**CSCI 317 Project 3 – Cluster Analysis**

1. **Dataset meets the requirement and is appropriate for cluster analysis. If your data set has redundant columns (e.g. “Temperature in Celcius” column is redundant to “Temperature in Fahrenheit” column) or irrelevant columns, you should remove them either in or outside of R or Weka . Describe your data. (20’)**

* The dataset contains 89,237 entries and 15 attributes.
* **age:** Continuous variable representing the age of the individuals.
* **hypertension:** Binary variable indicating the presence of hypertension (0 = no, 1 = yes).
* **heart\_disease:** Binary variable indicating the presence of heart disease (0 = no, 1 = yes)
* **bmi:** Continuous variable representing body mass index
* **hba1c\_level:** Continuous variable representing the HbA1c level, which is a measure of average blood glucose levels over the past three months.
* **blood\_glucose\_level:** Continuous variable representing the current blood glucose level.
* **diabetes:** Binary variable indicating whether the individual has diabetes (0 = no, 1 = yes).
* **gender\_female, gender\_male, gender\_other:** One-hot encoded variables for gender.
* **smoking\_history\_noinfo, smoking\_history\_current, smoking\_history\_former, smoking\_history\_never, smoking\_history\_notcurrent:** One-hot encoded variables representing different categories of smoking history.
* There are no missing values in the dataset.
* No redundant columns.
* Each variable in the dataset is relevant to predicting diabetes.

1. **Use comments in the script file(s) to specify your name and explain the problem (application domain, purpose etc) (20’)**

* R script provided in zip file (Only code) ☺
* Explanation of the results from the script are explained in part 4.

1. **Perform a Cluster Analysis. (30’)**

Preprocessing Steps

* We will need to normalize the numerical values (age, bmi, hb1ac\_level, blood\_glucose\_level)
* Filter -> unsupervised -> attribute -> Normalize -> Apply
* Attributes that are binary in nature do not need to be normalized because they are already on a scale that aligns well with other binary attributes

A screenshot of a computer

Description automatically generated

Performing Cluster Analysis

* Cluster -> SimpleKMeans -> K = 2A screenshot of a computer

  Description automatically generated

1. **Evaluate and explain the results. (20’)**

Weka

Result

*=== Run information ===*

*Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10*

*Relation: cleaned\_diabetes\_one\_hot\_encoding\_BMI\_Outliers-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0*

*Instances: 89237*

*Attributes: 15*

*age*

*hypertension*

*heart\_disease*

*bmi*

*hba1c\_level*

*blood\_glucose\_level*

*diabetes*

*gender\_female*

*gender\_male*

*gender\_other*

*smoking\_history\_noinfo*

*smoking\_history\_current*

*smoking\_history\_former*

*smoking\_history\_never*

*smoking\_history\_notcurrent*

*Test mode: evaluate on training data*

*=== Clustering model (full training set) ===*

*kMeans*

*======*

*Number of iterations: 3*

*Within cluster sum of squared errors: 113085.34875579618*

*Initial starting points (random):*

*Cluster 0: 0.974975,0,0,0.899958,0.545455,0.209091,0,0,1,0,0,0,0,1,0*

*Cluster 1: 0.937437,1,0,0.779319,0.490909,0.727273,1,1,0,0,1,0,0,0,0*

*Missing values globally replaced with mean/mode*

*Final cluster centroids:*

*Cluster#*

*Attribute Full Data 0 1*

*(89237.0) (58348.0) (30889.0)*

*==========================================================*

*age 0.5203 0.5258 0.51*

*hypertension 0.0691 0.0766 0.0549*

*heart\_disease 0.0375 0.0435 0.0262*

*bmi 0.4896 0.4975 0.4747*

*hba1c\_level 0.3658 0.3686 0.3605*

*blood\_glucose\_level 0.2611 0.2639 0.2558*

*diabetes 0.0743 0.0827 0.0584*

*gender\_female 0.5801 0.3579 0.9998*

*gender\_male 0.4198 0.642 0*

*gender\_other 0.0002 0.0002 0.0002*

*smoking\_history\_noinfo 0.3779 0.2626 0.5958*

*smoking\_history\_current 0.0966 0.0683 0.1502*

*smoking\_history\_former 0.0953 0.0728 0.1378*

*smoking\_history\_never 0.363 0.5552 0*

*smoking\_history\_notcurrent 0.0671 0.0411 0.1162*

*Time taken to build model (full training data) : 0.28 seconds*

*=== Model and evaluation on training set ===*

*Clustered Instances*

*0 58348 ( 65%)*

1. *30889 ( 35%)*

Evaluation and Explanation

* 2 clusters were formed from the dataset.
* Cluster 0 contains 65% of the instances (58,348 instances).
* Cluster 1 contains 35% of the instances (30,889 instances).
* The algorithm converged after 3 iterations, which indicates a quick stabilization of cluster centers.
* The within-cluster sum of squared errors (SSE) is 113,085.35. This metric reflects the total variance within the clusters, where a lower number typically indicates a model with clusters that are more compact and well-separated.

