Neighbourhood Based Fire Risk Effect Module

Dashboard for Petrol Stations around Amsterdam

Group D2 - Faaris Akbar (11899573), David Chen (13642111), Josephine Chrisiani (13815091), Hualiya Li (13872893), Mengqi Zhang (13792369)
Data System Projects 2021
University of Amsterdam

ABSTRACT

This project aims to improve the fire effect estimation of the existing fire risk model and provide insights for inspections on venues with a significant effect on fire incidents. A neighbourhood-based fire effect module is proposed, and the prototype is focused on petrol stations in Amsterdam. Data and features are collected from reliable sources and could be updated depending on the range of effects via Google API. Based on the literature, the effect range is suggested but configurable as user input. State-of-the-art clustering algorithms are adopted to group petrol stations based on feature similarities. A weighted effect-scoring model is introduced in parallel for inspection suggestions. The prototype is presented on a user-interactive dashboard, which is tested by project stakeholders.

CCS CONCEPTS

• **Data Science** → *Unsupervised Machine Learning*; • **Information Systems** → *User Interface Design.*

KEYWORDS

Fire Risk Model, Unsupervised Clustering, Effect Score, Neighbourhood Features, Interactive Systems, Google Maps API

1 INTRODUCTION

Amsterdam municipality takes care of the non-residential fire inspection, and they determine the inspection frequency based on the capacity and the function of the building. The type of buildings is divided into four categories, the first category needs inspection once per year, and the second category needs inspection once per two years. Kindergarten and elder houses could fall into category one as the builder users are less mobile. The category derives from the building's risk score, roughly a three-factor model that considers the chance of the fire incident (object or building), effects (economic, cultural, social, and environmental), and observations based on the inspectors' work. However, the problem is that there are limited resources for the inspections. The Amsterdam municipality is looking to optimize the inspection planning to improve its current working methods.

The project is focused on one specific aspect of the fire risk model. The initial idea was to explore the fire risk of a chemical plant explosion in the research phase. The idea came based on the news that there were chemical plant explosions in Germany[1]. However, based on the research, only a few chemical plants in Amsterdam are available. Petrol stations are hazardous workplaces because they store and sell flammable material (petrol, diesel, and compressed natural gas) [1]. Vehicle refueling stations are built in crowded urban, suburban, and motorway sides, with massive traffic and people [2]. Availability of flammable materials at petrol stations poses a constant hazard to the staff, public, assets, and environment. Based on research, it is decided that the project is focused on assessing the fire risk on petrol stations in Amsterdam. This project aims to predict the use case of fire and explosion risks in petrol stations by creating a fire risk model.

First, the relevant literature is described in this document. In the second section of this paper, the method is described. The results are elaborated on in the Fire Risk Model section. After the results, the prototype is described, and limitations are mentioned in the discussion section. The paper concludes with suggestions for future work.

1.1 Problem Description

In 2019, the municipality of Amsterdam took over the fire safety inspections. Several departments are working together on the model that calculates the fire risk. More specifically, the research department and the department of permits, surveillance and enforcement (in Dutch: vergunningen, toezicht en handhaving, hereafter referred to as VTH) work together to obtain more risk-based insights so the fire safety inspections can be organised in a more efficient way. Naturally, the fire inspectors are unable to assess all non-residential venues in the city and therefore, they have to prioritise which venues to inspect. This is done based on a model that calculates the risk of a fire incident. Risk can be calculated as the probability of a fire incident multiplied by the effect of a fire incident. However, the effect can be categorised again into social, economic and cultural effects. The current model is a summation of a checklist, but it is desired to get extended by estimating or predicting the effect within different scenarios. In order to estimate the effect for different scenarios, new features need to be collected which are able to describe buildings based on their characteristics. In addition to this, the characteristics and feature information of the venues, probability and effect of a fire incident has to be made accessible to fire inspectors and planners of the VTH. The ultimate goal would be to contribute to the daily work of the VTH department such that inspectors and planners can make optimal, data-driven and risk-based decisions.

1.2 Requirements List

The main requirements for a solution for the above-stated problems are threefold, namely, (1) collecting new features that can describe the venue features and (2) using these features, extending the current risk model within several scenarios. (3) Combining (1) and (2) into a visualisation that the fire safety inspectors and planners can use.

[1]https://www.bbc.com/news/world-europe-57978043

Regarding (1), the solution requires descriptive data for the venue itself, for example, fields such as venue size in square meters and business index. Secondly, it would be beneficial if the real-time solution data, or dynamic data fields, get updated if parameters change. Finally, the solutions need to include certain user inputs, such that the solution can be adjusted according to the user experience and expertise.
For (2), the model should be extended in the sense that the effect score calculation can be extended. As mentioned before, there could be different effects, such as social, economic, and cultural. The solution must focus on at least one effect, such that the current model can be extended and a more sophisticated score calculation can be achieved. Finally, for (3), the solutions must provide a visualisation in which the fire safety inspectors and planners can see the features of the venues and the effect score of a potential fire incident. For this, the solution is required to be a dashboard with user-defined parameters.

1.3 Solution proposition

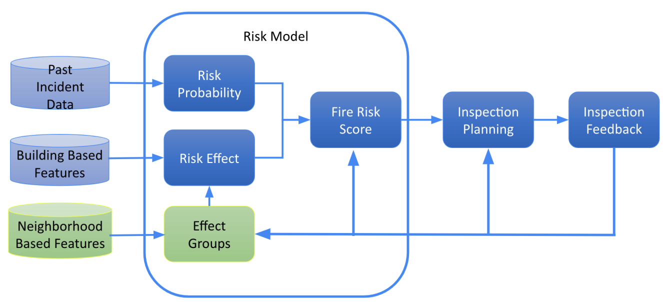

Figure 1: Proposed Neighbourhood Based Module

Figure 1 shows our proposed solution - a neighbourhood-based fire effect module (in green), in addition to the existing fire risk-effect model (in blue). Apart from the self-features of target venues, this module also collects neighbourhood features such as buildings, roads, and public facilities nearby. These neighbourhood features are subsequently used as input of unsupervised learning models to put the target venues into groups and suggest inspection prioritisation. The module should also be a dynamic system that takes input as parameters and feedback for clustering fine-tuning.

