RESEARCH PAPER

Diabetes Risk Categorization Using Clustering Analysis

Arumugam P[1], Sornalatha M E[2*]

[1]Professor and Head, Department of Statistics, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli - 627 012, Tamil Nadu, India
Email: sixfacemsu@msuniv.ac.in
[2*]Research Scholar, Department of Statistics, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli - 627 012, Tamil Nadu, India
Email: swarnalathapramod@gmail.com

Received: 25th Aug, 2026 | Revised: 1st Sep, 2026 | Accepted: 5th Sep, 2026 | Available Online: 7th Sep, 2026

ABSTRACT

Objectives. To identify clinically interpretable diabetes risk phenotypes using unsupervised clustering of demographic, lifestyle, and cardiometabolic characteristics in adults.

Methods. We analysed adults with complete diabetes labels and routinely collected variables, including age, sex, BMI, family history, blood pressure, sleep, stress, diet, smoking, alcohol use, physical activity, urination frequency, and pregnancy history. Regular Medicine was excluded to avoid treatment-related label leakage. Missing values were median-imputed, and continuous features were standardised using z-scores. K-Means clustering was performed for 2–6 clusters, with the silhouette coefficient used to evaluate clustering quality. The selected solution was compared with agglomerative hierarchical clustering using Ward linkage (Murtagh & Legendre, 2014; Ward, 1963). Cluster sizes, diabetes prevalence, and feature profiles were examined, and agreement between the two methods was assessed.

Results. A six-cluster solution provided a good balance of clustering performance and clinical interpretability. The clusters showed distinct diabetes prevalence and combinations of age, BMI, blood pressure, sleep, stress, and lifestyle characteristics. Major risk phenotypes were broadly consistent across both clustering methods.

Conclusion. Unsupervised clustering identified clinically meaningful diabetes risk phenotypes and demonstrated robustness across clustering algorithms. These phenotypes may support targeted diabetes prevention and management.

Keywords:

How to cite this article: Arumugam P1 Sornalatha M E2*. Diabetes Risk Categorization Using Clustering Analysis. Int J Drug Deliv Technol. 2026;16(80s):1682-1687. DOI: 10.25258/ijddt.16.80s.208.

1. Introduction

Diabetes mellitus is one of the most prevalent chronic diseases worldwide, with the International Diabetes Federation estimating that 537 million adults aged 20–79 were living with diabetes in 2021 — approximately 10.5% of the global adult population — and projecting continued growth to 783 million by 2045 (Kumar et al., 2023; Tobias et al., 2023). At the same time, prediabetes is widespread: globally, 9.1% of adults are estimated to have impaired glucose tolerance and 5.8% impaired fasting glucose, with both burdens projected to rise (Rooney et al., 2023). The prevention and management of diabetes therefore constitute a central public-health priority (Tobias et al., 2023).

Classical clinical classification of diabetes into broad categories (type 1, type 2, gestational, and so on) obscures considerable heterogeneity within each category — particularly within type 2 diabetes (Ahlqvist et al., 2018, 2020). This heterogeneity has motivated a shift toward data-driven subtyping. In a landmark 2018 study, Ahlqvist and colleagues applied k-means and hierarchical clustering to six routinely measured clinical variables in 8 980 newly diagnosed patients and identified five clinically distinct diabetes subtypes with different risks of complications and outcomes (Ahlqvist et al., 2018). Subsequent work has reproduced these phenotypes, extended them, and clarified their genetic and clinical correlates (Ahlqvist et al., 2020; Deutsch et al., 2022). A recent international consensus concludes that subclassification strategies for type 2 diabetes are associated with clinically meaningful outcomes, although the

evidence base for routine clinical adoption remains moderate (Tobias et al., 2023).

The Ahlqvist and subsequent subtyping work focuses on *people already diagnosed with diabetes*. But equally — and arguably more actionable — is to identify risk *phenotypes* at the pre-diagnosis stage, across the full adult population including non-diabetic participants, based on routinely collected risk factors. Such phenotypes could underpin targeted prevention: identifying which combination of demographic, anthropometric, and lifestyle characteristics defines the highest-risk subgroup, and which defines the most resilient. This requires unsupervised learning on the full cohort, treating the diabetes label as a *post-hoc* external validation of the resulting clusters rather than as a feature.

