[[1]](#footnote-1)

Predicting Hospital Readmission of Diabetes Patients

Minh Quan Do

**Introduction:**

According to the Health Care Cost Institute (HCCI), total spending on healthcare per person is now growing at faster rates than prior years with a 4.6% growth in 2016 compared to a 4.1% growth in 2015 which is also an increase from just over 3% growth from 2012-2014 [1]. Furthermore, according to the National Diabetes Statistics Report, 30.3 million people in the United States (9.4% of the U.S. population) has diabetes [2]. In 2014, 7.2 million patients discharged from hospitals were diagnosed with diabetes [2]. Also, diabetes is the seventh leading cause of death in the United States in 2015 [2]. Helping hospitals know which patients are most at-risk for readmission into the hospital can help hospitals better prepare and save those more patients’ lives.

Because of data repositories such as the UCI Machine Learning Repository, statistical models have been used to analyze the large amounts of data in these repositories [3]. These techniques involve splitting the dataset up into three smaller datasets based on the patients’ age (0-29 years old, 30-69 years old, 70-100 years old) and then training different machine learning models on each of those smaller datasets [4]. The machine learning models used are a combination of decision trees, random forests, and support vector machines [4]. This technique is fairly complicated and does not take into account the heterogeneity of the patient population (at least not enough because the data is only categorized into three categories and the categories are only based on the patients’ age); therefore, the proposed method involves using clustering to organize the dataset into more homogenous groups and then using the data in those groups to train separate artificial neural networks [5]. The hypothesis is that breaking the diverse heterogenous group of patients into more homogenous groups will result in more accurate predictions and a lower error because the patients in these homogenous groups would have more similar features and theoretically, the patients will have more similar lifestyle patterns which should lead to a lower percent error (Formula I) [5].

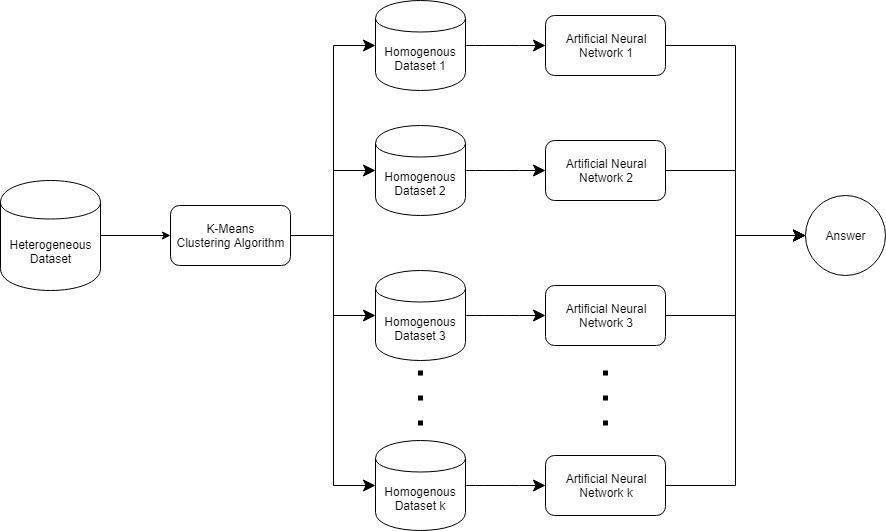
Formula I:

**Methods:**

To do this, the *Diabetes 130-US hospitals for years 1999-2008 Data Set* in the UCI Machine Learning Repository to build a dataset that would include race (nominal), gender (nominal), age (numeric), weight (numeric), the number of emergency visits in the year preceding the encounter (numeric), the number of inpatient visits in the year preceding the encounter (numeric), glucose serum test result (nominal), A1C test result (nominal), and if diabetic medication was prescribed (nominal). Because the dataset has some missing values, data entries with missing values will be discarded to clean the data. Once the data is cleaned, the dataset will be split up into three portions: one portion will be used as a training set to train the machine learning model, one portion will be used as a validation set to validate the accuracy of the model, and the last portion of the dataset will be used to as a test set. The accuracy of the model will be determined using the equation described in Formula I.

From there, as previously discussed, a clustering algorithm (K-Means) will be used to organize the dataset into homogenous groups. Once the homogenous groups have been determined, artificial neural networks will be used to analyze each homogenous group separately. The artificial neural networks would take in race, gender, age, weight, the number of emergency visits in the year preceding the encounter, the number of inpatient visits in the year preceding the encounter, glucose serum test result, A1C test result, and if diabetic medication was prescribed as its inputs (9 inputs) in its input layer; and it will have one output in its output layer, which is if the patient will be readmitted into the hospital within 6 months, after 6 months, or not readmitted at all.

Diagram I:



The diagram above is a flow chart of the system. The heterogenous dataset is sorted into k number of smaller homogenous datasets; from there, for each homogenous dataset, a separate artificial neural network will be trained with the data in that homogenized dataset. Therefore, each artificial neural network will output a different answer. When classifying data from the test set, a point will first be categorized into a particular homogenous group, and then the class label will be determined using the homogenous group’s respective artificial neural network.

References

1. United States, Congress, “2016 Health Care Costs and Utilization Report.” *2016 Health Care Costs and Utilization Report*, 2016.
2. United States, Congress, Division of Diabetes Translation. “National Diabetes Statistics Report.” *National Diabetes Statistics Report*, 2017.
3. Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” *BioMed Research International*, vol. 2014, Article ID 781670, 11 pages, 2014.
4. Mingle, Damian. “Predicting Diabetic Readmission Rates: Moving Beyond HbA1c”. *Current Trends in Biomedical Engineering & Biosciences*, 2017, 7. 10.19080/CTBEB.2017.07.555715.
5. Miriam Seoane Santos, Pedro Henriques Abreu, Pedro J Garcia-Laencina, Adelia Simao, Armando Carvalho, “A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients”, *Journal of biomedical informatics*, 58, 49-59, 2015.

1. [↑](#footnote-ref-1)