Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Clustering analysis using Joint Dirichlet Mixture Model

Dr. G. Lakshmi Kameswari

Assistant Professor, Department of Applied Sciences & Humanities, MVSR Engineering College, Nadergul, Hyderabad, India.
Email: lakshmik_sh@mvsrec.edu.in

Abstract

Diabetes is a major public health challenge in India and worldwide, with increasing prevalence and diverse risk patterns across populations. Traditional statistical methods, such as histograms, kernel density estimation, and maximum likelihood estimation, often fail to fully capture the complexity of diabetes-related data. These methods typically assume fixed parametric forms or provide limited uncertainty quantification, which restricts their applicability in clinical and epidemiological decision-making. Recent studies have demonstrated that Bayesian approaches can overcome these limitations by allowing flexible estimation of probability density functions (PDFs), integrating prior knowledge, and explicitly quantifying uncertainty through posterior distributions. In particular, Dirichlet Process Mixture Models (DPMMs) and Bayesian networks have been successfully applied to estimate joint distributions of glycaemic markers, risk factors, and complications in diabetes datasets. However, despite global advances, there is limited application of Bayesian density estimation techniques to Indian diabetes data, where missing values, heterogeneous risk factors, and regional variation present additional challenges. In this paper the authors have attempted to model PIMA Indian Diabetes data set using Joint DPMM and dynamic clustering of the data is performed and tabulated. It is reported that out of available 768 rows of data, they are segregated into 03 significant clusters out of 12 dynamic clusters obtained and emphirical relations of glucose levels interms of the other variables such as Age, Sex, pregnancies, BMI, Insulin and Skin thickness with diabetes pedigree function is presented for prediction and forecasting.

Keywords: Joint Dirichilet Mixture model, Diabetes, PIMA Dataset, Bayesian nonparametric tool.

1. Introduction:

India carries a large and growing burden of diabetes, and many studies as reported by peer researchers have used Bayesian methods to estimate key distributions such as the density of glycaemic markers, risk factors, and complications. Recent research papers by Zhou et al[1] and K.L.Ong et al [2] relied on Bayesian meta-regression to estimate age and sex specific prevalence across countries, including India, stating flexible, uncertainty-aware density estimation is required for public health planning. These works modelled heterogeneity across studies and locations and produced full posterior intervals, which is directly aligned with goal of estimating probability density functions (pdfs) for diabetes-related variables with quantified uncertainty. Dirichlet process mixture models (DPMMs), a standard Bayesian nonparametric tool for density estimation have been used to learn complex joint distributions of predictors. Cardoso et

al. [3],[4] combined regression with a DPMM to model the joint density of clinical features while imputing missing predictors for counterfactual treatment selection in type 2 diabetes (T2D). Their approach produces posterior predictive distributions for both missing covariates and outcomes, and it improved decision support when data is incomplete. Bayesian density and mixture modelling have also supported population-level estimation for diabetic eye disease. Liang et al.[5] used a Bayesian mixture framework to estimate the prevalence of diabetic retinopathy from partially labelled EHR data, addressing a practical issue in which many patients remain unlabelled in routine care. Several authors have reported the direct relevance to probability density function (pdf) estimation for diabetes data. Initially, flexible Bayesian non-parametric provided density estimates without strong parametric assumptions and then Bayesian networks and belief networks have been applied to diabetes risk modelling, learning joint distributions among lifestyle and clinical factors and supporting probabilistic inference under uncertainty [11], [12]. These structures effectively encoded and estimated high-dimensional pdfs via factorization, which can be adapted to Indian cohorts where risk patterns vary by region and urbanicity.[7],[8]. Most of the Research studies explicitly framed their objective of research as "density estimation," many recent diabetes research papers have concentrated on estimating underlying distributions.