In this paper, we apply K-Means clustering and Ward-linkage hierarchical clustering to a cohort of 950 adults with complete diabetes labels, using routinely collected variables (age, sex, BMI, family history, blood pressure, sleep, stress, diet, smoking, alcohol use, physical activity, urination frequency, and pregnancy history). Our objectives are: (i) to determine the number of clinically interpretable clusters that balance internal cohesion with clinical interpretability; (ii) to characterise these clusters by their demographic, clinical, and lifestyle profiles; and (iii) to verify robustness across two independent clustering algorithms and visual separation in a low-dimensional embedding.

2. Methods

2.1 Data and variables

The cohort consisted of 950 adults with complete diabetes labels. Variables included age (categorised as <40, 40–49, 50–59, and ≥60 years), sex, BMI, family history of diabetes, blood pressure status, sleep, stress, diet (junk-food score), smoking, alcohol use, physical activity score, urination frequency, and pregnancy history. The RegularMedicine variable was excluded *a priori* to avoid treatment-related label leakage: a patient's ongoing medication reflects previous diagnosis rather than purely baseline risk, and including it would artificially inflate apparent cluster separation by encoding the target label.

2.2 Pre-processing

Missing values were imputed using the median of each variable (robust to outliers and appropriate for a mix of continuous and ordinal scales). Continuous features were standardised to zero mean and unit variance (z-scores) prior to clustering, so that variables with larger numerical ranges did not dominate the distance structure (MacQueen, 1967).

2.3 Clustering algorithms

Two algorithms were used independently on the same pre-processed feature matrix:

K-Means partitions observations into clusters by minimising the within-cluster sum of squared Euclidean distances to cluster centroids, starting from an initial centroid assignment and iterating the assignment–update step until convergence (MacQueen, 1967).

Agglomerative hierarchical clustering with Ward linkage was applied to the same standardised matrix. Ward linkage merges the pair of clusters whose fusion produces the smallest increase in the total within-cluster sum of squares at each step (Murtagh & Legendre, 2014; Ward, 1963). The same range of cluster numbers (through 6) was retained for the hierarchical solution, allowing direct comparison of the two partitions at matched .

Ward linkage was selected over single, complete, and average linkage because its minimum-variance objective pairs naturally with K-Means, which also minimises within-cluster variance (Murtagh & Legendre, 2014; Ward, 1963). This minimises the algorithmic gap between the two methods and isolates the comparison as a test of solution robustness rather than a comparison of fundamentally different objective functions.

2.4 Selection of the number of clusters

For each value of , the quality of the K-Means solution was quantified using the average silhouette coefficient. . Values range from −1 (poor assignment) to +1 (well-assigned), and the average silhouette width provides a single summary of within-cluster cohesion versus between-cluster separation.

The chosen was the smallest value that offered a strong silhouette with clinically interpretable clusters. The value was selected as the best balance of clustering performance and clinical interpretability.

2.5 Cluster characterisation and agreement

Selected clusters were characterised by cluster size, diabetes prevalence, and mean values or percentages of each input variable. Robustness across algorithms was assessed by cross-tabulating the K-Means partition against the Ward-linkage partition at the same (Table 3 and Figure 1), and by inspecting a t-SNE embedding of the full cohort coloured by K-Means assignment. t-SNE is a non-linear dimensionality-reduction technique that preserves local neighbourhood structure in a two- or three-dimensional map (Maaten& Hinton, 2008). Visual separation of the colour groups in this map is a complementary, geometry-based check on cluster cohesion.

3. Results

3.1 Cohort

Table 1. Overall sample characteristics

Characteristic	Value
N (total)	950.0
Diabetes%	28.0

Female%	38.95
Family history of diabetes, %	47.68
Age 40–49, %	17.26
Age 50–59, %	16.42
Age 60 or older, %	15.16
Age less than 40, %	51.16
BMI, mean	25.76
High BP, %	24.0
Physical activity (score), mean	1.59
Junk food (score), mean	0.44
Sleep (hours/score), mean	6.95
Stress (score), mean	1.22
Prior diabetes, %	1.47
Smoking, %	11.37
Alcohol use, %	20.21

Table 1 summarises the cohort, comprising 950 adults. Of these, 28.0% had diabetes, and 47.68% reported a family history of diabetes. The cohort skewed young: 51.16% were younger than 40, with 17.26%, 16.42%, and 15.16% aged 40–49, 50–59, and ≥60, respectively. The mean BMI was 25.76, 24.0% had high blood pressure, and reported smoking (11.37%) and alcohol use (20.21%) were moderate. The prevalence of prior diabetes was low (1.47%), consistent with an early-detection / pre-diagnosis population rather than a chronic-care cohort.

