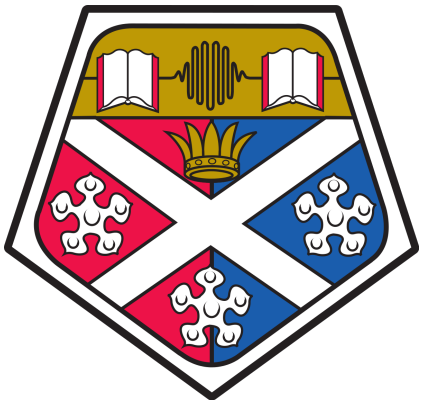
Generation and Management of Household Waste in Scotland

CS982 Big Data Technologies Coursework



University of Strathclyde
Glasgow
9th November 2020

Contents

List of Figures

List of Tables

1 Introduction

We live in an age of acute environmental crisis. Between its many aspects is the issue of global waste management, which in recent years has emerged as a deeply concerning and complex problem. Over-consumerism and indiscriminated production of non recyclable waste by both developed and developing nations are having a devastating effect on nearly every habitat on Earth. Now more than ever is critical to carefully monitor the production and management of waste, and to develop strategies to tackle this problem.

In Scotland, the Scottish Environment Protection Agency (SEPA) is responsible for the collection, assessment and reporting of all national waste data [1]. This activity is critical for policy making and the development of waste reduction strategies such as the national Zero Waste Plan [2].

In the following report a dataset containing information on the generation and management of household waste in Scotland was taken into consideration. Data is sourced from Scottish local authority returns as reported using the WasteDataFlow system [3], and published by SEPA on the official government statistics website [4]. It falls under the open government licence (version 3) [5] which allows for free reuse and adaptation.

After introducing the dataset structure and characteristics a few key questions and challenges are highlighted in the next section. Section 2 also contains an in-depth look into some of the statistics that emerge from the data. In the sections that follow two further kinds of analysis are conducted on the data: unsupervised clustering and supervised classification, with the aim to extract more sophisticated insights and model the overall behaviour of the dataset. The report concludes with some reflections on the methodology and approach used and concluding remarks.

2 Dataset

2.1 Structure and key challenges

The term Household Waste used in this dataset and its accompanying documents is defined in Paragraph 1.2 of the *Zero Waste Plan - guidance for local authorities* document [6] from the WasteDataFlow website as:

> Waste from households including household collection rounds, other household collections such as bulky waste collections, waste

deposited by householders at Household Waste Recycling Centres and recycling points/bring banks.

It does therefore not include any waste generated by commercial or manufacturing endeavours, but only the waste individual citizens can directly be called accountable for. The original dataset as obtained from the statistics.gov.scot website contains almost 40.000 observations organised in 7 columns, briefly described in Table 1.

Column Name	**Description**
FeatureCode	Unique code identifying one of the 32 Scottish Councils
DateCode	Year of the observation (2011 - 2018)
Measurement	Type of entry, can be either Count or Ratio
Units	Type of units of entry, can be either Tonnes or Percentage
Value	Numerical value of the entry
Waste Management	Way in which the entry was managed
Waste Category	Type of waste of the entry

Table 1: Description of the columns of the original dataset

In order to carry out a more focused analysis a number of pre-processing steps were carried out. Firstly only the entries reflecting the amount of waste generated in tonnes were kept (i.e. the ratio/percentage information was dropped), secondly the rows containing zero values were also dropped, alongside the measurements that rounded up the total waste generated for each year and the nationwide numbers (as this is redundant information that will nonetheless emerge from the analysis of the time and geographical features). The FeatureCode were replaced with the corresponding and more intuitive Councils names.

At this stage the "Waste Management" column included 6 kinds of entries: Landfilled, Other Diversion, Other Diversion (pre 2014 method), Recycled, Recycled (pre 2014 method). This distinctions captured a change in management practices which took place in 2014. To simplify the analysis they were merged into just three categories:

- **Landfilled**: waste disposed to a landfill facility.
- **Recycled**: waste prepared for reuse, recycled and organics recycled.