A Dirichlet Process Mixture Model (DPMM) is a Bayesian nonparametric method used for flexible probability density estimation and clustering. In traditional parametric models we assume a fixed number of components ie., in Gaussian mixture models, whereas DPMMs allow the number of mixture components to be potentially infinite and automatically determined by the data. This property makes them especially powerful when dealing with complex biomedical datasets, such as those arising from diabetes studies, where the underlying data distribution is often heterogeneous and unknown.

The DPMM is built on the Dirichlet Process (DP), which serves as a prior distribution over probability measures. The DP introduces flexibility by allowing new mixture components to be created as more data points are observed, while also reusing existing components.

Mathematically, a DPMM can be expressed as:

1. Each data point x_i is assumed to come from a mixture distribution $x_i \sim \sum_0^\infty \pi_k f(x_i|\theta_k)$, where $f(x_i|\theta_k)$ is component density with parameter θ_k , and π_k is the mixture density.
2. The parameter θ_k is drawn from a base distribution G_O, while the mixture weights are governed by a stick-breaking construction of the Dirichlet Process.

DPMMs are highly suited for medical datasets because, they adopt to multimodal distributions, which is common in glycaemic markers. Even they naturally handle uncertainty and missing data, producing posterior predictive distributions rather than single-point estimates. DPMM Models enable clustering of patient subgroups without requiring the number of clusters to be fixed in advance, which is useful for identifying hidden patterns among diabetic and non-diabetic populations.

Diabetes datasets indicate the Multimodality, Skewness & heavy tails, Heterogeneity across age/sex/region, it also consists of missingness and zero-inflation data with 0 values.

A DPMM naturally Learns complex, multi-peak pdfs without fixing the number of modes, and gives soft, uncertainty-aware clusters , that Provides posterior predictive distributions to impute missing values

coherently. It Scales to hierarchical setups and can be extended to mixed data types such as continuous and categorical data with zero-inflated features.

The objective of present work is estimating the univariate and multivariate probability distribution functions for various clinical important variables to discover latent clusters (ie., any hidden clusters) and quantification of uncertainty using DPMM. In the present work R-Software is used for analysis of the data .

2. Overview on the Indian Diabetes Data PIMA.

India is facing one of the largest diabetes burdens globally, with more than 100 million diagnosed cases and around 130 million individuals with prediabetes. Type 2 diabetes is more dominant, accounting for over 90% of cases, while Type 1 and gestational diabetes also contribute significantly. A concern in India is the early onset of Type 2 diabetes, where individuals in their 20s and 30s are increasingly being diagnosed, unlike the Western trend of onset after 40. This rise is driven by rapid urbanization, lifestyle, high-carbohydrate diets, and genetic susceptibility. Indians are also prone to central obesity and insulin resistance even at lower body mass indexes, making them especially vulnerable. Nearly half of cases remain undiagnosed, and complications such as cardiovascular diseases, kidney failure, nerve damage, and blindness impose a huge health and economic burden. Urban areas reported prevalence rates of 15–20%, nearly double that of rural regions, though rural cases are rising sharply due to changing lifestyles. Government initiatives like the National Programme for Prevention and Control of Cancer, Diabetes, Cardiovascular Diseases and Stroke (NPCDCS) and Ayushman Bharat aim to improve screening and care, but barriers such as cost, awareness, and access persist. Addressing the epidemic requires large-scale awareness campaigns, early screening, and lifestyle modifications, making diabetes not just a medical challenge but also a pressing public health and societal issue for India's future.

The PIMA Indian Diabetes dataset contained eight clinical variables that were important in diagnosing and predicting diabetes. Pregnancies recorded the number of times a woman had been pregnant, since gestational diabetes increased long-term risk. Glucose measured blood sugar concentration and served as a primary diagnostic indicator. Blood pressure was included because hypertension often coexisted with diabetes and aggravated cardiovascular and renal complications. Skin thickness was used as an estimate of body fat, and along with BMI (a measure of obesity), it reflected the role of excess weight in insulin resistance. Insulin levels indicated pancreatic response to glucose, where abnormal values suggested impaired regulation. The Diabetes Pedigree Function (DPF) quantified genetic risk based on family history, while age acted as a general risk factor, with diabetes being more prevalent in older individuals, though Indians showed earlier onset compared to other populations. Finally, the Outcome variable (0 = non-diabetic, 1 = diabetic) represented the target for classification. Collectively, these variables captured both physiological markers (glucose, BMI, insulin) and risk factors (pregnancies, blood pressure, age, heredity), making the dataset valuable for diagnostic modelling and early prediction of diabetes.