Final Cluster Centroids

* Represents the mean of all points in that cluster
* Age: Both clusters have similar age distributions, slightly higher in Cluster 0
* Both hypertension and heart disease are more prevalent in Cluster 0 than in Cluster 1.
* BMI, HbA1c Level, and Blood Glucose Level: Slightly higher on average in Cluster 0, suggesting that this cluster might represent individuals with slightly poorer health metrics.
* The diabetes class has higher prevalence in Cluster 0 (8.27%) compared to Cluster 1 (5.84%), reinforcing the suggestion that Cluster 0 might consist of individuals with more health issues.
* Cluster 0 is mostly male (64.2% male, 35.79% female).
* Cluster 1 is mostly exclusively female (99.98% females)
* Cluster 0 has more individuals who have never smoked (55.52%) and fewer with no information about smoking (26.26%), whereas Cluster 1 has a higher proportion of individuals with no information on smoking (59.58%).
* In conclusion, cluster 0 appears to represent a demographic with more significant health issues (higher average BMI, HbA1c, blood glucose levels, and a higher prevalence of hypertension, heart disease, and diabetes). This cluster is predominantly male, which might suggest certain lifestyle or genetic predispositions. Lastly, cluster 1 seems to consist primarily of females, with lower average values for health issue indicators and a lower prevalence of conditions like diabetes, hypertension, and heart disease. The high rate of missing smoking data might indicate a lack of complete health profiling in this group, which could be an area for further investigation or data collection.

R Script

Result:

*> print(kmeans\_result$centers) # Print cluster centroids*

*age bmi hba1c\_level blood\_glucose\_level hypertension heart\_disease*

*1 0.5088425 0.4946483 0.3703892 0.2652131 0.07379040 0.05446879*

*2 0.5286686 0.4859054 0.3625019 0.2581470 0.06568018 0.02519028*

*diabetes gender\_female gender\_male gender\_other smoking\_history\_noinfo*

*1 0.08761442 0.000000 0.9996531 3.46935e-04 0.4090897*

*2 0.06469497 0.999942 0.0000000 5.79531e-05 0.3553877*

*smoking\_history\_current smoking\_history\_former smoking\_history\_never*

*1 0.10634891 0.11339436 0.3071709*

*2 0.08961481 0.08223544 0.4034115*

*smoking\_history\_notcurrent*

*1 0.06399616*

*2 0.06935054*

*> table(kmeans\_result$cluster) # Frequency table of clusters*

*1 2*

*37471 51766*

Evaluation and Explanation:

Cluster 1

* Mean Age: Normalized to 0.51 (slightly younger demographic)
* BMI: Mean normalized value of 0.494
* HbA1c Level: Normalized mean of 0.370
* Blood Glucose Level: Normalized mean of 0.265
* Hypertension Prevalence: 7.38%
* Heart Disease Prevalence: 5.45%
* Diabetes Prevalence: 8.76%
* Predominantly Male: 99.97%
* High proportion of unknown smoking history: 40.91%
* Number of Instances: 37,471

Cluster 2

* Mean Age: Normalized to 0.53 (older demographic)
* BMI: Mean normalized value of 0.486
* HbA1c Level: Normalized mean of 0.363
* Blood Glucose Level: Normalized mean of 0.258
* Hypertension Prevalence: 6.57%
* Heart Disease Prevalence: 2.52%
* Diabetes Prevalence: 6.47%
* Predominantly Female: Nearly 100%
* High proportion of never smokers: 40.34%
* Number of Instances: 51,766
* Cluster 1 has 37,471 instances, while cluster 2 has 51,766 instances, indicating a larger cluster primarily composed of older females. Cluster 1 shows a slightly higher prevalence of diabetes, hypertension, and heart disease, which aligns with its slightly younger, male-dominated demographic. Cluster 2, with its predominantly female demographic, shows a lower prevalence of the previously mentioned conditions and a higher proportion of never smokers, suggesting potentially different health profiles and risk factors.

1. **Submit a zip file that contains your data file, script file and a WORD document with descriptions and screen shots for each of the above 4 items . File name format is lastname\_project3.zip. (10’)**