Smoke Detection

Group: D12
Members: Merit Laidroo, Hannogert Otti
Link to project: https://github.com/Hannogert/IDS2022_D12

Business understanding

Background: A smoke detector is a device that warns of a fire in its initial stages. Because most tragic cases happen when people are asleep, the smoke detector needs to be accurate and be able to warn people before it is too late. In Estonia, it is mandatory by law to have a smoke alarm in every household.

Business Goal: The project benefits everyone because our goal is to improve the smoke detectors' ability to detect a fire by creating an accurate model that decreases the number of false alarms. Another goal is to get an insight into what features affect the triggering of the fire alarm.

Business Success Criteria: We will consider the project a success if we create at least one model that can accurately predict the triggering of the fire alarm.

Inventory of resources: Our group consists of two people. Each person has a personal computer with Internet access. We will use Jupyter notebook to create the project, which members will maintain in a GitHub repository. We also have access to the dataset from Kaggle and access to teaching assistants in case of any problems.

Requirements, assumptions, and constraints: The project has only time constraints. The deadline for the poster is 12.12.2022, and the deadline for the project itself is 15.12.2022. We assume that we will complete the project in time without significant problems.

Risks and contingencies: Due to life happening, there might be a few factors that might delay the completion of the project. The main factor that might hinder the realization in time is a team member getting sick, Internet outages, or power outages.

Terminology:
Data mining terms:

1. **Data -** collected values.
2. **Accuracy -** accuracy refers to the rate of correct values in the data.
3. **Data mining -** information extraction activity that aims to discover hidden facts in data.
4. **Cleaning -** step in preparing data for a data mining activity.
5. **Decision Tree -** a tree-like representation of a collection of hierarchical rules that lead to a class or value.
6. **Random Forest -** a machine learning algorithm that constructs many decision trees at training time.

7. **K-Nearest Neighbor -** a machine learning algorithm that tries to predict the correct class for the test data by calculating the distance between the test data and all the training points.
8. **Test Data -** a data set independent of the training data set, used to fine-tune the estimates of the model parameters
9. **Training Data -** a data set used to estimate or train a model.
10. **Training -** another term for estimating a model's parameters based on the data set.

Our project does not contain any business terms.

Costs and benefits: We will not be benefiting from the project financially. Expenses related to the project are non-existent.

Data-mining goals: The project's goals are to find correlations between the measured data and create a model to predict the triggering of the fire alarm. We will make models using four algorithms K-nearest neighbors, Random Forest, Decision Tree, and K-Means, with different variables. Also, we will be creating charts to illustrate the correlations and gather information for the poster.

Data-mining success criteria: Our success criterion for the data-mining is to find an accurate model for triggering the smoke alarm. The trained model should have an accuracy of 0.9 or higher.

Data understanding

Outline data requirements: For the project, we need to know air temperature, humidity, particulate matter concentration in the air, air pressure, amount of raw hydrogen and ethanol in the surroundings, Total Volatile Organic Compounds and CO2 concentration in the air. All the values should be numerical.

Verify data availability: Dataset is available for everyone on Kaggle. All the needed data for the project is in the dataset, so we do not need to gather any new data.

Define selection criteria: For the project, we will use smoke_detection_iot.csv, which contains all the needed features. We will not be using any other files. During the project, we will use all features except UTC and CNT. We will not use those features because those values are not crucial for training a model or finding the correlations between features and triggering the fire alarm.

Describing data: The dataset consists of 62630 rows and 16 columns. The dataset contains the following columns and data types:

- **UTC**: Time when the experiment was performed (Integer).
- **Temperature:** Temperature of surroundings, measured in Celcius (Float).
- **Humidity:** Air humidity during the experiment (Float).
- **TVOC:** Total Volatile Organic Compounds, measured in parts per billion (Float).
- **eCO2:** CO2 equivalent concentration, measured in parts per million (Integer).
- **Raw H2:** The amount of Raw Hydrogen in the surroundings (Integer).
- **Raw Ethanol:** The amount of Raw Ethanol in the surroundings (Integer).
- **Pressure:** Air pressure, Measured in hPa (Float).
- **PM1.0:** Particulate matter of diameter less than 1.0 micrometer (Float).
- **PM2.5:** Particulate matter of diameter less than 2.5 micrometers (Float).
- **NC0.5:** Concentration of particulate matter of diameter less than 0.5 micrometers (Float).
- **NC1.0:** Concentration of particulate matter of diameter less than 1.0 micrometers (Float).
- **NC2.5:** Concentration of particulate matter of diameter less than 2.5 micrometers (Float).
- **CNT:** Sample count (Integer).
- **Fire Alarm:** If the fire alarm was triggered, 1 if triggered, and 0 if not triggered (Integer).

Exploring data: Dataset doesn't contain duplicates or missing values. During observation, we couldn't find any data quality problems. Temperature values are between -22.01C and 59.93C. The mean temperature is 15.97C. The mean air humidity ranges from 10.74% to 75.2%, and the mean humidity is 48.539499%. The measured CO2 concentration during the tests was between 400 to 60000 ppm, and the mean was 670 ppm. Out of 62630 entries, the fire alarm was triggered 44757 times.

Verifying data quality: The dataset is good enough to support our goals. We should be taking into account when training our model that the data is a little imbalanced towards triggering the fire alarm. But other than that there are not any quality issues in the dataset.

Planning your project

Tasks:

1. **Planning the project -** ~8h. All team members will do this task. Members will divide the tasks between each other. Also, the members will meet every Tuesday to discuss the project's progress.
2. **Creating a repository -** ~1h. Hannogert will do this task. Member has to create a repository, invite other team members, create the notebook for the project and prepare the data for data mining.
3. **Understanding the data -** ~4h. All team members will do this to ensure everyone knows what to do.
4. **Visualizing the data -** ~4h. Both teammates will do this task. Members have to create plots to visualize the correlations between data. During the exercise, we will be using libraries such as matplotlib and seaborn and also Tableau. Each teammate must plot at least two graphs and write an explanation.
5. **Finding and reporting correlations between features -** ~6h. All team members will do this task. Both teammates will do the task together so teammates can discuss it. Each team member will need to try to find at least one correlation.
6. **Training a model for smoke detection -** ~6h. Both team members will do this task. Each teammate will train two models and try out different variables to ensure the desired accuracy. The aim is to train at least 1 model with an accuracy of 0.9.
7. **Design a poster -** ~4h. This task will is done by Merit.