## Applying data science and machine learning in development of a tool for sustainable waste management onboard cruise ship

### Abstract

### Background

In this work, a data product will be produced by using data science tools and machine learning (ML) algorithms that would help a cruise ship operator to plan ahead the future waste management operations by a most sustainable and feasible way.

The project work is done by constructing the ML-product by using real shipboard operational data and external web sources. The purpose is to create ML-system that would give a decent picture and prediction of the best option for the operations.

#### Cruise ship waste

A cruise ship is a passenger ship used for pleasure voyages when the voyage itself, the ship's amenities, and sometimes the different destinations along the way (i.e., ports of call), form part of the passengers' experience [1]. Modern cruise ships can be considered as a combination of hotel, shopping center, amusement park, casino, spa and a center of all kind of restaurants with big galleys for preparation of all kind of meals. All these onboard activities produces high volumes of wastes that is around 2 to 3 times the amount produced by municipal centers per capita [2]. Especially, as all the meals (around 8 per each day) are prepared onboard the ship in the central galleys. Thereby, organic and nutrient loading of the waste, if discharged to the sea, is very high.

The waste management system of a modern cruise ship is confined to ship machinery spaces. There are basically three type of solid waste produced onboard: Dry waste (DW), food waste (FW) and biosludge waste(BW). The DW is mainly solid packing materials from the provisions and unsorted waste from the cabins and restaurants. The FW is produced mainly from the food preparation from the gallies but also as a food rejects from bars, restaurants and crew messes of the ship. The BW is a biological surplus sludge from the ship wastewater treatment plant.

The management and the discharge of the waste materials is strictly regulated by both international and national legislations and guidelines. International Maritime Organisation (IMO) [3] The discharge of garbage is strictly prohibited in all sea areas. The discharge of FW and BW is also strictly regulated and also prohibited in many special sea areas. In addition to this the storage silos for these waste materials is very limited and is sufficient only for couple of days.

#### Motivation for the work

Evac Oy supplies integrated cleantech solutions, including e.g. waste management systems, to all types of ships. The cruise ships are our biggest market segment and for past few years Evac has had more than 80 % of market share in waste systems. The sustainability is the key value for all major cruise ship owners. Not only because of regulations and guidelines, but because of acceptability of the cruise business as a whole. Basically, the cruise ships are operated by zero-discharge principle.

The cruise ships are operated on a strictly defined itinerary and and in the given schedule. The shipboard waste management is not the main operation of the ship and should by not means affect the operation of a ship. On the other hand, volumetric production of the waste streams (DW, FW, BW) are not constant and are fairly unpredictable. As mentioned the onboard storage silos are very small in volume. For that reason, cruise ships are many times forced to store waste container in corridors and luggage storages. For the same reason, ships are frequently discharging high proportion of untreated FW and BW directly into the sea. This poses high risk for the reputation of the ship owners as well as for the whole cruise business.

### Purpose and objective

The purpose of this work is to carry out an exploratory analysis on the waste data produced from one of the ships of TUI Cruises to produce an insight of the dynamics of the waste production. From that knowledge a machine learning tool is trained and validated to predict the production of the waste streams. Ultimately, the project should produce a machine learning tool that will predict the production of the three streams and thereby help the operator to plan ahead the waste management operations of the ship.

The goal is to study how to extract usable data from the existing sources, clean and explore the data so that a usable predictive ML models could be produced and trained with new and fresh data in the future from the IoT-system. When this goal is achieved, it may be worth of considering to apply and further develop the extent of the work to an usable data science product.

## Materials and methods

### Data sources

#### Shipboard data

There are both internal and external data sources that would be used. The internal data sources are the waste data produced from the ship IoT. In the first stage of data exploration, a set of readily produced CSV-files containing all the waste data from Mein Schiff 6 covering couple of days. The raw data set consists of:

- volume as cubic meters (m3) of dry waste produced onboard cruise ship into a dry waste silo. The CSV file contains status information of the gate valves that needs to be converted into a volumetric values of waste production. This data is converted to kilograms (kg) of dry waste and eventually rate of production.

- volume as cubic meters (m3) of food waste produced continuously onboard cruise ship into a food waste silo. There are two FW silos. The data is produced as filling %-level of the tanks. The data is converted into cumulative production of kilograms (kg) of food waste and eventually as rate of production

- volume as cubic meters (m3) of biosludge produced continuously onboard cruise ship into a BW silo. The data is produced as filling %-level of the tanks. The data is converted into cumulative production of kilograms (kg) of BW and eventually as rate of production.

In addition to the waste data, there is data of the waste management system including pumping information of the FW solids and liquid and data from waste incinerator units.

The first set of data is produced from Mein Schiff 6. As said, the data of the waste production is first cleaned and plotted to understand the nature of the data. These data sets will be used as labeled data for prediction of future waste production trends. The first development and training of the ML-model is done with this data set.

#### Operational data

For predicting the future waste production, we need to understand the basic facts and activities affecting the waste production rate in order to select sensible and most influential features for our ML-model. First one need to ask that is the ML-model really needed to know the future waste production. Basically, the prediction model would not be needed if the the production rate is either constant or the pattern of production rate is well known.

Is the production rate constant and/or is the production pattern know? To answer that question one need to know the variables affecting the waste production. The pattern of production is dependant mainly on:

- Activity of passangers that is dependant on the time of a day: That would basically be solved with Pandas datetime-data type.