- **Other diversion**: waste diverted from landfill, apart from waste recycled, comprising waste disposed by incineration, recovered by incineration, recovered by co-incineration and waste managed by other methods.

The "Waste Category" column includes 20 different kinds of waste ranging from Animal and mixed food waste to Textile, Plastic and many others. A class which will be revealed to be of particular interest is "Household and similar wastes", this sounds counter-intuitive within a dataset already comprising of household waste, but an investigation of the WasteDataFlow documents revealed this category to be including "Bric-a-brac, Furniture, Mattresses and General residual waste ("black bag" waste)", perhaps the results of fly-tipping or improper disposal.

At this stage the dataset was ready for analysis. A few preliminary questions immediately emerge: which councils recycle the most? what is the most landfilled type of waste? what is the most common type of waste produced? so, more generally, what are the statistical trends present in the data?

These questions are answered in the following section. Another kind of question also emerges: does the data contain unseen patterns? can we built a model able to predict how waste will be dealt with based on its kind and where it was produced in Scotland? A critical attempt at answering these questions was made in section 3 and 4.

2.2 Summary statistics

Figure 1 shows how the time evolution of the total amount of waste produced and how it was managed nationwide, from 2011 to 2018. From this figure we can see that the amount of waste sent to landfill has been steadily reducing since 2011, while the recycling and "other diversion" rates have been increasing, which is perhaps a positive trend. The total amount of waste produced has nonetheless been fairly constant, at a mean value of 3.7 million tonnes per year. The mean amount of waste sent to landfill was 1.2 million tonnes per year, with a maximum of 1.45 million tonnes in 2011 and a minimum of 1.03 million tonnes in 2018. The amount of waste managed via "other diversion" (this is mainly incineration facilities) has increased from a minimum of 75,394 tonnes in 2011 to a maximum of 5.7 million tonnes in 2018, while the amount recycled has ranged from a minimum of 2.03 million tonnes in

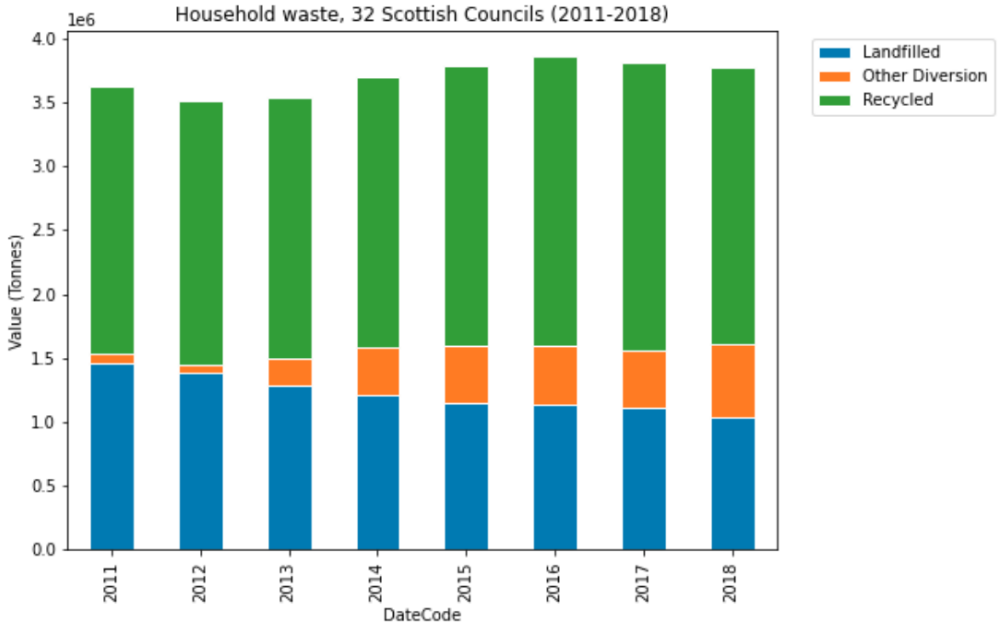


Figure 1: Household waste produced and managed - distribution over the years 2011-2018

2013 to a maximum of 2.26 million tonnes in 2016, with a mean value of 2.15 million tonnes per year.