2 RELATED WORK

2.1 Fire risk model

This section presents an overview of the relevant studies performed in the fire risk model.

Ahmed et al.[3]developed a safety and risk assessment criterion for petrol stations based on one-year data to see the hazard trends and contributing factors. The data were analysed using a statistical package to perform regression and Pearson correlation to determine the correlation between these contributing factors. The author argues that the safety and risk assessment model helps the Health Safety and Environment professionals prioritise hazard contributing factors.

Further, Lau et al. [4] create a scoring system for various parties and individuals to assess the possibility of a fire outbreak. The study proposes a fire risk scorecard based on a scoring system used in the financial industry. The objective is to identify the risk level of industrial buildings and residential areas. Different fire risk factors are weighted by Analytic Hierarchy Process (AHP). The weighting is an important indicator to rank the factors in terms of the inherent risk from the professionals' points of view. SVM is chosen as a risk assessment model to provide prediction, and it has shown the accuracy of the scorecard. The author argues that the results of the scorecard and support vector machine (SVM) model are proven to be effective.

Besides risk assessment, Damaraju et al.'s [5] study claim that timely response is crucial to reduce the damage caused by fires. The study analyses how the data can tackle the problem by reducing the response time of firefighters. An interactive system developed using open-source tools has been introduced that uses visual analytics to understand fire incidents and their attributes better. Fire incidents are clustered using k-means to determine the ideal fire station locations, which results in a significant reduction in response times. The author argues that their approach is limited to fire incidents and can easily be extended to include multiple domains such as police stations, schools, and hospitals.

Based on the literature review, the fire risk module with a machine learning approach and an interactive system is proposed to tackle the municipality problem in the inspection.

2.2 Range of effect

The *Veiligheidregio* (Safety Region) in the Netherlands has developed a handbook of hazardous material accident scenarios containing information about incidents scenarios involving dangerous substances [2]. The handbook shows scenarios in which an accident involving hazardous substances can happen and what you can do to prevent, limit and combat it. The handbook can guide initiators of spatial developments and governments to include the dangers of hazardous substances in decision-making processes. Three scenarios that are relevant to this project are selected and explained in the following:

- Petrol A pool fire occurs when the tanker collides with another vehicle. As a result, the tank leaks, leading to a short and intense fire. A pool fire can cause heat, radiation, and smoke.
- LPG One of the Liquefied Petroleum Gas (LPG) stations incident scenarios is a Cold Boiling Liquid Expanding Vapor Explosion (Cold BLEVE) caused by external damage, such as a collision. Cold BLEVE is a ground-level cloud fire associated with a weak fireball tangential to the ground level[6]. The incident can cause heat radiation, overpressure, and sharding.
- CNG A flare fire occurs when a valve in one of the cylinders breaks after an incident, it causes Compressed Natural Gas (CNG) to flow out and burn all the cylinders in the buffer. The effect of a flare fire is heat radiation, and it can cause casualties, damage, and fire in the area.

[2]https://www.scenarioboekev.nl/tankwagen-lpg-koude-bleve/#fn-252-10

Cluster Group	Fuel Type	Effect Distance (Meters)	Consequences
Cluster 1	Petrol LPG CNG	30 80 11	99% Fatality Irreparable damage
Cluster 2	Petrol LPG CNG	30-50 80-200 11-14	1% Fatality Average damage
Cluster 3	Petrol LPG CNG	50-75 200-330 16	People who present are victims Minor damage

Table 1: Effect Distance and Consequences

The distance effect and damages are summarised in Table 1 and calculated based on various factors: fuel types and amount, size of leaks, ambient temperature, wind speed and direction, and surrounding area. The incident scenario in LPG shows that it has a significant effect distance; therefore, the default radius distance value will be 200 meters in this project. In this project, the system proposes a changeable radius distance as user input.

3 METHODOLOGY

3.1 Feature engineering

3.1.1 Feature selection.

We started feature selection based on literature and related work. According to the scenarios handbook developed by *Veiligheidregio* (Safety Region), the effect distance of incidents from different fuel types (petrol/LPG/CNG) ranges from 30 to over 300 meters. [3] Thus, the first feature selected for the prototype is fuel type(s) offered at the petrol stations. It is transformed into a series of binary variables using one-hot encoding to better utilize this feature in the clustering and scoring model. As a total of four types of fuels, namely diesel, petrol, LPG, and CNG, are offered in our target petrol stations around Amsterdam, four binary features are introduced into our prototype.

The effect of incidents at a petrol station would largely depend on the type of area where the station is place. Therefore, it is introduced as a categorical feature with categories *Residential*, *Business*, *Leisure* and *Industrial*. This feature is also transformed into four binary features using one-hot encoding. A series of numerical features are also included to gain more accurate results from clustering and scoring. Considering patients, children, and the elderly more vulnerable to the effects of fire incidents, numbers of hospitals, schools, children's daycare centres, and elderly houses are included as features. The numbers of hotels, stores, restaurants, gyms, and cinemas are also considered indicators of people needing to be evacuated in case of fire. Highways and public transport facilities, such as metro and tram stations, are also considered, as traffic and transport services could be suspended due to fire at petrol stations.

