A picture containing text, clipart

Description automatically generatedRIGA TECHNICAL UNIVERSITY

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

INSTITUTE OF APPLIED COMPUTER SYSTEMS

Practical assignment #2

Fundamentals of Artificial Intelligence

**Machine Learning**

[Model](https://drive.google.com/drive/folders/1sIZG9V7FVDzrW3097S3OSi9mKr96pi9v?usp=sharing) /[Database](https://vincentarelbundock.github.io/Rdatasets/csv/openintro/fastfood.csv)

Author: Balasuriyage Aritha Dewnith

KUMARASINGHE

Student ID Number: 203AEB014

[1. Exploring and Pre-processing Data 3](#_Toc104414404)

[Description of the dataset. 3](#_Toc104414405)

[Description of content and features. 3](#_Toc104414406)

[Analysis of Dataset 6](#_Toc104414407)

[Imputation 7](#_Toc104414408)

[Scatter Plots 7](#_Toc104414409)

[Histograms 7](#_Toc104414410)

[Distributions 8](#_Toc104414411)

[Heatmap/Corelation 9](#_Toc104414412)

[2. Unsupervised Learning 11](#_Toc104414413)

[Hierarchical Clustering 11](#_Toc104414414)

[Hyperparameters 11](#_Toc104414415)

[Iterations of Hierarchal Clustering 11](#_Toc104414416)

[K-Means Clustering 14](#_Toc104414417)

[Hyperparameters 15](#_Toc104414418)

[Iterations of K-Means Clustering 15](#_Toc104414419)

[Conclusions regarding the Unsupervised Learning Algorithms 16](#_Toc104414420)

[3. Supervised Learning 16](#_Toc104414421)

[Artificial Neural Networks 16](#_Toc104414422)

[Description of Hyperparameters 16](#_Toc104414423)

[Iterations of the neural network 16](#_Toc104414424)

[Support Vector Machine (SVM) Algorithm 18](#_Toc104414425)

[Description of Hyperparameters 19](#_Toc104414426)

[Iterations of the SVM 19](#_Toc104414427)

[k-Nearest Neighbour (kNN) Algorithm 21](#_Toc104414428)

[Description of Algorithm 21](#_Toc104414429)

[Description of Hyperparameter 21](#_Toc104414430)

[Iterations of the kNN 21](#_Toc104414431)

# Exploring and Pre-processing Data

## Description of the dataset.

Name: Nutrition in Fast Food . [link to csv. file](https://vincentarelbundock.github.io/Rdatasets/csv/openintro/fastfood.csv)

Source: R data sets on GitHub [link](https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/datasets.csv)

Author/Licence: This information could not be found.

Problem Domain: This dataset provides data regarding the nutritional values of different menu items from popular fast-food chains.

Data Collection Methods: Specific information about data collection methods could not be found but an education guess would be that information about nutritional content of menu items could be found on the menus at the various restaurant chains.

## Description of content and features.

Count of data objects: 515

Number of classes: 8

Class descriptions:

|  |  |
| --- | --- |
| Restaurant Name/Label | Number of menu items |
| Arby’s | 55 |
| Burger King | 70 |
| Chick Fil-A | 27 |
| Dairy Queen | 42 |
| McDonalds | 57 |
| Sonic | 53 |
| Subway | 96 |
| Taco Bell | 115 |

Features:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature name | Description | Type | Minimum Value | Maximum Value |
| calories | Number of calories in kCal | Numeric | 20 | 2430 |
| cal\_fat | Calories from fat in kCal. | Numeric | 0 | 1270 |
| total\_fat | Total fat in g. | Numeric | 0 | 141 |
| sat\_fat | Saturated fat in g. | Numeric | 0 | 47 |
| trans\_fat | Trans fat in g. | Numeric | 0 | 8 |
| cholesterol | Cholesterol levels in grams. | Numeric | 0 | 805 |
| sodium | Sodium levels in milligrams. | Numeric | 15 | 6080 |
| total\_carb | Total carbohydrates in grams. | Numeric | 0 | 156 |
| fiber | Amount of fiber in grams. | Numeric | 0 | 17 |
| sugar | Amount of sugar in grams. | Numeric | 0 | 87 |
| protien | Amount of protein in grams. | Numeric | 1 | 186 |
| vit\_a | Amount of Vitamin A in milligrams. | Numeric | 0 | 180 |
| vit\_c | Amount of Vitamin C in milligrams. | Numeric | 0 | 400 |
| calcium | Amount of Calcium in milligrams. | Numeric | 0 | 290 |