- Amount of passangers onboard: Basically, it is safe to assume that the ship is always fully booked, that is the situation almost always for all cruise ships in real life. This may be thereby either neglected or use constant integer, i.e. known capacity of the ship derived from the web site of the ship by web scraping.

- Proportion of passangers present onboard: Cruise ships visit the ports on a daily basis and most of the cruise ship customers disembark to the port of call. So, the relevant question affecting the waste production; is the ship in a port side or cruising in the sea conditions? This information can be derived from the web site of the ship by web scraping. The data type is boolean.

- The amount of food waste is dependant on the meal time, meaning that two hours before and until the meal time the main galleys are fully operational. We can assume the meal times of a day based on our knowledge about the ship operation. The data type is boolean.

- During staying in the turnaround port (i.e. home port). The waste production is higher because the whole ship is cleaned between 8:00 am and 01:00 pm. This increases substantially the amount of DW.This information can be derived from the web site of the ship by web scraping. The data type is boolean.

As we can see, there is at least five most relevant attributes affecting the waste production rates. Thereby, one can safely conclude that the production rates aren't constant nor known. We can also conclude that the rates are not same throughout one day. Most likely the the production patterns are very ship-specific and cannot be generalised from single attribute. Therefore, the development of the predictive machine learning model is a relevant task.

### Method for development of Predictive Machine Learning model

Typical situation for creating a predictive model is that in a set of inputs (X) are readily available, but the output (Y) cannot be easily obtained [4]. The output can be predicted with an estimate of function describing the unknown relationship between X and Y. Basically, a model, that is an estimate of the relationship, is at first treated as a black box that would be called as hypothesis space, meaning all computationally reasonable relationships between X and Y.

#### Definition of the setup for the machine learning task

In this project, the high-level insight produced by the model is the production of the waste streams, DW, FW and BW, that will tell whether or not the storage capacities are sufficient. This will return recommendation of actions whether to treat the waste onboard, store in the silos or if the waste needs to be discharged to the sea. As an ultimate goal would be to combine the return value of sustainability in terms of environmental impact or feasibility information in terms of cash flow or cost from waste disposal.

The inputs (X) as relevant features for the ML-model are the operational data described above. The labelled data as relevant higher-level insights are the known shipboard data described above. This data will be supplemented by additional data from the IoT-system onboard Mein Schiff 2, that will be live after 11th April 2019. This data will be also used for furter training of the model as well as for the validation of the model.

The output (Y) predictions will be the predicted waste productions.

#### Development of the ML-model

The relevant question for development of the model is that is it possible to predict the waste production from the operational data listed above. The main prerequisite for development and training of predictive regression model is a feature space that is quantitative. However, as we can see our given feature space contains mainly categorical features with a boolean values.

In this case, one proven technique is to use one-hot encoding, which effectively creates extra columns indicating the presence or absence of a category with a value of 1 or 0, respectively [5]. In this project that would simply mean, e.g. ships location is indicated as one category a sea condition as 1 (True) and portside as 0 (False). Also, the meal time is one category indicated as either 1 (True) and rest of the time as 0 (False).

After preprosessing the features into useable feature space, the model is developed. The labelled data is joined with the another dataset. Basically, try to determine the extent to which the above listed operational data affect the volume of waste streams. A simple linear regression is performed to relate the operations and the labeled data.

The linear regression functionality of the Scikit-Learn-package will be utilised for model development and training the model as well as for the validation of the model

### Computational infrastructure, software product considerations

The work is done with Jupyter Lab of Anaconda distribution package. The code will be written in Python meaning that all the usable Python libraries, incl. Pandas for data preparation, Matplotlib and Seaborn for visualisation of cleaned data and Scikit-Learn for model training and validation.

The model development is the main goal of the project. As I have only basic level of programming knowledge, there is no guarantees that the work will be finalized as data product to be utilised by end-customer. The primary goal of the work is to learn the methods of data science and machine learning. If I get to the production stage with this work, the GUI would be developed PyQt-GUI-tool. Alternatively, a web application could be developed with Python Flask-package. As there is lot of real time data available, it would be preferable to have it running real time. For the employer company point of view, a software product would be the desired option. However, most of the efforts during these studies should be put to data management and the ML-model building. As long as there is no connection between a specific ship or the ship owner, there should not be any privacy issues. The model should basically be a tool for evaluating sustainability of human actions towards the nature.

## Preliminary Conclusions

Preliminary, it can be concluded that the data from the ship IoT-system is both dense and thereby noisy. The data sampling interval on 3 to 5 seconds is emphasising the underlying noise caused mainly by ship movement. Also, the data is derived mainly only as status information that causes multi-stage data transformation before one can make any meaningful visualisation of the data.

Deriving the operational data by web scraping is a good rehearsal for practice of data hacking from the website. However, the reliability of the data is dependant on whether or not the data on the webpage is correct or not. This entails a risk for model development stage.

The feature space contains mainly categorical data and needs preprocessing prior to model development. Also, it remains to be seen if this categorical feature space is applicable for training a reliable model.

## References:

[1] https://en.wikipedia.org/wiki/Cruise\_ship

[2] Cruise ship waste management database. Evac Oy. 2019.

[3] IMO Marpol Annex V, Resolution MEPC. 295(71).

http://www.imo.org/en/OurWork/Environment/PollutionPrevention/Garbage/Pages/Default.aspx

[4] An Introduction to Statistical Learning with Applications in R. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. 2017. eBook. Corrected at 8th printing.

[5] Python Data Science Handbook. Essential Tools For Working With Data. Jake VanderPlas. 2017. 1st Edition. O'Reilly.