The duration of the fire is another main factor when the *Veiligheidregio* scenarios handbook calculates the fire effect. The duration would mainly depend on the arrival time of fire brigades. The travel time to a petrol station from its nearest fire station is added as another numerical feature as an indicator.

3.1.2 Data collection.

A cruel fact is that little information regarding petrol stations is available to the public from oil enterprises. The stakeholders cannot provide any data set due to a lack of details in the petrol station registrations at the municipality. Therefore, all data used in the prototype is collected from our efforts.

Data collection was started as a manual process. Basic information of petrol stations, such as locations and belonging companies, is collected from *Brandstof-Zoeker.nl* [4]. Subsequently, fuel types offered at each station are collected from official websites of respective oil companies, e.g. *Royal Dutch Shell* [5], *BP* [6] and *Esso* [7].

For the neighbourhood features, we started manual collection on *Google Maps* [8], utilising the *Measure distance* and *Search nearby* functions. This manual process is later replaced by automated and dynamic data update using the *Google Places API* [9] service. This API service takes the location (latitude and longitude) of a place, radius in meters, search types and search keywords as input and returns a list of places that fulfil the search criterion, which exactly automates our manual collection process. With radius removed as input, the API returns a list of places in ascending order of distance. By this means, we find the nearest fire station to a petrol station and, with *Google Distance Matrix API* [10], we get the travel time from the nearest fire station to the petrol station.

The initial range of neighbourhood features collection is set at 200m. As *Google Places API* is implemented, the range is designed as a dynamic input by the user. We choose not to calculate the radius based on the fuel types for two reasons: (1) we could not find scenarios on diesel from literature, and (2) we want to keep the possibility of this model to be extended and generalised for other venues apart from petrol stations.

Table 2 gives a summary of all the features in the prototype. Static features are features collected manually, while dynamic features are features collected using *Google API.*

3.2 Clustering and Scoring Analytics

3.2.1 Overview and Motivation.

The neighbourhood based fire effect module is designed to provide insights into the fire effect of target venues and give suggestions for inspection planning. The module consists of two main parts: clustering and scoring, as shown in figure 2.

[3] https://www.scenarioboekev.nl/overzichtstabel-scenariokaarten/

[4] https://www.brandstof-zoeker.nl/

[5] https://www.shell.com/

[6] https://www.bp.com/nl$_{n}l/netherlands/home.html$

[7] https://www.esso.nl/nl-nl/

[8] http://maps.google.com/

[9] https://developers.google.com/maps/documentation/places/web-service/overview

[10] https://developers.google.com/maps/documentation/distance-matrix/overview

Feature	Type	Collection
Fuel Type: Diesel	Binary	Static
Fuel Type: Petrol	Binary	Static
Fuel Type: LPG	Binary	Static
Fuel Type: CNG	Binary	Static
Neighbourhood Area: Residential	Binary	Static
Neighbourhood Area: Business	Binary	Static
Neighbourhood Area: Leisure	Binary	Static
Neighbourhood Area: Industrial	Binary	Static
# Highways	Numerical	Static
# Hospitals and Medical Centers	Numerical	Dynamic
# Schools	Numerical	Dynamic
# Children Day care Center	Numerical	Dynamic
# Elderly Houses	Numerical	Dynamic
# Hotels	Numerical	Dynamic
# Stores, Restaurants, Gyms, Cinemas	Numerical	Dynamic
# Public Transport Stations	Numerical	Dynamic
Travel Time from Nearest Fire Station	Numerical	Dynamic

Table 2: Features in Prototype

The clustering part takes the geocoding data as input, analyses the neighbourhood features, and uses four state-of-the-art clustering algorithms, namely K-means, DBSCAN, Mean-Shift and K-Prototype, to cluster the target venues: petrol stations in our prototype, into groups based on feature similarities. The user would select a certain number of petrol stations in each group for inspection and avoid putting all inspecting resources into similar venues in the same group.

As clustering is an unsupervised machine learning method, an effect score model is introduced in this module with weights from user input. This allows our target users, e.g. fire inspection planners, to put their expertise into the system and provide references to the clustering results.

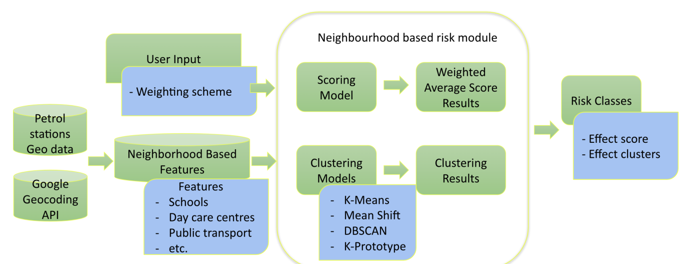

Figure 2: Details of Neighbourhood Based Module

3.2.2 *K-Means.*

The K-Means algorithm divides the clusters into a given number of clusters based on the mean distance between each cluster formed. [7] The K-Means algorithm identifies the k number of centroids, which are the imaginary centre point of a given cluster. The learning objective of k-Mean is to minimise centroids of a given k number of clusters. K-Means has been widely used for various clustering tasks because of their fast convergence. However, the K-value needs to be determined beforehand, which directly affects the convergence result. We have analysed the K-value determination using Elbow Method and Silhouette analysis to solve this problem.

The Elbow method used the square of the distance between points in each cluster and the centroid of the clusters to determine the optimum number of clusters in K-means.[8] Within Cluster Sum of Squares (WCSS) is defined as the sum of the squared distance between each member of the cluster and its centroid. By visualising the relationship between the number of clusters and WCSS, we select the number of clusters where the change in WCSS shows a rapid decline.