The marked increase in the amount of waste managed via "other diversion" reflects Scotland's investment into so called "Energy form Waste" (EFW) incineration facilities [7], several of which have been developed by private companies in different Scottish councils. EFW facilities produce electricity and heat via the incineration of not-recyclable residual waste. This has drawn both praise and criticism from experts, mainly because the incineration of waste is known to release significant amounts of greenhouse gases, which can potentially be more damaging than landfill if not dealt with correctly [8].

Figure 2 shows how the total amount of waste produced nationwide was distributed across the 20 waste types, as well as how it was managed. The most recent year 2018 is taken as a reference, but the same trends are present in the previous years. The most evident insight from this graph is that by far the most common type of waste produced is the one labelled "Household and similar", 1.39 million tonnes of it were produced only in 2018. This is

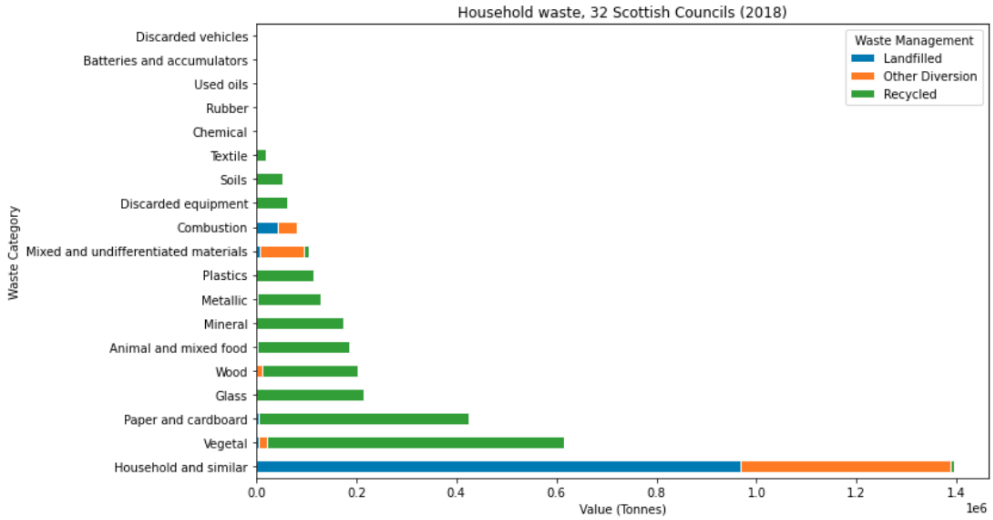


Figure 2: Household waste produced and managed - distribution over waste types (2018)

also the type of waste with the highest landfill rate, nearly 70% of it was sent to landfill in 2018, while only 0.5% was recycled and 30% was sent to incineration facilities.

This makes for a worrying trend. As mentioned above, Household and similar waste is defined as general residual waste or "black bag" waste. This kind of waste is perhaps the result of improper or illegal disposal, and it is almost certainly going to be sent to landfill. It is waste which is non-recyclable or that has not been identified and sorted correctly, and it is therefore rarely recycled. On the other hand, Figure 2 also shows that when the waste has been correctly classified and sorted it has a high chance of being recycled, like paper and cardboard, 84% of which was recycled in 2018.

Finally Figure 3 shows the distribution of the data across the 32 Scottish councils for the reference year 2018. An initial insight into this graph reveals that the most populous regions such as Glasgow City and City of Edinburgh account for the highest amount of waste produced nationwide, as one would expect. This together with the fact that not all councils have access to the same management facilities or have the same budget leads to highly different rates of recycling, landfill or other diversion, and makes the different councils not directly comparable.

