

# ARTIFICIAL INTELLIGENCE FOR IMPROVING CANCER SUPPORT

## Unsupervised machine learning of integrated health and social care data from the Macmillan Improving the Cancer Journey service in Glasgow

Kean Lee Kang<sup>1#</sup>, Margaret Greer<sup>2#</sup>, James Bown<sup>1</sup>, Janice Preston<sup>2</sup>, Judith Mabelis<sup>2</sup>, Leigh-Anne Hepburn<sup>3</sup>, Miriam Fisher<sup>3</sup>, Ruth Falconer<sup>1</sup>, Sandra McDermott<sup>4</sup>, Stuart Deed<sup>3</sup>

<sup>1</sup>Abertay University, <sup>2</sup>Macmillan Cancer Support, <sup>3</sup>Digital Health and Care Institute, <sup>4</sup>Glasgow City Council, #Contributed equally

### Background



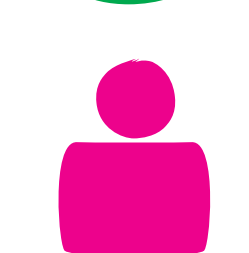
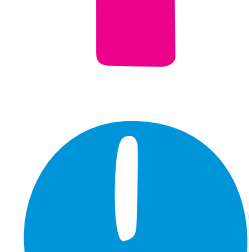
Macmillan Improving the Cancer Journey (ICJ) was launched in 2014 by Glasgow City Council with funding from Macmillan Cancer Support. Shortly after diagnosis, people with cancer are invited to take part in a person centred conversation, where they will complete a Holistic Needs Assessment (HNA) with a key worker to establish any physical, emotional, social, financial, family, spiritual or practical needs. The key worker then either signposts or refers on to relevant agencies to support the person and their individual needs.

Routinely, service data is collected on ICJ clients including demographic and health information, results from the HNA assessment and quality of life score as measured by EQ-5D health status. There is also data on the number and type of referrals made and feedback from service users on the overall service. By applying artificial intelligence (AI) and interactive visualization technologies to these data, we seek to improve service provision and optimise resource allocation.

### Methods

An unsupervised machine-learning algorithm was deployed to cluster the data. The k-modes method<sup>1</sup> used for categorical data is an extension of the classical k-means<sup>2</sup>. Initialization maximized both data point density and inter-cluster dissimilarity<sup>3</sup>, and the algorithm automatically imputed missing data<sup>4</sup> and identified the number of clusters<sup>5</sup>.

The resulting clusters are used to summarize complex data sets and produce three-dimensional visualizations of the data landscape. In addition, the traits of new ICJ clients are predicted by approximately matching their details to the nearest existing cluster centre. Finally, based on client data, a deep neural network (multilayer perceptron) is trained to recommend agencies that the client should be referred to out of a list of 125 agencies.

-  139 input fields
-  125 referral choices
-  7 milliseconds/person
-  1/2 million answers/hour

### Conclusions

A key strength of this system is its ability to rapidly ingest new data on its own, and derive new predictions from the data. This means the model can guide service provision by forecasting demand based on actual or hypothesised data. The neural network in particular, has the potential for