Silhouette analysis is used to study the separation distance between the results of a clustering task and estimate the optimum number of clusters in a clustering task. [9] The silhouette coefficient is calculated based on the mean intra-cluster distance and the mean nearest-cluster distance. The range of the silhouette coefficient is between 1 and -1, differentiating the best clustering model and the worst. The higher the silhouette coefficient, the better clustering results have been computed as further the distances between the clusters have been achieved. Values close to 0 indicate that there are overlapping clusters. Negative values suggest that there are points that have been assigned to the wrong cluster.

Although We decided to improve the model's accuracy by not relying on only one evaluation method, the result of the elbow method and silhouette analysis have arrived at the same result in our K-Means model. Using the elbow method, we have derived the optimum number of the cluster as four. The optimum number of clusters suggested by the silhouette method is 4. The silhouette analysis of the K-Means model has been shown on the figure 3 and 4. We have a similar average silhouette score when the number of clusters is three and four. The individual score of the clusters, when K is equal to three, is below the average score. When the number of clusters is four, the cluster size is more evenly distributed, and the individual silhouette score is above the average score.

The fact that the results of K-means could be computed relatively faster than other clustering methods makes it an attractive option for large datasets. Furthermore, K-mean can produce more concentrated clusters compared to hierarchical based clustering. It makes the clustering result more compact to the user as the expected number of clusters will be small.

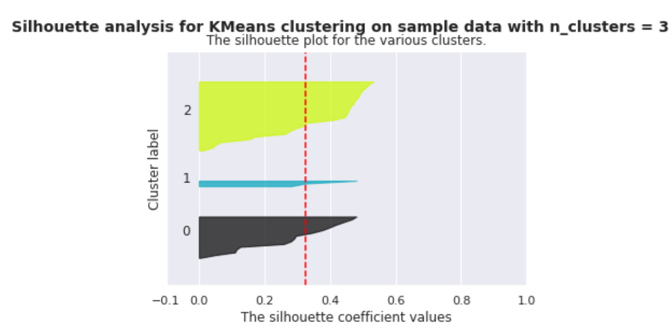

Figure 3: Silhoette Method

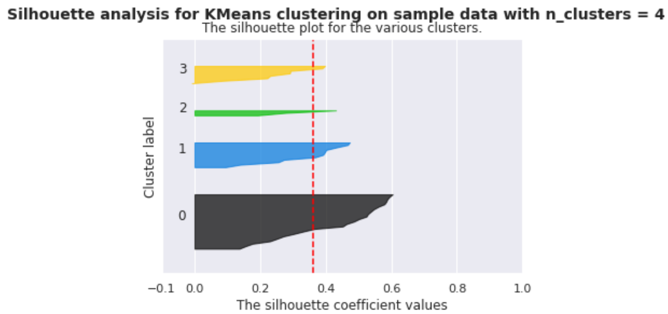

Figure 4: Silhoette Method

3.2.3 DBSCAN.

Although K-means is an extensively used model, it only works well for spherical clusters. Further, K-means are sensitive to outliers as outliers significantly impact the centroids before the centroids are converged and stabilized. On the other hand, DBSCAN works well with arbitrarily shaped clusters, and it is not sensitive to outliers. Density-based spatial clustering of applications with noise (DBSCAN) determined the clusters by separating the clusters with high density and low density. It views the clusters as areas with various densities; clusters found by DBSCAN are not constrained to convex shapes like k-means.

Two hyper-parameters for the DBSCAN are the epsilon and minimum points. Epsilon is defined as the maximum radius of the cluster in which the distance between two data points must not exceed. In other words, two data points will be considered as neighbours if their distance falls within the given epsilon range. Logically, a larger epsilon produces a larger cluster as the required radius has a wider range, the chance of more data points falling into the cluster is larger. On the other hand, a large part of the data might not be clustered if the epsilon is set to be too small. One way we could determine the epsilon of DBSCAN is to use the nearest neighbour distance. The average distance between each point and the nearest neighbours could be calculated. Based on the elbow graph, we could plot the average distance and determine the epsilon value. [10]

Another hyper-parameter for DBSCAN is the minimum point, defined as the minimum number of data points within the radius of a cluster. When the minimum point is low, more clusters will be formed with a small number of data points, and the clusters produce more noise. A higher minimum point ensures more robust clusters as more data points have been considered when forming the clusters. The minimum points are set to be greater or equal to the number of dimensions of the dataset.

Although DBSCAN can detect arbitrary shapes and handle outliers and data sets well, the tuning process of DBSCAN is, in fact, more complicated than K-Means. DBSCAN will categorise a high number of data points into the noise cluster if the data is highly sparse or data points with varying densities. A further limitation of DBSCAN is its low efficiency in handling high dimensional datasets.

3.2.4 Mean-Shift.

Similar to K-Means, Mean-Shift is a centroid based on the clustering algorithm. The Means-Shift is a non-parametric algorithm that generates clusters iteratively by finding the densest region in a feature space. The idea of Mean-Shift is to find the centre of a probability density. Given the region of interest, we expect that the centre of the mass is the mean. However, it might not be that case. The mean shift vector is derived from the distance between the centre of the mass and the actual mean of the region of interest. With the actual mean being identified, the region of interest will be shifted to the actual mean. This iterative process until the centre of the mass converges to the actual mean.[11] Mean-Shift is not sensitive to initialisation as it will converge at the end. However, the algorithm is not highly scalable regarding the computational time for large datasets as multiple nearest neighbour searches will increase the time consumption during the execution.