Snippet of data table:

Table

Description automatically generated

## 

## Analysis of Dataset

### Imputation

Of the 14 features 9% of the data was missing most of which was in the last three columns vat\_A, vat\_C and Calcium. These missing values were replaced with the Average value or the most frequent value from those respective columns.

### Scatter Plots

Chart, scatter chart

Description automatically generated

Figure 1.1

Chart, scatter chart

Description automatically generated

Figure 1.2

The scatter plot shows that for these chosen variables(portien,total\_carb and sat\_fat,calories) there is some separation between the different class objects with five distinguishable colour regions, but there is a significant amount of clustering of the data specially towards the origin. By looking at this we can conclude that it will be highly likely that the classification will be less accurate.

### Histograms

Chart

Description automatically generated with low confidence

A picture containing graphical user interface

Description automatically generated

Figure 1.3

Trans fat appears to have the highest dispersion meaning that it would result in the larger variance between values from different classes.

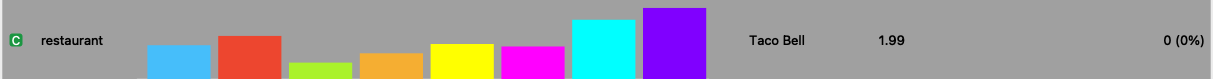


Figure 1.4

We can conclude from this graph that the data isn’t well balanced due to the fact that there is a lot of variances in the number of items belonging to each group. With Taco Bell and Subway having a much larger number of items.

### Distributions

A picture containing histogram

Description automatically generated

Figure 1.5

The calories feature has a right skewed normal distribution, and most of the other features follow a very similar distribution.

Chart, histogram

Description automatically generated

Figure 1.6

The distribution for trans\_fat on the other hand follows a distribution closer to a binomial one.

### Heatmap/Corelation

Chart, treemap chart

Description automatically generated

Figure 1.7

This heatmap was created using the python seaborn library, and with it you can see that most of the features seem to have at least a moderate corelation with one another.

Graphical user interface

Description automatically generated

Figure 1.8

Therefor a closer analysis reveals that the features that have a strong correlation(>=0.7) have to removed. Therefor cal\_fat, total\_fat, sodium, tran\_fat, cholesterol have to be removed.

The final features used:

Graphical user interface, text, email

Description automatically generated

Figure 1.9

# Unsupervised Learning

## Hierarchical Clustering

### Hyperparameters

Distance Metric: This algorithm is used to calculate the distances between data objects.

Linkage metric: How the distance between elements of two clusters are decided.

Pruning: Does not affect the clustering but visually changes how many leaves of the dendrogram are shown

Selection: Changes the highlighting of the levels of the dendrogram.

## Iterations of Hierarchal Clustering

1. Distance Metric: Euclidean

Linkage: Average

Chart, box and whisker chart

Description automatically generated

Figure 2.1

1. Distance Metric: Euclidean

Linkage: Single

Chart

Description automatically generated

Figure 2.2

When using single as the linkage instead to average we can see that the C8 in the dendrogram became a single dairy queen stead leaf of multiple subway items of the subway cluster were close to each other but did not have a significant enough familiarity when it comes to features. The cluster C7 is also interesting because it contains leaves items from the same class but in different clusters meaning that they are probably close to another but not still have certain features that are different.

1. Distance Metric: Manhattan

Linkage: Average

Chart

Description automatically generated

Figure 2.3

1. Distance Metric: Manhattan

Linkage: Weighted

Chart, box and whisker chart

Description automatically generated

Figure 2.4

When comparing the Manhattan and the Euclidean clusters using average linkage you can see that number of single leaf clusters have decreased and meaning that there is a greater variance when every single feature is considered when calculating distance.

The weighted linkage seems to make it a bit easier to cluster subway items together demonstrated by C10.