Glasgow City accounts for the highest percentage of waste sent to landfill,

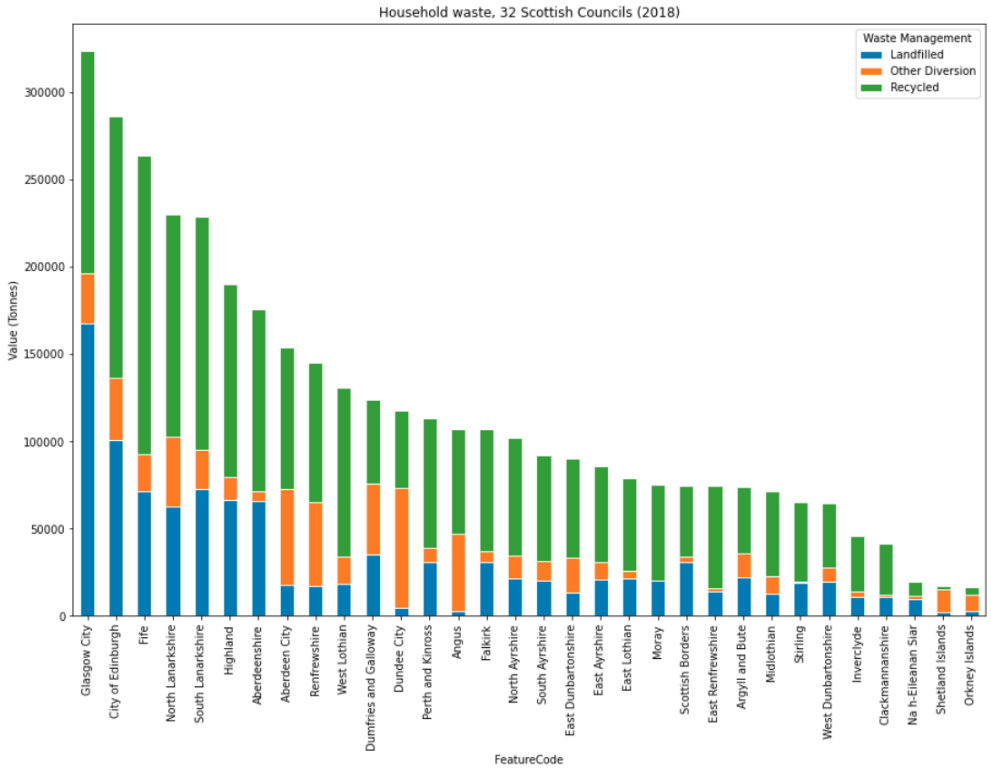


Figure 3: Household waste produced and managed - distribution over Scottish councils (2018)

52% of its total in 2018; while the best recycling performance belongs to East Renfrewshire, with 79% of its waste recycled in 2018 (Glasgow City council has roughly 6.5 times the population of East Renfrewshire council). An interesting insight is the large percentage of waste managed via other diversion in Dundee City, 59% of its total and the highest rate in the dataset. A quick investigations reveals that this is probably due to a large investment into a Energy From Waste incineration facility, which happened in 2017 [9].

3 Unsupervised Learning

The initial rationale of this part of the analysis was to attempt to bring to light intuitive groupings within the dataset, to investigate for example, if the different waste types can be clustered by waste management strategy,

or geographical and time information. This kind of information could then be useful in targeted planning of new management facilities, if for example, a significant cluster of one type of waste is found in a specific local council (this information is also deducible from summary statistics, but being able to look at the dataset as whole rather than slices of it, can often be more insightful).

What becomes apparent is that this specific dataset is not particularly suited for clustering analysis. This is because many clustering algorithms, such as hierarchical and k-means, are centred around the concept of *distance*, i.e. the metric or separation between different data points in n-dimensional space, where n is the number of attributes associated with each point. Clustering analysis is then a process of grouping data points which are closest to each other, as this is taken to be a measure of similarity. This is one particular understanding of what a cluster is.