3.2.5 K-Prototype.

A problem of K-Means is that the cost function is calculated with Euclidean distance, which is only suitable for numerical data. Although the existence of K-mode provides a solution for categorical data, in situations where datasets contain both numerical and categorical data, a different solution is required. K-Prototype is a clustering algorithm that provides clustering solutions for mixed data types. It integrates the K-Means and K-Modes algorithms by combining the squared Euclidean distance measure on the numeric attributes and the dissimilarity measures for categorical objectives as the cost function.[12] K-prototypes removed the limitation of the K-means and extended the centroid-based clustering model into a generic data partitioning operation. The scalability has proved to be similar to the performance of K-Means on large datasets.

3.2.6 Feature Importance per Cluster.

Once the clustering results are obtained, meaning that it is known to which cluster the petrol station belongs, the results are used as labels to obtain the feature importance in each cluster. The cluster number has k different categories, then, for each petrol station, we apply logistic regression to calculate whether the petrol station can be classified into cluster i or not, where $i = 1, 2, ..., k$. It means we run for each cluster number a logistic regression. We use the normalized regression coefficients as an indicator for its feature importance from the regression. Normalization is done in the following way:

$$\text{Normalized Coefficient} = \frac{X_j - X_{\min}}{X_{\max} - X_{\min}},$$

where X_j is the coefficient of the jth feature. By normalizing, we transform the coefficient to a range $[0, 1]$, which provides a better interpretation of the magnitude of importance since 1 indicates maximum importance and 0 indicates minimum importance. The feature importance provides insight into how the clusters are composed, as it allows the user to see which features are highly important and which are less important. In this way, the user can see if, for example, the number of schools and public transportation is of higher importance in the respective cluster. It is important to note that higher feature importance is not associated with a higher count of buildings within the cluster. The feature importance merely tells the user what magnitude the feature played during the composition of the clusters.

3.2.7 Effect Score Calculation.

The calculation of the effect score depends on two elements: (1) the weights of the features, predefined and set by the user (the default weight is 0.5, meaning each feature is of equal weight), and

(2) the value of the feature. Then, the effect score is calculated of a weighted sum of the J features:

$$\text{Effect Score} = \sum_{j=1}^{J} w_j \cdot A_j,$$

where w_j is the user-defined weight and the A_j is the feature value. The weights are in a range $[0, 1]$, where 1 indicates maximum weight and 0 indicates minimum weight. Furthermore, the user should set the weights since the system assumes that the user values the weights from an experience and expertise perspective. This means that the user can set whether the values, for example, the number of hospitals more than the number of public transportation within a predefined and user set radius. It is important to note that the weights are only used for the effect score calculation and not in any other part of the module. The calculated effect score can be used as an additional module on top of the currently used effect score and shows the magnitude of a potential fire incident. By sorting the petrol stations based on the effect score, fire safety inspectors and planners can prioritise the venue visits. Also, the effect scores are shown within each cluster. Therefore, the user is able to view and sort high-effect petrol stations within each cluster.

4 DESIGN DECISION AND USER INTERFACE

The first design decision to make is what form the data system should be. To answer this question, the system's user must first be identified. The fire inspection resources like manpower are very limited, and the stakeholders need to decide the most efficient way to allocate those resources. Meanwhile, the inspection team has lots of experience in fire inspection that can be used for decision making. Therefore, the data system aims at combining the data science techniques and planner's expert knowledge about fire incidents to provide a view of all petrol stations from a neighbourhood effect based perspective. The form of the data system was then decided to be a clear presented dashboard, which can do all background calculations base on the user input parameters in real-time to combine the user's knowledge into the presented results. Figure 5 shows the main interface of the dashboard

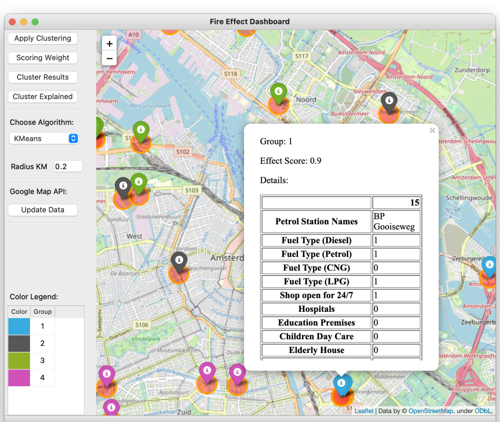

Figure 5: Main Interface of Dashboard

The dashboard user interface (UI) is created based on python to allow the calculation results to be updated in real-time based on the user's input. As shown in figure 5, An interactive map of all petrol stations in Amsterdam is shown, in which petrol stations are marked in different colours based on the clustering group they belong to. The user can choose from four alternative clustering algorithms to replace the recommended ones on the left side. The weights used for effect score calculation are also changeable in a dedicated weight edit window. This weight editor window is shown in figure 6, in which users can easily edit, save changes or reset the weight to default.

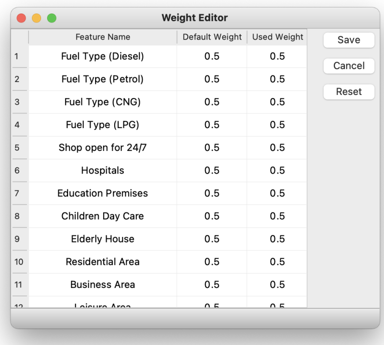

Figure 6: Weight Editor

To make it easier for the user to have an overall knowledge about the calculation results, the results are presented in two different perspectives: Sort by effect score in descending order grouped by clustering results. By switching between the two perspectives shown in figure 7, users can easily find out the petrol stations with the highest effect score among similar stations and the petrol stations that have the overall highest effect score.