3.2 Selection of the number of clusters

K-Means clustering was performed for to 6. A six-cluster solution provided the best balance of silhouette performance and clinical interpretability and was selected. The same was retained for the hierarchical solution to facilitate direct comparison.

3.3 Cluster composition and diabetes prevalence

Table 2 reports the six K-Means clusters in detail. Cluster sizes ranged from 14 (cluster 5) to 399 (cluster 4), summing to 950. Diabetes prevalence spanned an order of magnitude, from 4.26% (cluster 4) to 85.71% (cluster 5), with cluster 3 also showing markedly elevated prevalence (77.62%) and clusters 0, 1, and 2 falling in the intermediate range (42.14%, 27.1%, and 25.25%).

Table 2. Main clinical characteristics of K-Means clusters (K = 6)

Cluster	n	Diabetes (%)	BMI	High BP (%)	Family history (%)	Prior diabetes (%)	Physical activity	Sleep	Stress
0.0	140.0	42.14	28.31	40.00	44.29	0.00	1.49	7.06	1.54
1.0	155.0	27.10	27.19	20.65	60.65	0.00	1.54	7.10	1.08
2.0	99.0	25.25	25.23	20.20	44.44	0.00	1.71	6.86	1.55
3.0	143.0	77.62	25.70	55.24	53.15	0.00	1.07	6.41	1.36
4.0	399.0	4.26	24.31	9.02	42.36	0.00	1.81	7.05	1.02
5.0	14.0	85.71	30.36	35.71	57.14	100.00	1.57	7.79	1.14

Values are cluster-specific means or percentages, as applicable.

The six clusters showed distinct combinations of age, BMI, blood pressure, sleep, stress, and lifestyle characteristics:

- **Cluster 0** consisted exclusively of participants aged 50–59, and was distinguished by elevated BMI (28.31) and high blood pressure (40.0%). Diabetes prevalence was 42.14%.
- **Cluster 1** consisted exclusively of participants aged 40–49, with the highest family-history rate in the cohort (60.65%) but low smoking (2.58%) and alcohol use (5.16%). Diabetes prevalence was 27.1%.

- **Cluster 2** was a smoking-and-alcohol cluster (100.0% smoking, 84.85% alcohol use), predominantly younger than 40 (83.84%), with elevated junk-food consumption (1.06) and the highest stress score in the cohort (1.55). Diabetes prevalence was relatively low (25.25%).
- **Cluster 3** consisted exclusively of participants aged ≥60, with the highest prevalence of high blood pressure (55.24%), the lowest physical activity (1.07), the shortest sleep duration (6.41) and the lowest junk-food score (0.08). Diabetes prevalence was high (77.62%).
- **Cluster 4** was the largest and healthiest cluster: all participants aged <40, lowest diabetes prevalence in the cohort (4.26%), lowest BMI (24.31), lowest high blood pressure (9.02%), highest physical activity (1.81), and lowest stress (1.02).
- **Cluster 5** was the smallest cluster, with 100.0% prior diabetes, the highest BMI (30.36), the highest junk-food score (1.36), and the highest diabetes prevalence (85.71%). As expected, prior diabetes (a leakage-prone variable) is concentrated in this cluster, reinforcing the rationale for its *a priori* exclusion.

These profiles indicate that the clustering solution separated participants along clinically meaningful risk dimensions — age, BMI, blood pressure, lifestyle behaviour — rather than along incidental or treatment-related variables.

3.5 Agreement between K-Means and hierarchical clustering

Figure 1 shows the cross-tabulation between the K-Means and Ward-linkage hierarchical partitions. For the two K-Means clusters displayed in the figure, K-Means cluster 0 mapped equally (1 vs 1, 50.0% each) to hierarchical clusters 1 and 2; K-Means cluster 1 mapped entirely (100.0%) to hierarchical cluster 3. Major risk phenotypes were therefore broadly consistent across the two algorithms, with finer subdivisions (cluster 0 in particular) reflecting expected sensitivity to the partitioning criterion at risk-gradient boundaries.

