# 7CCSMDM1 Data Mining

Coursework 1

**George R.E. Bradley**
K20113376

February 2021

# 1 Classification

## 1.1 Data information table

| Adult Data Information | |
|---|---|
| Number of instances | 48842 |
| Number of missing values | 6465 |
| Fraction of missing values over all attribute values | 0.009454684785342825 |
| Number of instances with missing values | 3620 |
| Fraction of instances with missing values over all instances | 0.07411653904426518 |

## 1.2 Attribute nominal values

All instances of each of the selected *Adult* data attributes were encoded using the *Scikit-learn LabelEncoder*. A set of the returned values for each of the attributes is displayed.

| Nominal Attribute Values | |
|---|---|
| Age | 0, 1, 2, 3, 4 |
| Workclass | 0, 1, 2, 3, 4, 5, 6, 7, 8 |
| Education | 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 |
| Education-num | 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 |
| Marital-status | 0, 1, 2, 3, 4, 5, 6 |
| Occupation | 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 |
| Relationship | 0, 1, 2, 3, 4, 5 |
| Race | 0, 1, 2, 3, 4 |
| Sex | 0, 1 |
| Capitalgain | 0, 1, 2, 3, 4 |
| Capitalloss | 0, 1, 2, 3, 4 |
| Hoursperweek | 0, 1, 2, 3, 4 |
| Native-country | 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41 |

## 1.3 Decision tree classifier creation

A decision tree classifier was created using the *DecisionTreeClassifier* class from *Scikit-learn*. Once instances with missing values had been excluded, the resulting dataset was comprised of 45222 instances, of which 20% (9045) was used for testing. A common split ratio of 80% training data and 20% test data was used in order to prevent overfitting the model. The resulting error rate was 17.62% of the 9045 values tested, 7451 were correctly predicted.

## 1.4 Comparison of missing value handling approaches

The data set $D^{'}$ was constructed from the original data D. $D^{'}$ contains all the instances with at least one missing value from the original data and an equal number of randomly chosen instances without missing values.

Two new data sets where constructed using the $D^{'}$ data set:

- $D_1^{'}$ was constructed by filling all missing values in D' with "missing"
- $D_2^{'}$ was constructed by populating the missing values in D' with the most popular value for each attribute

Once the two data sets were constructed, they were encoded using *Scikit-learn LabelEncoder* and two decision trees were trained using each of the constructed data sets. The trained trees where tested using all remaining data left in the data set D, once instances which also appear in $D_1^{'}$ and $D_2^{'}$ had been removed respectively. The final test data set is circa four times larger than the training set. Which ensures a more accurate calculation of model performance whilst ensuring that the model is not tested on any training data.

| **Data Set** | **Error Rate (%)** |
|---|---|
| $D_1^{'}$ | 21.98% |
| $D_2^{'}$ | 21.67% |

**Findings:** The range of the error rates between the two decision trees was 0.31%. It could be concluded that the decision tree models, due to be trained on small data sets of 7240 instances and tested with 41602 instances are suffering from underfitting, causing the models to not capture and reflect the underlying trend of the data. Underfitting could explain the high error rate and low variance of the results.