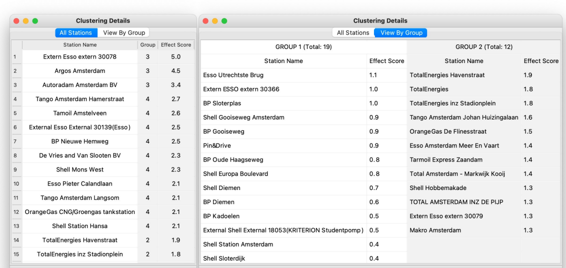

Figure 7: Two Perspectives of Clustering Results

The target user of the dashboard is the fire inspection planner of fire inspection and is usually not a data scientist. The terms used in the dashboard are carefully chosen to be understandable to users unfamiliar with clustering algorithms and the effect score calculation process. Besides, according to the stakeholders, decisions made by the government usually need to be supported by reasons.

Therefore, the feature importance of clustering was added to the dashboard to explain the clustering process. As shown in figure 8, a dedicated window shows the explanation to keep the main interface simple and easy to understand.

Importance of Features

1	Feature Name	Group 1 Importance	Group 2 Importance	Group 3 Importance	Group 4 Importance
2	Fuel Type (Diesel)	0.88	0.04	0.64	0.3
3	Fuel Type (Petrol)	0.88	0.04	0.65	0.46
4	Fuel Type (CNG)	0.88	0.04	0.82	0.67
5	Fuel Type (LPG)	1.0	0.02	0.68	0.08
6	Shop open for 24/7	0.84	0.06	0.8	0.49
7	Hospitals	0.65	0.06	0.64	0.35
8	Education Premises	0.58	0.19	0.52	0.67
9	Children Day Care	0.88	0.04	0.56	0.34
10	Elderly House	0.88	0.04	0.56	0.34
11	Residential Area	0.76	0.0	1.0	0.0
12	Business Area	0.81	0.06	0.72	0.65
13	Leisure Area	0.8	0.0	0.94	1.0
14	Industrial Area	0.92	0.04	0.22	0.98

Figure 8: Clustering Explanation

The ability to generalise is also an important demand from stakeholders. All data used in the dashboard are stored in Excel files so that the system can be applied to other venues easily by replacing data. Data can also be updated through integrated Google Map API according to user-specified effect radius, which can also increase generalisation because the dashboard can stay up to date and meet different data requirements automatically.

5 APPLICATION ON PROTOTYPE

In this section, a scenario will be discussed on how the data system could be used to obtain more insights and calculate the effect score of the fire risk safety model.

As mentioned before, the radius can be adjusted according to the user input. The user must define which radius to use beforehand, so the Google API updates the feature data accordingly. We will assume the default radius of 200 meters for the remainder of this scenario. Also, we will consider that the user is a fire safety inspector or planner.

5.1 Inserting Weights

First, the user is expected to insert the weights for the features that will be used for the effect score calculation. The weights could be set according to expertise and experience. The default weight is set to 0.5 for each feature, but let us assume that the user values the feature "Fuel Type - Diesel" slightly higher, i.e. the user sets its weight equal to 0.6. This may be because the user knows that diesel might bring higher effects in a fire incident. All other weights are kept to 0.5.

5.2 Clustering

After the user inserts the weights, the fire inspector or planner will choose the appropriate clustering method, i.e. the grouping of the petrol stations based on their characteristics. The user can choose four algorithms, but the user might not be aware of the details of these algorithms. Hence, the recommended method would be to use the k-prototype grouping method since that is the most appropriate method for the data type. If the user wants to see the effects of outliers in the data, DBSCAN would be an option to use. The other two algorithms would be recommended to use when there is only numerical data.

Petrol Station	Group	Effect Score
Extern Esso extern 30079	2	5.0
Argos Amsterdam	2	4.5
Autoradam Amsterdam BV	2	3.4

Table 3: Top 3 Petrol Stations with Highest Risk

	Feature Importance Cluster 1	Feature Importance Cluster 2	Feature Importance Cluster 3
Fuel_Type_Diesel	0.058047	0.000932	0.177148
Fuel_Type_Petrol	0.058050	0.000933	0.215236
Fuel_Type_CNG	0.258272	0.001029	0.092149
Fuel_Type_LPG	0.063754	0.026352	0.097685
Shop_Open_24/7	0.050541	0.003167	0.544168
Hospitals	0.211350	0.000134	0.398507
Education_Premises	0.123840	0.176778	0.000000
Children_Day_Care	0.258661	0.247130	0.058754
Elderly_House	0.033156	0.102840	0.133056
Residential_Hotel_Area_R200	0.089642	0.043217	0.245029
Business_Area_R200	0.093790	0.025394	0.226988
Leisure_Area_R200	0.200964	0.050471	1.000000
Industrial_Area_R200	0.122096	0.000000	0.201090
Public_Transport_R200	0.079710	0.021976	0.033180
Highways_A	0.144726	0.050471	0.317865
Highways_N	0.000000	0.050623	0.431050
Highways_S	0.008878	0.009884	0.278356
Hotels	0.082220	0.035467	0.617858
Store_Restaurant_Gym_Cinema	1.000000	1.000000	0.060248
Distance_FireStation_Min	0.225430	0.048449	0.011257

Figure 9: Feature Importance from Clustering

5.3 Result Interpretation

Now, let us assume the user selects the k-prototype grouping method and applies the algorithm to the data. Then, the user will be able to look into the cluster results, where first, all petrol stations are shown together with their effect score. The petrol stations are sorted from high to low based on the effect scores. Table 3 shows the top three petrol stations with the highest effect score. Also, the user can see the results per group. Then, the user would view each group and see which petrol stations have the highest effect score in a group. It would help the user by prioritising the fire safety visits, meaning that the user could visit the top 3 petrol stations in each cluster based on the effect score.