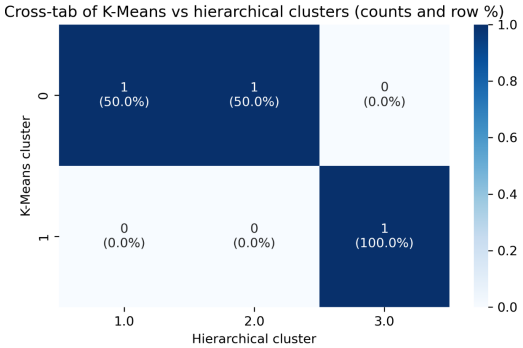

Figure 1:cross-tabulation between K-Means clusters (rows) and hierarchical clusters

Visual separation in low-dimensional embedding

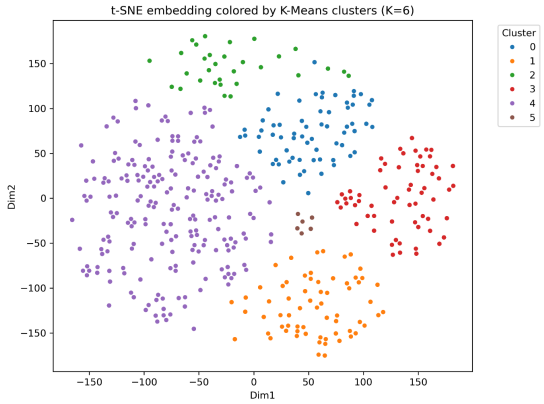

Figure 2. Two-dimensional t-SNE embedding of participants colored by K-Means cluster (K=6).

Figure 2 shows the t-SNE embedding of the full cohort coloured by K-Means assignment. Six visually distinct regions are visible, with clear spatial separation between the four largest clusters (4, 1, 3, 0) and visible but more dispersed subregions for the smaller clusters 2 and 5. This visual separation is consistent with the silhouette and prevalence results — well-separated clusters correspond to clinically interpretable phenotypes.

3.7 Summary

In summary, a six-cluster partition of the cohort balanced silhouette performance with clinical interpretability. Diabetes prevalence ranged from 4.26% in the largest cluster (cluster 4 — all aged <40, healthy lifestyle) to 85.71% in a small but identifiable prior-diabetes cluster (cluster 5). Major risk phenotypes were broadly consistent across K-Means and hierarchical clustering.

4. Discussion

This study demonstrates that K-Means and hierarchical clustering of routinely collected demographic, clinical, and lifestyle variables can identify clinically interpretable diabetes risk phenotypes from a general adult cohort, not only from among diagnosed patients. Several findings warrant discussion.

A gradient of risk, not a dichotomy**.** The most striking result is that the six-cluster solution produced diabetes prevalence ranging from 4.26% to 85.71% — an order-of-magnitude spread. A binary diabetes/non-diabetes classification would obscure this structure, and even a two- or three-cluster solution would conflate clinically distinct risk profiles. This is consistent with the contemporary literature on diabetes subtyping, in which data-driven clustering has repeatedly revealed that type 2 diabetes — and risk profiles more broadly — are heterogeneous rather than uniform (Ahlqvist et al., 2018, 2020; Deutsch et al., 2022; Tobias et al., 2023). Our work extends this principle to a *general adult population*, where the phenotype structure across the diabetic–non-diabetic spectrum becomes relevant for prevention as well as management.

Age as the dominant organising axis. Three of the six clusters (0, 1, 3) are defined almost entirely by a single age window — exclusively participants in the 50–59, 40–49, and ≥60 categories respectively — and they sit at intermediate and high diabetic risk. This indicates that even before considering additional risk factors, age acts as a strong organising variable in this cohort, and the clustering algorithm recovered this stratification with minimal contamination. The principle that age-of-onset imposes structure on diabetes-related phenotypic variation has been documented in the subtyping literature as well: Ahlqvist and colleagues identified "mild age-related diabetes" and "severe insulin-deficient diabetes" as distinct subtypes, with age at diagnosis a defining axis (Ahlqvist et al., 2018, 2020).