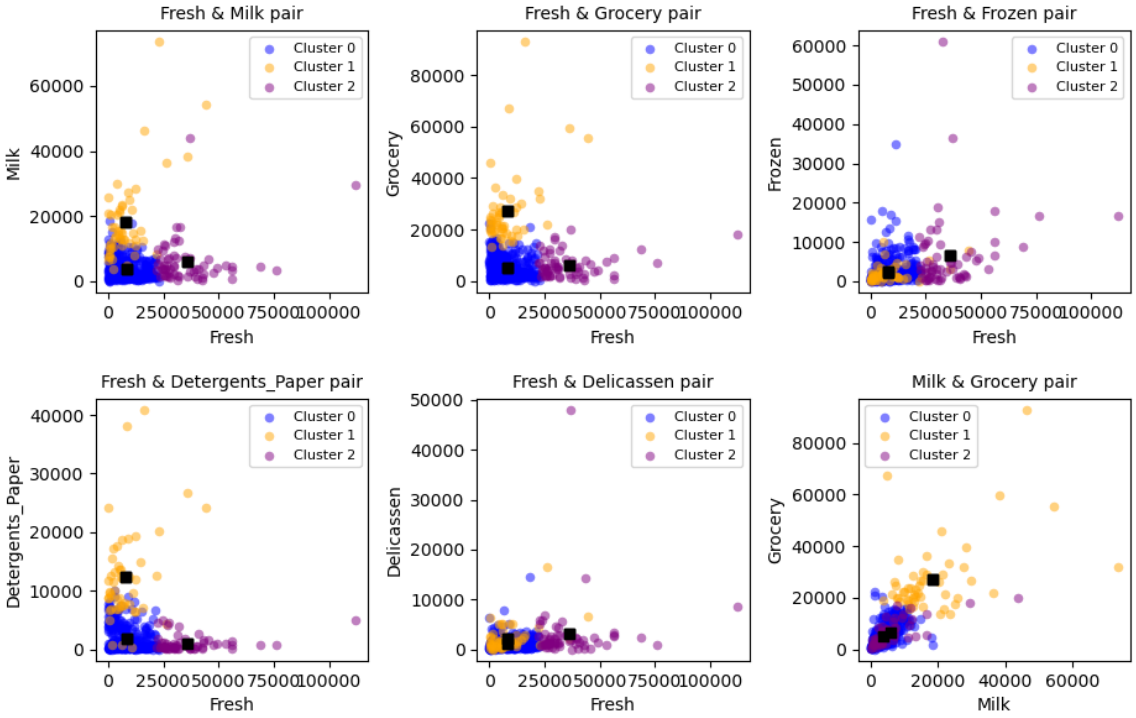
# 2 Clustering

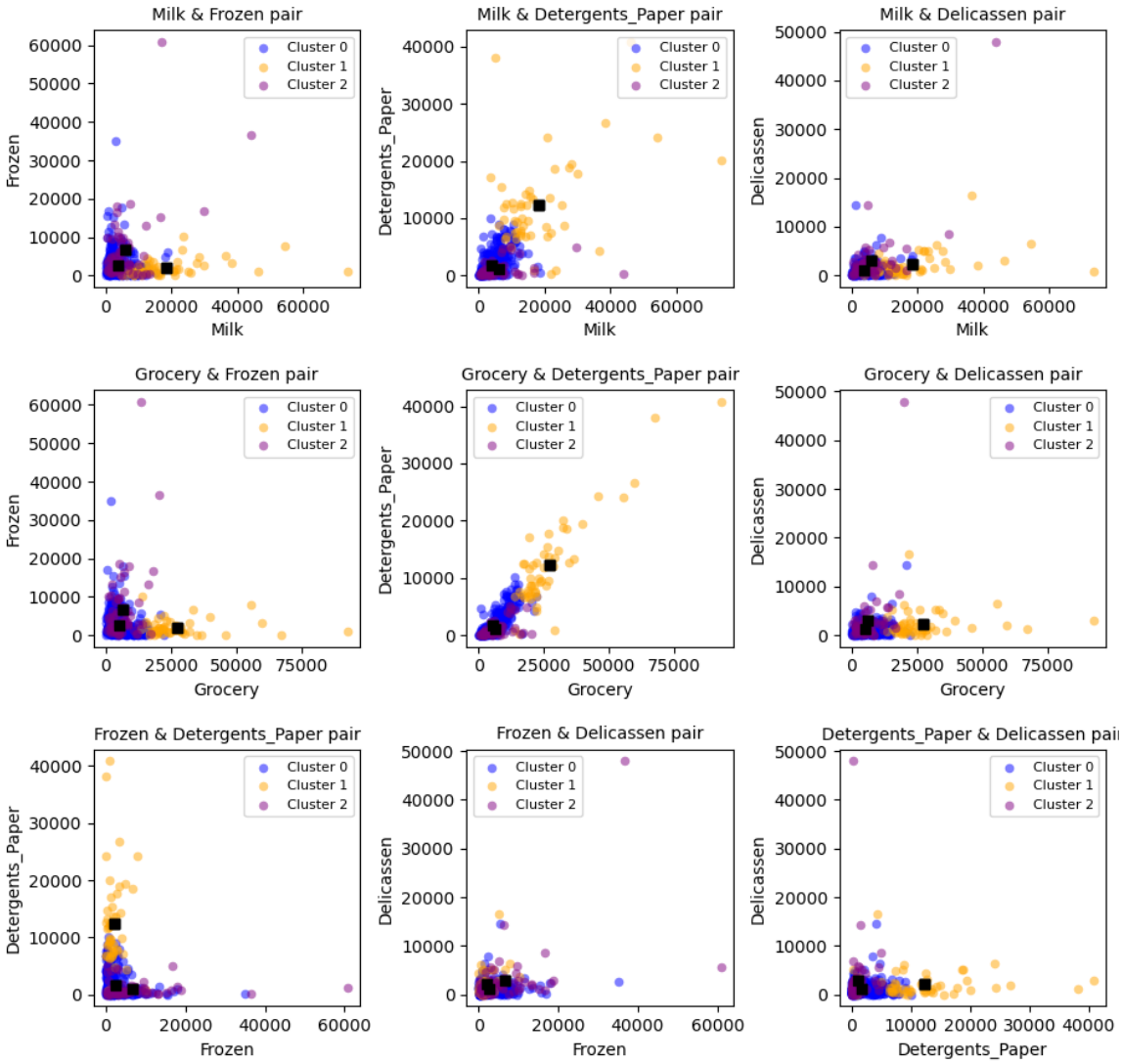
## 2.1 Data attribute's mean and range

Six numeric attributes of the *Wholesale Customer* data were selected (i) Fresh (ii) Milk (iii) Grocery (iv) Frozen (v) Detergents (vi) Delicatessen. The mean and range of each attribute is displayed below. The mean values have been rounded to 2 decimal places for formatting, it was deemed no meaningful insight would be been lost.

| Attribute | Mean (2dp) | Min | Max |
|---|---|---|---|
| Fresh | 12000.3 | 3 | 112151 |
| Milk | 5796.27 | 55 | 73498 |
| Grocery | 7951.28 | 3 | 92780 |
| Frozen | 3071.93 | 25 | 60869 |
| Detergents | 2881.49 | 3 | 40827 |
| Delicatessen | 1524.87 | 3 | 47943 |

## 2.2 Scatterplots






**Observations:** It can be inferred that clustering was more successful with the pairs, (i) Fresh & Milk (ii) Fresh & Grocery (iv) Fresh & Detergents in comparison with the other pairs. Excluding possible outliers, the clusters within these pairs seem to have minimised within cluster distance and maximised between cluster distance. In contrast, the pair Frozen & Delicassen despite having a low within cluster distance appears to have a low between cluster distance, resulting in poor clustering performance. It could be suggested that the pairs which include the Fresh attribute yielded more optimal clustering results, due to the mean value of the Fresh attribute being meaningfully higher than the other attributes, causing a skew of importance toward the Fresh attribute.

### 2.3 Cluster distance values at different $k$ values

| | $k$=3 | $k$=5 | $k$=10 |
|---|---|---|---|
| $WC$ | 80333726672.53984 | 52928148942.57614 | 29650104827.92608 |
| $BC$ | 3132296567.4367013 | 25621025530.77019 | 216486277704.1037 |
| $BC/WC$ | 0.03899105266465127 | 0.4840718227000053 | 7.301366351332598 |

**Observations:** K-means clustering aims to minimize within cluster distance and maximise the between cluster score. From the table we are able to deduce that as the number of clusters increases the within cluster score reduces whilst, increasing the between cluster score, the BC/WC ration increasing from 0.038 with $k$=3 to 7.301 when $k$=10. We could therefore conclude that strictly within this set of $k$ values, the higher the $k$ value the better the clustering performs, due to the small sample size we cannot however infer that a larger $k$ value performs more optimally overall.