When dealing with purely numerical datasets the distance between points is calculated using metric measures such as Euclidean or Manhattan. In the case of categorical variables such as the ones present in this dataset, the very concept of distance is not so straightforward. It becomes important to preserve the logical relationship between the variables, as for example, there is no direct measure of distance between "plastic" and "paper and cardboard", because these attributes have no ordinal relationship.

In order to carry out clustering the dataset was split between attributes ("Value", "FeatureCode", "DateCode", "Waste Category") and outcome ("Waste Management"). The "Value" variable, being on a continuous scale, was binned into 10 bins (aka deciles), and was then considered as a discrete categorical variable. The attributes data was encoded using one-hot encoding, a choice that preserves the logical independence of each variable by creating new binary columns indicating the presence of each possible value from the original data. The outcome labels were encoded as integer values: 0 = Landfilled, 1 = Other, 2 = Recycled.

In order to assess the goodness of the clustering three different metric were used:

1. **Silhouette Score**: measures how similar an object is to its own cluster compared to other clusters. Ranges from -1 (poor fit) to 1 (perfect fit).

2. **Completeness Score**: measures proportion of data points of a single type that are elements of the same cluster. Ranges from 0 to 1 (all items are in the correct cluster)

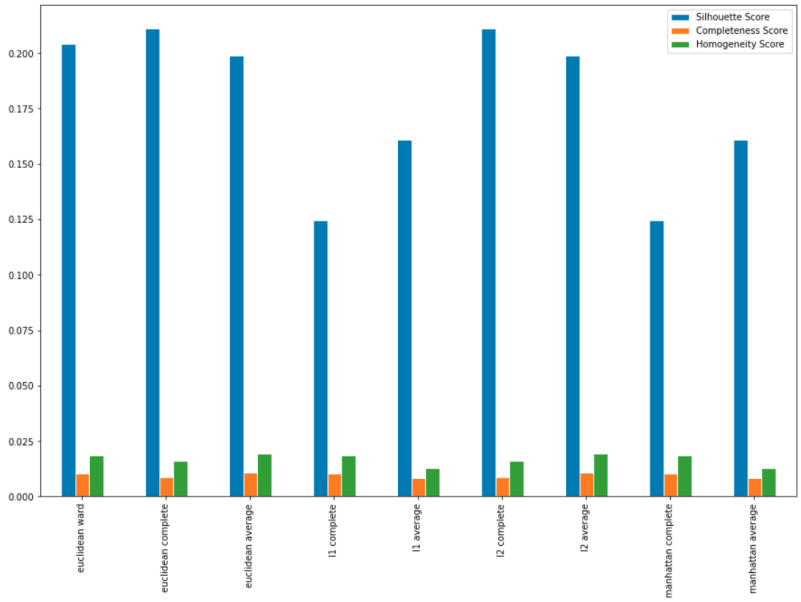


Figure 4: Comparison of agglomerative clustering methods

3. **Homogeneity Score**: measures proportion of of data points within a cluster that are of the same type. Ranges from 0 to 1 (all items in a cluster are of the same type)

Two different attempts at clustering were carried out.

3.1 Hierarchical Clustering

Hierarchical agglomerative clustering is a "bottom up" approach in which each datum starts in its own cluster, the clusters are then progressively joined according to their measured similarity. This is achieved via a linkage criterion, which in the following analysis can be *ward*, *complete* or *average*.

Figure 4 shows the resulting scores for each of the clustering methods, with different distance metrics. All of the metric scores are poor, and especially the completeness and homogeneity scores, which do not rise above 0.025, cast doubts on the efficacy and meaningfulness of this type of analysis for this dataset.

