| 1 | 2 |
| 3 | 5 |
| 6 | 7 |
| 9 | 10 |
| 4 | 11 |
| 13 | 14 |
| 12 | |

**Picture 1**. Frequency of occurrence of each value of parameters

Table 2. Analyzing of picture 1

|  |  |  |
| --- | --- | --- |
| Name of the columns from (.csv) file | Numbers of items that are equeled to -200 | Persentage of data without errors |
| CO(GT) | 1683 | 82,013 |
| PT08.S1(CO) | 366 | 96,088 |
| NMHC(GT) | 8443 | 9,768 |
| C6H6(GT) | 366 | 96,088 |
| PT08.S2(NMHC) | 366 | 96,088 |
| NOx(GT) | 1639 | 82,484 |
| PT08.S3(NOx) | 366 | 96,088 |
| NO2(GT) | 1642 | 82,452 |
| PT08.S4(NO2) | 366 | 96,088 |
| PT08.S5(O3) | 366 | 96,088 |
| T | 366 | 96,088 |
| RH | 366 | 96,088 |
| AH | 366 | 96,088 |

Analyzing the histograms on picture 2 and table 2 it could be done the important conclusion about the degree of representativeness of given dataset. It means that there is an opportunity to work only with some data.

|  |  |
| --- | --- |
| (a) | (b) |
| D:\-200\23.png | D:\-200\5.png |
|  |  |
| (c) | (d) |
| D:\-200\2.png | D:\-200\15.png |

**Picture 2.** Time distribution of failures. (a) – PTO8.S1.C0, C6H6.GT, PTO8.S2.NMHC, PTO8.S3.NOx, PTO8.S4.NO2, PTO8.S5.03, T, RH, AH (b) – NMHC.GT, (c) – CO.GT,(d) – N0x.GT, NO2.GT

“*Failure*” – the special sign which indicates that one or more items are equal to -200. During data analyzing it has been noticed that graphics for failure of some features are similar to each other, so there has been made only four groups for graphics see Picture 2. On the graphic 2.b it is only 9.8% values without failure, so it is not any ability to predict something basing on it.

## *Data quality report*

Customer is responsible for filling in gaps in their data:

Table 3. Data gaps information

|  |  |  |
| --- | --- | --- |
| Feature | Replacement NA | clarification |
| All features | -200 | If there are some problems with data measuring |

This should be taken into account for the correct processing of the gaps.

Testing conditions, data quality and errors are given in table 4.

Table 4. Information about data quality and errors

|  |  |  |
| --- | --- | --- |
| Feature | Test condition | Format errors |
| Data | dd.mm.yyyy | no |
| Time | hh:mm:ss | no |
| Temperature | -50..50 0C | no |
| Relative humidity | 0…100% | no |
| Others parameters | >0 | no |

## *Hypotheses*

* As it is seen on picture 2 (c, d), the large number of failure values by GT device are happened in nearly same time (the period 3 - 5 a.m.). The first hypothesis is that in this time period regular sensors’ update happens.
* The quantity and distribution of errors coincide for sensors PT08 for different parameters that suggest that there is failure of whole device or failure in data transmission.
* Also it can be suggested that failure values (-200) can influence on average value that is presented in .csv file
* Also it can be made independent hypothesis that PT08 device is more reliable, because quantity of all failures is less than quantity in GT device.
* It cannot be built any suggestions about the sensors’ accuracy, because it cannot be compared the measurements by each sensor, because the units are unknown and range of values is quite extensive.

## *Conclusion*

The team “123” hopes about successfully overcoming of all stages of the studying the project Data Mining 2016. This paper specifically reflects the Data Understanding part basing on AirQuality dataset. All key moments of data understanding have been studying, analyzing, also some skills in R working have been acquired, especially graphical display of big massive of input data has been studied.

Given data has been analyzing, some defects has been found and reviewed, interesting dependencies have been found, and some hypotheses have been formed.

Comparisons between measurements of two completely different sensors have been made in detail.

More results are to be discussed at the lessons.

\*\*\*