Another functionality is that the user is able to view the cluster explanation. In this view, the user will find the feature importance per cluster. Figure 9 shows the feature importance in each cluster. From here, the user is able to see how the features are valued within a cluster. As we saw that the top 3 petrol stations with the highest effect score are from cluster 2, the user could see the feature importance to understand the cluster composition. We see from the Figure that the number of Stores, Restaurants, Gyms and Cinemas are the most important in cluster 2 and that the presence of an industrial area is the least important in cluster 2. The second and third highest important features are the number of children daycare and the number of education premises, respectively. It means that these three features had played the most important role during the cluster composition.

6 USER VALIDATION

A usability test is an essential part of the evaluation process to validate the design of the dashboard developed for this project. The data scientist from the Amsterdam municipality who has little knowledge of fire risk in the petrol station tested the dashboard. The test was carried out to assess their experiences using the dashboard. The main purpose of this test was to discover any problems in usability. A list of tasks (see Appendix A.5) was created and given to the participant. And the participant needs to accomplish and comment on their thoughts in the process. Overall, the user was positive about the design and functionality of the dashboard. The participant was able to navigate within the pages and identify possible actions and complete the tasks given to them in full. However, an issue was encountered when doing task 2 to find the important feature. The user was confused between two buttons, "Cluster Results" and "Cluster Explained.". For this, the user suggested using the right word in label buttons to make it user-friendly. Another remark from the user is that there are a lot of functions in the dashboard that requires some explanation for other user groups (e.g., inspectors in fire departments) if they use this dashboard. The user suggested having the radius by the fuel types offered at the stations.

Further, the test was done only once due to the limited number of respondents and time. The remarked about the label buttons were adjusted. The user handbook explaining some functionalities and algorithms will be a future work recommendation.

7 EXTENSION AND SCALABILITY

7.1 Target Venues

Although petrol stations are chosen as the target venues of our prototype, the neighbourhood based fire effect module could be extended to all non-residential venues. One reliable data source of non-residential addresses in Amsterdam is the *Functiekaart* (Function Card) from the municipality as shown in figure 10. Different types of clustering models and scoring weights could be applied to various types of venues in a generalised system.

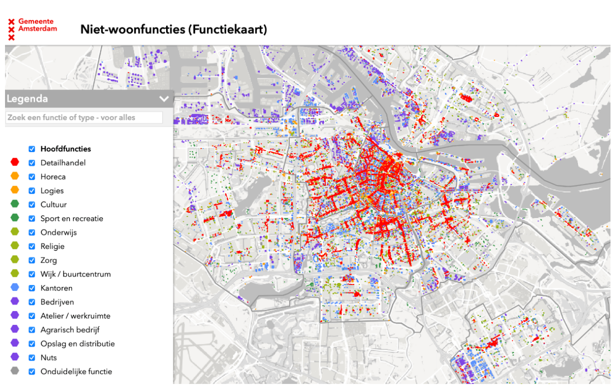

Figure 10: Non-residential Addresses in Amsterdam

7.2 More Features and Details from Google API

As almost all addresses accessible to the public are available on *Google Maps API*, neighbourhood features could be added or removed from the module if necessary. In addition, more details could be added to the existing features to improve clustering and scoring results. E.g. typical traffic data of roads (figure 11) could be an enhancement to the current *number of highways* feature. Popular times (figure 12) could also improve the *number of stores, restaurants, gyms and cinemas* feature. Still, unfortunately, it is yet to be available on the *Google Places API*.

Figure 11: Traffic on Google Maps

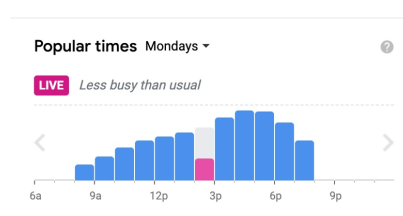

Figure 12: Popular Times on Google Maps

8 REFLECTION AND LIMITATIONS

8.1 Features and Data

One of the biggest challenges of this project is the lack of details of the target venues. In the *Veiligheidregio* handbook, fire effects at petrol stations are estimated with assumptions of various factors such as mass of fuel [11] and volume of fuel tank [12]. These factors determine the range of effects and should be essential information of oil enterprises. But unfortunately, they are unavailable to the public. To further develop the neighbourhood based module, detailed information from the owners is necessary for petrol stations

[11] https://www.scenarioboekev.nl/tankwagen-benzine-plasbrand/#fn-885-4
[12] https://www.scenarioboekev.nl/lng-tankwagen-fakkelbrand/#fn-240-1

or other target venues.

8.2 Clustering model

8.2.1 Distance measures of mixed data types in the clustering model. As we the prioritized tasks using scrum methodology, the prototype was built in short period, evaluated and improved periodically. For the distance measure of the clustering task, we have chosen Euclidean distance for simplicity. Euclidean distance is the most common and intuitive measure, most known as the shortest distance between two points. It is the square root of the sum of squared differences between two vectors. However, it does not work well with binary variables. One possible distance measure for binary variables is the Manhattan distance, which is the actual path of the points. We encountered a mixed dataset consisting of both numerical and binary features in our situation. Splitting the data into two groups and using different distance measures could be a solution, but not the best. Gower has proposed heterogeneous distance measures to solve the problem of mixed data types [13]. Gower's distance is computed as the average of partial dissimilarities across individuals. We decided to follow a more recent development of distance measure which is an ensembles based distance measure proposed by Huang [12], known as K-prototype clustering. With unprecedented development in machine learning, there are more extensions of Huang's model available for further development.