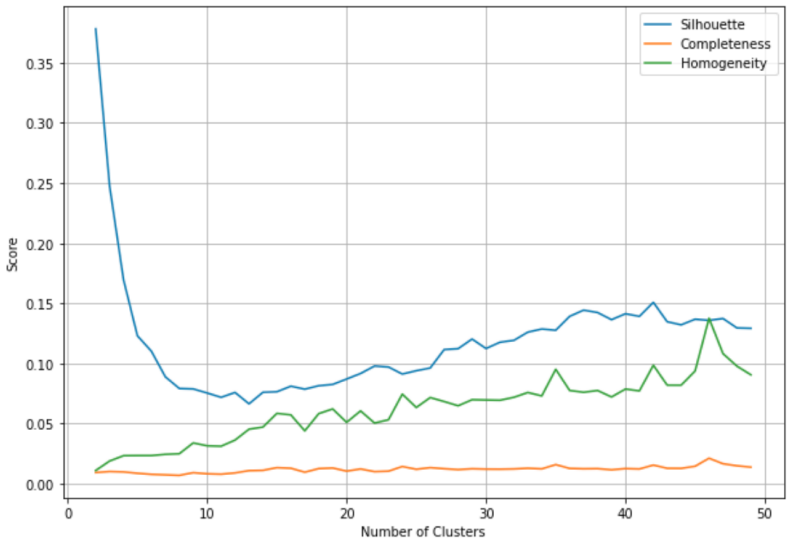


Figure 5: K-Means clustering for different numbers of clusters

3.2 K-Means Clustering

A second attempt at clustering was made using a k-means approach. This method needs the number of clusters n to be specified in advance. A randomly selected group of n centre points (centroids) for each cluster are selected by the algorithm, which then assigns each other point to its closest cluster, and calculates new centroids. This process is repeated iteratively for all points, and the algorithm will continue to optimise the centroids positions either until a defined number of iterations has been achieved or the centroids have stabilised.

Figure 5 shows the result for the k-mean clustering attempt, which was repeated for 2 to 50 clusters. If the method was successful one could expect to see 3 clusters scoring best, as this is the number of unique outcome variables (Landfilled, Other, Recycled). But this is not reflected by the result. The silhouette score is highest for 2 and 3 clusters, but again the completeness score does not rise above 0.05. The homogeneity score stays below 0.1 for most number of clusters, and has a peak for 45 clusters. This results points again to the fact that this kind of analysis might not be returning meaningful results.

4 Supervised Learning

In the final part of the analysis we wish to build a classifier model capable of predicting how waste will be managed depending on its kind, its amount, and where and when in Scotland it was produced. This is achieved via supervised learning, an approach in which the algorithm learns to classify unseen samples by being exposed and optimised on a training set of correctly labeled examples.

In the present case the same preprocessing steps outlined in section 3 were applied, and the task was then to correctly classify the entries as "Recycled", "Other" or "Landfilled". To carry out this process the data was split in a random training set (80% of total) and testing set (20% of total). Several different multiclass classification algorithm are available but for the present case a Decision Tree algorithm was selected.

4.1 Decision Tree

A decision tree is an algorithm which uses an intuitive recursive decision process to map set of attributes to outcomes in a flowchart-like structure. The branches of the tree then represent chains of relationships or decisions.

The algorithm starts from one root attribute and assesses the data underneath it. If the attribute in the considered split can be unequivocally linked to one outcome, e.g. "all paper and cardboard entries, regardless of where and when they were produced, are always recycled", then the split is considered to be pure. If the split is not pure, which is often the case, then the algorithm calculates the most advantageous further split and again assesses the data underneath it. This process is repeated until all the splits are pure or another stopping criterion is applied to prevent overfitting. The result of this process is then a tree-like model, which represents chains of decisions that can be followed to predict the outcome of unseen data.

The performance of any classifier is evaluated by applying the trained model to unseen test data and creating a confusion matrix of the predicted outcomes. This is a table that contains the number of correctly and incorrectly classified entries for each class. Table 2 shows the confusion matrix for our decision tree attempt.

As a reference the test data used comprised 2001 entries, 162 labelled Landfilled, 171 labelled Other Diversion and 1668 labelled Recycled. Just by glancing at this table we can see that the classifier succeeds in correctly

		Predicted		
		Landfilled	**Other Diversion**	**Recycled**
Actual	**Landfilled**	119	31	12
	Other Diversion	21	137	13
	Recycled	3	18	1647

Table 2: Confusion matrix of the decision tree classifier

predicting the correct class for most of the unseen data. To further probe the performance, the precision, recall and f1-scores are calculated for each predicted class. These are defined as follow and presented in Table 3.