1. Distance Metric: ﻿Mahalanobis

Linkage: Average

Chart

Description automatically generated

Figure 2.5

1. Distance Metric: Mahalanobis

Linkage: Average

Chart

Description automatically generated

Figure 2.6

The Mahalanobis there are a few more clusters and also a higher number of single class clusters this can probably be explained by the fact that Mahalanobis take correlation into account which works in a dataset with high amounts of correlating features.

## K-Means Clustering

### Hyperparameters

Number of Clusters: Number of clusters to be formed when sorting the data.

Initialization: How the initiation of the clustering begins either with centre being randomly selected or clusters being assigned randomly to the data objects.

Reruns/ Maximum Iterations: How many times that algorithm must do start over with a new random starting point or random clustering and the number of iterations that should be included in each run.

### Iterations of K-Means Clustering

Random Initialization with 10 reruns and max iterations set to 50

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| K\_value | 3 | 5 | 8 | 10 | 20 |
| Silhouette plot | A picture containing chart  Description automatically generated | | | | |
| Silhouette score | 0.525 | 0.461 | 0.419 | 0.362 | 0.308 |

The reason that there is a single plot to represent all five of the values is that changing the k\_Mean hyperparameters had no effect on the Silhouette plot even though the silhouette scores change. Arby’s, Taco Bell, Chick Fil-A, Burger King and Sonic are the only classes with inliers, and therefore we can conclude that the k\_means algorithm finds it easier to classify these three classes when compared with other classes. Although those classes too have only weak classifications due to the silhouette score being under 0.2.

### Conclusions regarding the Unsupervised Learning Algorithms

It’s obvious that the separability of the classes is very poor. All the classes that show some separability in the hierarchical clustering(McDonalds and Subway) appears to have a low silhouette score, and Taco bell which appears to have apparent highest silhouette scores doesn’t have a single leaf unique leaf in any of dendrograms thereby the results of the two algorithms appear to contradict each other.

# Supervised Learning

The dataset will be divided into the training set using random sampling in order to ensure that the test dataset isn’t skewed towards a signal class. Where 50% of the data is used as test data.

Graphical user interface, text

Description automatically generated

## Artificial Neural Networks

### Description of Hyperparameters

Neurons in the hidden layer: this metric is self-explanatory, for each of the hidden layers there needs to be a specific number of neurons.

Activation: The function that decides if a neuron should be activated or not.

Solver: Optimizes the weight that each input value carries.(weight optimization)

Alpha: The learning rate.

Maximum number of Iterations: The number of times that the optimizations occur.

### Iterations of the neural network

1. Parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Neurons in hidden layer | Activation | Solver | Alpha | Max Iterations |
| 100 | Logistic | Adam | 1 | 500 |

Average over all classes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AUC | CA | F1 | Precision | Recall |
| 0.860 | 0.549 | 0.482 | 0.492 | 0.549 |

Table

Description automatically generated

Figure 3.1

The neural network follows a similar pattern to the supervised learning algorithm where very sew of the data items are categorized correctly with chick fill-A having a very high turn positive percentage. But none of the data items have been sorted as into dairy queen.

1. Parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Neurons in hidden layer | Activation | Solver | Alpha | Max Iterations |
| 100 | Logistic | L-BFGS-B | ﻿1 | 500 |

Average over all classes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AUC | CA | F1 | Precision | Recall |
| 0.903 | 0.660 | 0.631 | 0.633 | 0.660 |

Table

Description automatically generated

Figure 3.2

Changing the solver has increased the values of all the evaluation metrics as well as the True positive values showing that the L-BFGS-B is better than the Adam algorithm.

1. Parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Neurons in hidden layer | Activation | Solver | Alpha | Max Iterations |
| 100 | tanh | L-BFGS-B | ﻿1 | 500 |

Average over all classes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AUC | CA | F1 | Precision | Recall |
| 0.925 | 0.693 | 0.684 | 0.682 | 0.693 |

Table

Description automatically generated

Figure 3.3

Changing the activation function has further improved the performance of the algorithm although the precision and recall hasn’t improved significantly meaning that classification itself is better it is still a hit and miss.

1. Parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Neurons in hidden layer | Activation | Solver | Alpha | Max Iterations |
| 100 | tanh | L-BFGS-B | ﻿0.1 | 500 |

Average over all classes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AUC | CA | F1 | Precision | Recall |
| 0.925 | 0.714 | 0.710 | 0.708 | 0.714 |

Table

Description automatically generated

Figure 3.4

Decreasing the alpha makes further improves the algorithm.

Note: The number of neurons was not demonstrated due to there not being enough of a change in the evaluation metrics even with say a thousand neurons or adding additional layers.

## Support Vector Machine (SVM) Algorithm

In essence the SVM algorithm finds the hyperplane that maximises the distance between data points of different classes. The hyper plane is placed in such way to ensure that the data points on either side of the hyperplanes called support vectors are used to give local to measure the distance of different classes. To maximise the distance between the data points between the data points on either side of the hyperplanes is the hinge loss function as known as the cost function. The higher the deviation from zero on the cost function value the more inaccurate the prediction.

### Description of Hyperparameters

Cost: the evaluation function that is used to define the how the distance between the adjacent items.

Kernel: define the shape in which the hyperplane.

Regression loss: the acceptable distance between the two data points where a penalty is not given.

Numerical tolerance: this is the permitted deviation between the actual and predicted value.

Iterations: number of iterations that are done where higher values normalize more.

### Iterations of the SVM

1. Parameter

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cost | Kernel | Numerical  Tolerance | Regression  Loss | Iterations |
| 50 | Linear | 0.002 | 0.1 | ﻿10000 |

Average over all classes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AUC | CA | F1 | Precision | Recall |
| 0.905 | 0.669 | 0.672 | 0.676 | 0.669 |

Table

Description automatically generated

Like the neural network this algorithm also is able to classify the subway and taco bell items are classified correctly but the rest of the classes appear to be mediocre at best.

1. Parameter

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cost | Kernel | Numerical  Tolerance | Regression  Loss | Iterations |
| 100 | Linear | 0.002 | 0.1 | ﻿10000 |

Average over all classes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AUC | CA | F1 | Precision | Recall |
| 0.899 | 0.640 | 0.645 | 0.653 | 0.640 |

Table

Description automatically generated

Increasing the cost has improved the evaluation metrics but the true positive percentages have

degraded.

1. Parameter

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cost | Kernel | Numerical  Tolerance | Regression  Loss | Iterations |
| 100 | RBF | 0.002 | 0.1 | ﻿10000 |

Average over all classes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AUC | CA | F1 | Precision | Recall |
| 0.926 | 0.702 | 0.702 | 0.703 | 0.702 |

Table

Description automatically generated

Changing the function of the hyperplane appears to have improved the performance metrics while maintaining the true positive rate. Meaning that the distribution of various features more closely fit a RBF distribution.

## k-Nearest Neighbour (kNN) Algorithm

### Description of Algorithm

This algorithm calculates the distance between a k number of neighbours using a certain function then groups them based how values from the distance function.

### Description of Hyperparameter

Number of neighbours: Number of nearest neighbours when clusters are forming.

Metric: The function used to calculate the distance between neighbouring items.

Weight: I which the neighbours.

### Iterations of the kNN

1. Parameter

|  |  |  |
| --- | --- | --- |
| Number of  Neighbours | Metric | Weight |
| ﻿8 | ﻿Euclidean | Uniform |

Average over all classes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AUC | CA | F1 | Precision | Recall |
| 0.716 | 0.330 | 0.306 | 0.327 | 0.330 |

Table

Description automatically generated

1. Parameter

|  |  |  |
| --- | --- | --- |
| Number of  Neighbours | Metric | Weight |
| ﻿8 | ﻿ ﻿Manhattan | Uniform |

Average over all classes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AUC | CA | F1 | Precision | Recall |
| 0.788 | 0.428 | 0.410 | 0.437 | 0.428 |

Table

Description automatically generated

1. Parameter

|  |  |  |
| --- | --- | --- |
| Number of  Neighbours | Metric | Weight |
| ﻿8 | ﻿ ﻿Mahalanobis | Distance |

Average over all classes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AUC | CA | F1 | Precision | Recall |
| 0.860 | 0.562 | 0.500 | 0.551 | 0.562 |

Table

Description automatically generated