8.2.2 Dimensionality in clustering.
Although we have found a solution dealing with mixed data types, we still have more than fifteen features. The concept of distance becomes less precise as the number of dimensions grows. One general approach is to reduce the dimension of the features, for example, using PCA. However, it might be challenging to retrieve the feature importance after reducing the dimension and as clustering result will be hard to interpret in accordance with the features. Some suggested carefully choosing the distance measures, and one alternative could be the cosine similarity as the orientation of the vectors could be taken into account compared with Euclidean distance. However, the curse of dimensionality is due to the small sample size, which makes it hard to infer given the large feature space. Next to that, the association/ correlation between the features will also negatively impact the distance measure.

8.3 User Validation Process

Due to public health policies by the government, usability tests had to be executed remotely on *Microsoft Teams*. In our original plan, the test user could gain access to our prototype dashboard via *Remote Control*, but unfortunately, the function was restricted by the municipality on its employees, and the test user had to verbally control the test facilitator to perform the test tasks, which was not the ideal way of validation. Given different times and circumstances, the usability test could have been arranged in a standard manner.

8.4 Reception of Target Users

Through the presentation and user validation process, although the prototype gained good reception on creativity and visualisation, we also found limitations in the prototype. On the one hand, four state-of-the-art algorithms are adopted for best clustering performance. Still, on the other hand, these algorithms might be too complicated for our target users - the fire inspection planners, who might have little knowledge of data science methods. Although we have introduced feature importance to interpret the clustering results according to the critical reviews obtained from the first demonstration to the stakeholders, model explainability is still a limitation on our final presentation.

REFERENCES

[1] Mirza Ahmed, S.R.M. Kutty, Dr Mohd Shariff, and Dr. Mohd Faris Khamidi. Petrol fuel station safety and risk assessment framework. pages 1–8, 09 2011.

[2] Hong-yu Zhang. The research about fire prevention of vehicle refuelling stations. *Procedia Engineering*, 71:385–389, 12 2014.

[3] Mirza Ahmed, S.R.M. Kutty, Dr Mohd Shariff, and Dr. Mohd Faris Khamidi. Petrol fuel station safety and risk assessment framework. pages 1–8, 09 2011.

[4] Chun Kit Lau, Kin Keung Lai, Yan Lee, and Jiangze Du. Fire risk assessment with scoring system, using the support vector machine approach. *Fire Safety Journal*, 78:188–195, 11 2015.

[5] Nandita Damaraju, Subhajit Das, Andrea McCarter, Joe Minieri, Sriram Padmanabhan, and Duen Horng Chau. Mitigating fire risks using visual and data analysis. *Bloomberg: Data for Good Exchange*, 09 2015.

[6] André Laurent, Laurent Perrin, and Olivier Dufaud. Consequence assessments of a cold bleve. can we do it better? *Chemical Engineering Transactions*, 48:211–216, 01 2016.

[7] Joaquín Pérez-Ortega, Nelva Nely Almanza-Ortega, Andrea Vega-Villalobos, Rodolfo Pazos-Rangel, Crispín Zavala-Díaz, and Alicia Martínez-Rebollar. The k-means algorithm evolution. In *Introduction to Data Science and Machine Learning*, chapter 5. IntechOpen, Rijeka, 2020.

[8] Chunhui Yuan and Haitao Yang. Research on k-value selection method of k-means clustering algorithm. *J*, 2(2):226–235, 2019.

[9] Meshal Shutaywi and Nezamoddin N. Kachouie. Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*, 23(6), 2021.

[10] Sander-J. Ester M. Kriegel H. P. Xu X Schubert, E. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3), 2017.

[11] Comaniciu Dorin and Meer Peter. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[12] Joshua Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2:283–304, 2004.

[13] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.

A APPENDIX

A.1 Prototype Repository

The repository of the prototype is located at *GitHub*: https://github.com/jiujiucs17/DSP2021-Team-D2.git

A.2 Project Planning and Task List

A copy of the project planning and task list is available at: https://docs.google.com/spreadsheets/d/17gdr6LMWPAQCPFO8HbtHVGoDPWIjLOliVFef0M_g04w/edit?usp=sharing.

A.3 Pitch Video

The pitch video is available at: https://www.youtube.com/watch?v=P2PhRxv2rwU

A.4 Demonstration Video

. The video with a demonstration of the prototype is available at: https://www.youtube.com/watch?v=mWjkjl_mmDk.

A.5 Usability Test - Task List

(1) Basic operations
Goal: The user should know where to find the features and the group information to which the petrol station x belongs after applying the K-means cluster.
Tasks:
  (a) Apply K-means clustering method
  (b) Choose one petrol station and find out which group it belongs to
  (c) Choose one petrol station and find out which group it belongs to
  (d) Find the effect score of the chosen petrol station

(2) Feature importance
Goal: The user should know where to find the feature importance of each cluster.
Task:
  (a) Find the most important feature(s) of Group 2

(3) Changing parameters
Goal: The user should know where to adjust the weight in Scoring Weight and check the highest effect score in Clustering Details.
Tasks:
  (a) Adjust weight for hospitals to 0.8 and recalculate the score
  (b) Switch to the K-prototype clustering method and apply clustering
  (c) Find the petrol station with the highest effect score

(4) Change of radius and data update
Goal: The user should know how to change the radius and update data through API.
Tasks:
  (a) Change the radius of 0.2 km to 0.5 km
  (b) Update data via Google API
  (c) Re-apply clustering

(5) Resetting parameters
Goal: The user should know how to reset the data that has been put from the previous task and enter a new weight number in the features.
Tasks:
  (a) Reset the weights to default values
  (b) Adjust weight for Elderly_House to 0.2 and recalculate the score
  (c) Find the petrol station with the lowest effect score