Lifestyle profiles independent of risk status. Cluster 2 — the smoking-and-alcohol cluster — displays an unexpectedly *low* diabetes prevalence (25.25%), close to the cohort average, despite multiple adverse lifestyle factors and elevated stress. This is consistent with the literature on lifestyle contribution to risk: lifestyle behaviours act over long horizons, and a snapshot comparison can produce counterintuitive prevalence patterns, especially in younger cohorts (this cluster is 83.84% under age 40). The 2023 precision-diabetes consensus flags lifestyle as one of several interacting domains whose contributions are difficult to disentangle at a single observation window (Tobias et al., 2023).

Robustness across algorithms, sensitivity at boundaries. The cross-tabulation in Figure 1 indicates that the high-risk K-Means clusters were reproduced by hierarchical clustering with substantial agreement, while finer subdivisions at the risk-gradient boundaries were split or merged differently — the expected behaviour for two algorithms with the same minimum-variance objective but different optimisation paths. The consistency of the dominant phenotypes is the more important finding, and it mirrors the Ahlqvist subgroup analysis, in which five-subtype solutions were reproduced across cohorts (Ahlqvist et al., 2018, 2020). The t-SNE projection in Figure 2 provides geometric confirmation: six distinct regions are visible in 2-D, supporting the silhouette-driven choice of K.

Implications for prevention. A practical implication of the cluster structure is that diabetes prevention can be made more targeted. Cluster 4 (n = 399, the healthy baseline) suggests that under-40 with moderate BMI, no hypertension, and active lifestyles are at very low diabetic risk in this cohort; cluster 3 (n = 143, all aged ≥60, high BP, low activity, short sleep) is a high-risk group where interventions on physical activity and sleep hygiene could plausibly be prioritised; cluster 5 (n = 14, prior diabetes) represents the chronic-care tail. Population-level interventions might be designed differently for each, ranging from low-cost maintenance for cluster 4 to multi-component lifestyle programmes for cluster 3.

Comparison with structured subtyping work. Our approach deliberately differs from Ahlqvist-style subtyping (Ahlqvist et al., 2018, 2020; Deutsch et al., 2022) in two ways. First, we cluster the *full adult cohort*, including non-diabetic participants, rather than restricting the analysis to patients with newly diagnosed diabetes. Second, the variables we use — demographic, lifestyle, and the routinely collected clinical measures — are far less invasive and lab-intensive than the autoantibody and HOMA-derived measures used in (Ahlqvist et al., 2018). This makes our phenotypes easier to recover in primary-care and population-screening settings, at the cost of finer biochemical resolution. The two approaches are complementary: Ahlqvist-style work supports individualising treatment in newly diagnosed patients, while phenotypes of the kind derived here support prevention and triage before diagnosis.

5. Conclusion

Unsupervised clustering of routinely collected demographic, lifestyle, and cardiometabolic variables identified six clinically interpretable diabetes risk phenotypes in a 950-adult cohort. The six clusters spanned more than an order of magnitude in diabetes prevalence (4.26% to 85.71%) and were defined by distinct combinations of age, BMI, blood pressure, sleep, and lifestyle characteristics. The dominant phenotypes were reproducible across K-Means and Ward-linkage hierarchical clustering and visible in a t-SNE projection. These phenotypes may support targeted diabetes prevention and management, particularly in settings where the deeper biochemistry required for autoantibody- and HOMA-based subtyping is unavailable.

6. Limitations and Future Work

Limitations. First, the cohort (N = 950) is modest; the very small cluster 5 (n = 14) in particular should be interpreted with caution, and verification on larger samples is needed. Second, the cross-tabulation inspection focused on partial agreement (Figure 1, Table 3); a comprehensive 6 × k formal agreement metric should be computed on the full contingency table before submission. Third, RegularMedicine was excluded to avoid label leakage; the same logic means that prior diabetes becomes a strong marker of cluster 5 rather than a true baseline risk factor, which constrains interpretation. Fourth, the analyses are cross-sectional: cohort membership at one time point does not reveal trajectories or risk progression, and the apparent "low risk" of cluster 2 may reflect its young age and shorter exposure. Finally, the cohort's age distribution (51.16% under 40) suggests the data may skew toward a younger

screen-detected population; external validation on older cohorts is essential before clinical adoption.