1. **Precision**: ratio of correctly predicted observations of each class to the total predicted observations - e.g. of all entries predicted as Recycled, how many were actually Recycled?

2. **Recall**: ratio of correctly predicted observations to the all observations in actual class - e.g. of all the entries that truly were Recycled, how many did we label?

3. **F1-score**: weighted average of Precision and Recall.

	Precision	**Recall**	**F1-score**
Landfilled	0.83	0.73	0.78
Other Diversion	0.74	0.80	0.77
Recycled	0.99	0.99	0.95

Table 3: Performance measures of the decision tree classifier

All of the performance measure presented in Table 3 range from 0 to 1, with higher values indicating a higher classification accuracy. This decision tree model then, once trained, performs very well and can be used for classification of future unseen data.

5 Reflections and Conclusion

A detailed analysis of data on the generation and management of household waste in Scotland was conducted and presented here. After exploring the

dataset and looking at different summary statistics measures, one unsupervised and one supervised learning method were applied to the data.

The figures summarising the data presented in section 2.2, proved to be the most useful tool for extracting information from this dataset. A wealth of different insights could be gained just by looking at how the data was distributed in time, space and across the different waste types. Combining this with other contextual data (such as for example Scotland's investment in EFW facilities) additional links were made. Future work can bring this even further, by for example including information about demographics and wealth distribution across the 32 Scottish councils, and perhaps finding more interesting links.

As mentioned throughout section 3 the unsupervised clustering approach was largely unsuccessful, this was probably due to how the data was structured. I know that clustering of categorical data is in principle possible by using distance measures such as the Hamming distance, but I struggled to implement this correctly, and in the end resorted to use one-hot encoding of the variables. Both hierarchical and k-means clustering returned results which I believe are not particularly meaningful. Future work in this direction could include the use of clustering algorithms more suited to categorical variables, such at example k-modes [10].

The supervised classification approach explored in section 4 returned a very positive result, scoring high in both precision and recall. The data distribution across the three labels was discovered to be very skewed towards Recycled, which can perhaps be explained by the fact that only correctly sorted waste can be recycled, while unsorted waste is largely sent to landfill or incinerated. This means that most of the different waste types were labelled Recycled, while the unsorted "Household and similar" waste type counted for almost all entries labelled Landfilled which were significantly less in number but included the largest values in tonnes. The trained model can now be used to predict the labels of future data.

As mentioned in the introduction, there is not a better time than now to gain insight and critically assess the waste management strategy of our country. Awareness of how the waste we citizens produce is or can be handled by our local authority is crucial, and it will hopefully lead to us taking responsibility of what we consume and throw away, and make positive steps to change our habits.

References

[1] https://www.sepa.org.uk/environment/waste/waste-data/

[2] https://www.gov.scot/policies/managing-waste/

[3] https://www.wastedataflow.org/

[4] http://statistics.gov.scot/data/household-waste

[5] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

[6] https://www.wastedataflow.org/documents/guidancenotes/Scotland/zero_waste_plan_recycling_guidance1.pdf

[7] https://www.sepa.org.uk/regulations/waste/energy-from-waste/

[8] https://www.letsrecycle.com/news/latest-news/energy-from-waste-could-become-worse-than-landfill/

[9] https://www.dundeecity.gov.uk/news/article?article_ref=3071

[10] https://pypi.org/project/kmodes/

A Appendix A

A.1 Software versions, environment and included packages

Python version: Python 3.8.3

Environment: Jupyter Notebook

Packages used:

- pandas
- matplotlib.pyplot
- numpy

- sys
- cluster from sklearn
- metrics from sklearn
- model_selection from sklearn
- tree from sklearn
- scale from sklearn.preprocessing
- LabelEncoder from sklearn.preprocessing