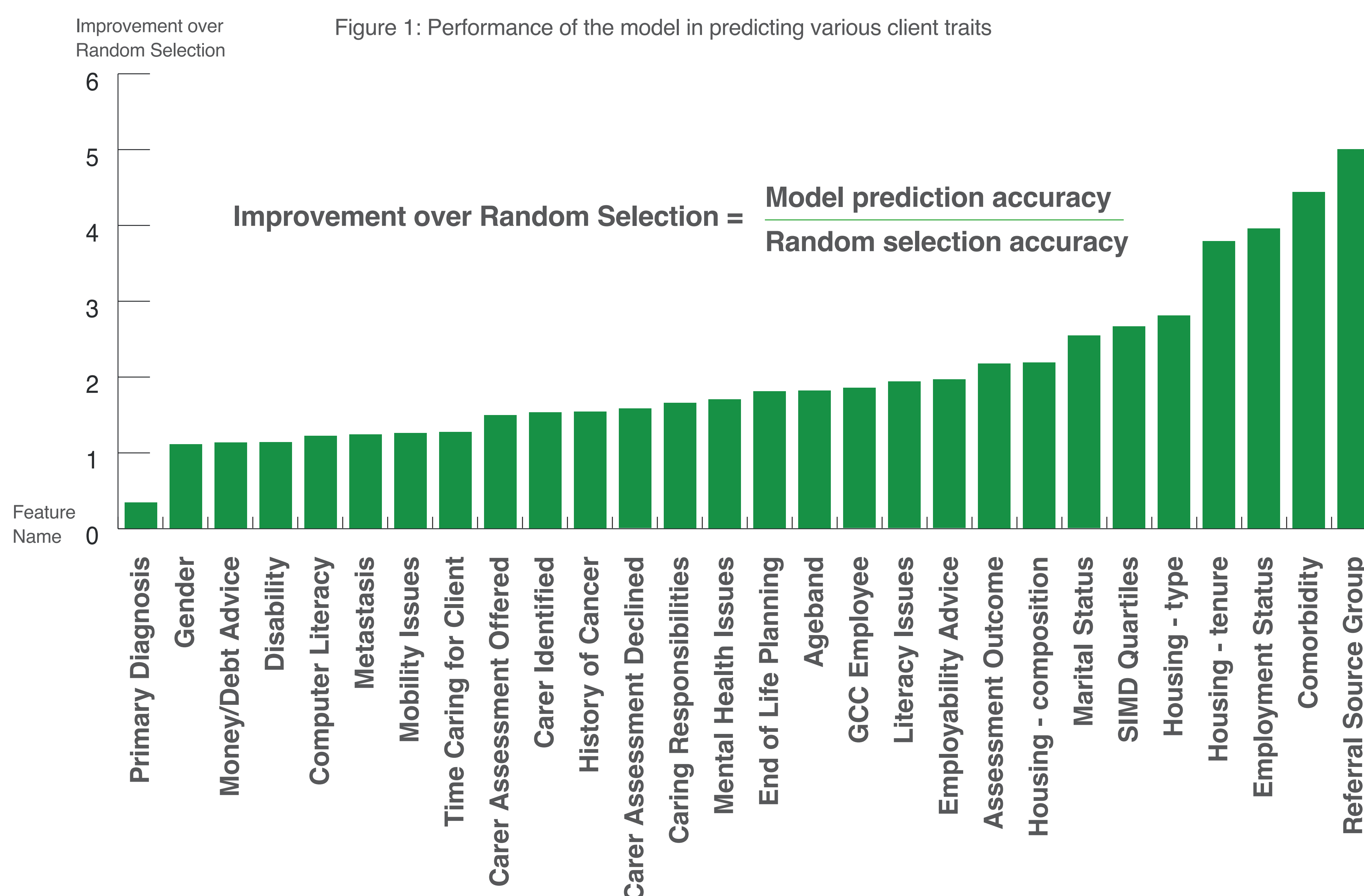


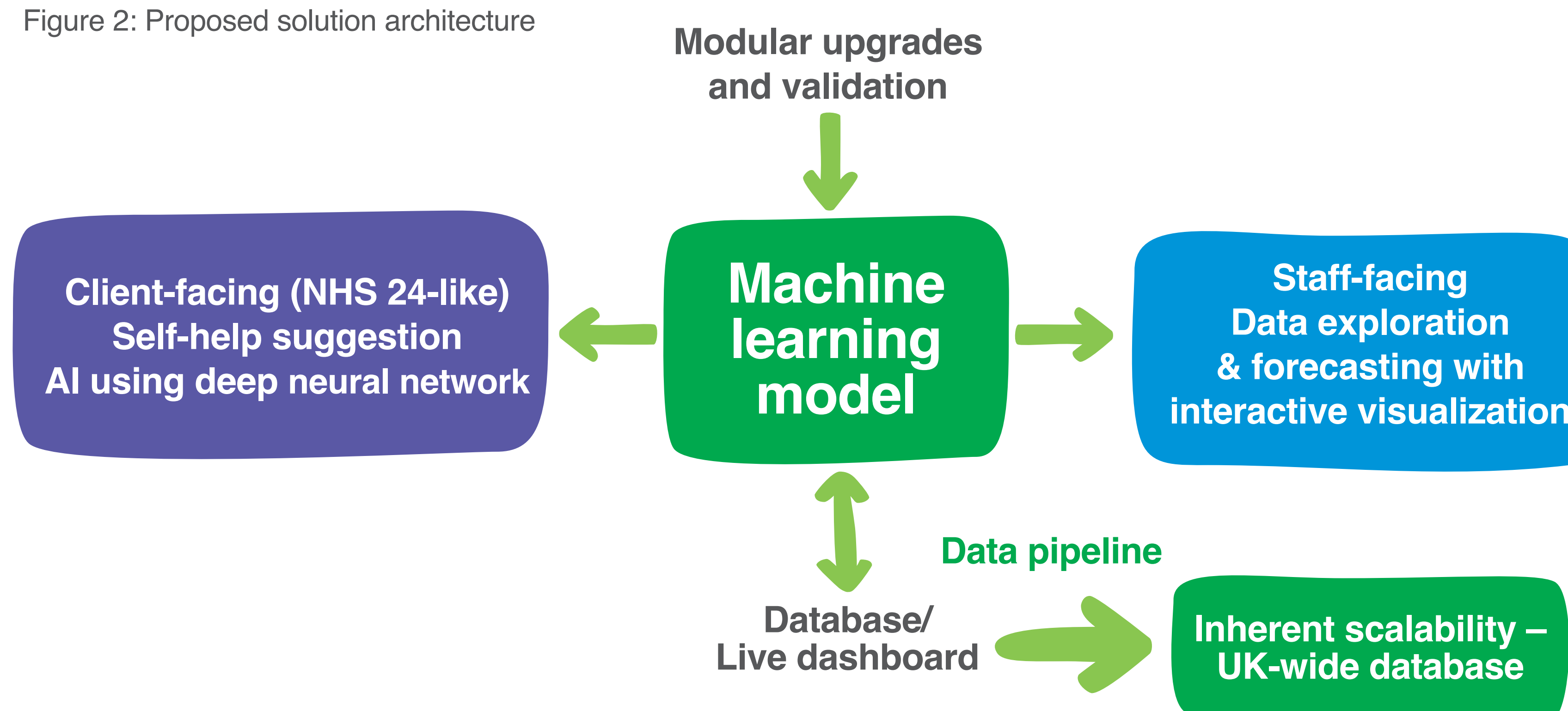
Figure 1: Performance of the model in predicting various client traits

### Results

K-fold cross-validation showed the model's effectiveness over a wide range of ICJ service client traits. For example, the model can predict marital status, employment status and housing type with an accuracy up to 5 times greater than random selection (Figure 1).

In terms of predicting the agencies that ICJ clients are referred on to, in 70.1% of cases, the neural network refers to at least one of the agencies that the key worker refers to. The trained neural network makes these referral decisions in an average of 7 milliseconds per person and this could potentially be used either to assist key workers with their work or as a platform for automated, intelligent self-service referral (Figure 2).

Figure 2: Proposed solution architecture



refinement, which should improve its accuracy. The aim is to provide intelligent person-centred recommendations. The machine-learning model described here is part of a prototype software tool currently under development for use by Macmillan Cancer Support and partners. Looking

forward, we plan to engineer a data pipeline linking the software to anonymised data collected by Macmillan services and feeding in to a live dashboard for staff to obtain a real-time view of the service status.



This work uses data provided by patients and collected by Glasgow City Council as part of their care and support.

#### References

- (1) Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery* 1998;2(3):283-304.
- (2) Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Second ed. New York: Springer-Verlag; 2009.
- (3) Cao F, Liang J, Bai L. A new initialization method for categorical data clustering. *Expert Systems with Applications* 2009;36(7):10223-10228.
- (4) Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2001;63(2):411-423.
- (5) Chi JT, Chi EC, Baraniuk RG. k-POD: A Method for k-Means Clustering of Missing Data. *The American Statistician* 2016;70(1):91-99.