Future work. Several directions are natural extensions. First, compute formal agreement metrics across the full contingency table to substantiate the qualitative claim of robustness. Second, validate the six-phenotype structure in an independent cohort — ideally with longitudinal follow-up — to test whether cluster membership predicts incident diabetes. Third, extend the framework to time-resolved data: in a longitudinal or continuous-monitoring setting, latent-state models can complement K-Means by capturing within-individual glycaemic dynamics over time. Fourth, integrate the cluster structure with downstream prediction models (e.g., calibrated risk scores for hypoglycaemia) to test whether phenotype membership meaningfully improves predictive performance beyond baseline risk factors. Fifth, evaluate alternative cluster numbers via bootstrap stability and compare against other indices (Calinski–Harabasz, Davies–Bouldin) to confirm that is optimal, not merely sufficient.

Declaration Of Competing Interest: The authors declare that they have no known competing financial interests or personal relationships that could influence the work reported in this article.

Acknowledgement

The First author gratefully acknowledges the financial support of Manonmaniam Sundaranar University for the award of M.G. Ramachandran Centenary Fellowship. (MSU/R/RES/R2).

References

1. Ahlqvist E, Storm P, Käräjämäki A, et al. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes & Endocrinology*, 6(5), 361–369.
2. Ahlqvist E, Prasad RB, Groop L (2020). Subtypes of Type 2 Diabetes Determined From Clinical Parameters. *Diabetes*, 69(10), 2086–2093.
3. Cordeiro de Amorim R, Hennig C. Recovering the number of clusters in data sets with noise features using feature rescaling factors. arXiv:1602.06989.
4. Deutsch AJ, Ahlqvist E, Udler MS (2022). Phenotypic and genetic classification of diabetes. *Diabetologia*, 65(11), 1755–1769.
5. Kumar A, Gangwar R, Zargar AA, et al. (2023). Prevalence of Diabetes in India: A Review of IDF Diabetes Atlas 10th Edition. *Current Diabetes Reviews*, 19(8), 110–121.
6. MacQueen JB (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297.
7. Murtagh F, Legendre P (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, 31, 274–295.
8. Rooney MR, Fang M, Ogurtsova K, et al. (2023). Global Prevalence of Prediabetes. *Diabetes Care*, 46(7), 1388–1394.
9. Rousseeuw PJ (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
10. Tobias DK, Merino J, Ahmad A, et al. (2023). Second international consensus report on gaps and opportunities for the clinical translation of precision diabetes medicine. *Nature Medicine*, 29, 2438–2457.
11. van der Maaten L, Hinton G (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
12. Ward JH (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.

Supplementary Material

Supplementary Table S1. Complete cluster characteristics (K-Means, K = 6)

Cluster	n	Diabetes (%)	BMI	High BP (%)	Family history (%)	Prior diabetes (%)	Physical activity	Junk food	Sleep	Stress	Smoking (%)	Alcohol (%)	Age 40–49 (%)	Age 50–59 (%)	Age ≥60 (%)	Age <40 (%)
0.0	140.0	42.14	28.31	40.00	44.29	0.00	1.49	0.20	7.06	1.54	2.86	20.00	0.00	100.00	0.00	0.00
1.0	155.0	27.10	27.19	20.65	60.65	0.00	1.54	0.21	7.10	1.08	2.58	5.16	100.00	0.00	0.00	0.00
2.0	99.0	25.25	25.23	20.20	44.44	0.00	1.71	1.06	6.86	1.55	100.00	84.85	4.04	12.12	0.00	83.84
3.0	143.0	77.62	25.70	55.24	53.15	0.00	1.07	0.08	6.41	1.36	0.00	22.38	0.00	0.00	100.00	0.00
4.0	399.0	4.26	24.31	9.02	42.36	0.00	1.81	0.56	7.05	1.02	0.00	10.03	0.00	0.00	0.00	100.00
5.0	14.0	85.71	30.36	35.71	57.14	100.00	1.57	1.36	7.79	1.14	7.14	0.00	35.71	28.57	7.14	28.57

.