3. Algorithm for Implementation of Joint DPMM

i. Input Data : Load the PIMA Diabetes dataset with 8 features: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age

ii. Model Setup : Assume data points $X=\{x_1,x_2,...,x_n\}$, where $x_i \in R^8$. Where each x_i belong to latent cluster Z_i.The prior is unknown hence use Dirichlet Process Prior.

iii. Generative process: $G \sim DP(\alpha, G_0)$ where α = concentration parameter. For each cluster k: $\theta_k \sim G_0$ ie., Multivariate Gaussian prior which is equal to sum of mean and co-variance of all 8 variables.

iv. For each patient i , assign a cluster equal to $Z_i \sim$ Categorical(π).Draw observations $x_i \sim N(\mu_{zi}, \Sigma z_i)$.

v. Randomly assign data points to clusters. Initialize the cluster means μ_k and covariances Σk. Set DP concentration parameter α

vi. Use Collapsed gibb's sampling technique : For each patient i: Remove x_i from the current cluster and compute the posterior probability of assignment to existing cluster k evaluated as $P(z_i = k|z_{-i}, \alpha)\ \aleph(x_i|\mu_k, \Sigma_k)$

vii. The New cluster is obtained by the relation : $\alpha . \int \aleph(x_i|\mu\ \ , \Sigma\ \ ) dG_o$. Then sample a new cluster Z_i.

viii. Update the Bayesian posterior using Normal-Inverse-Wishart.

ix. Repeat the process until convergence is obtained and report the final cluster of patients for Medical study.

4. Results

The simulations were performed using R-Software. R-Software is freely downloadable software and it supports all the statistical operations. Dirichilet process Library is installed from the nearest CRAN Mirror and PIMA Diabetes data as obtained from the Kaggle dataset is given as input to the code written in R-Language. Totally 8 variables are enumerated in the dataset namely, Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age

The Data is analysed using Prior unknown and posterior known using DPMM. Totally 768 rows of data is finally clustered into 12 clusters as shown in the figure 1.

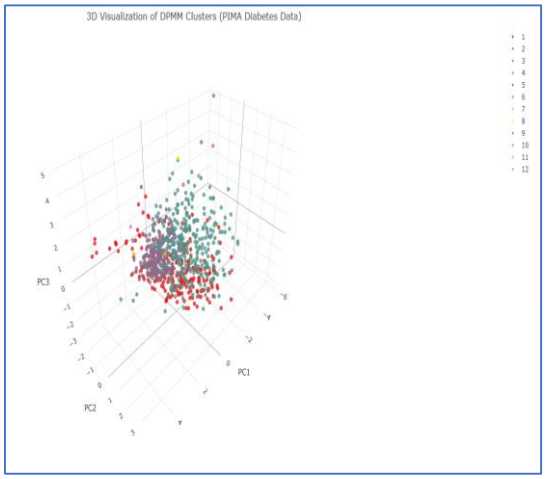


Figure-1: clusters as obtained by joint DPMM

Out of the 12 clusters 03 clusters are significant with 227 patients in cluster-01, 235 patients in cluster-02 and 295 patients in cluster-03. The remaining clusters are totally insignificant with patients either 1 or 2. Regression analysis is performed on the on all the clusters obtained from the DPMM and regression coefficients were tabulated and indicated in Table-01.

Empirical correlations were obtained for all the variables and out of them only relations for cluster-01 cluster-02 and cluster-03 is significant.

For Cluster-01: The empirical equation is given for glucose in-terms of others variables as given by :

Table 1: Regression Coefficients

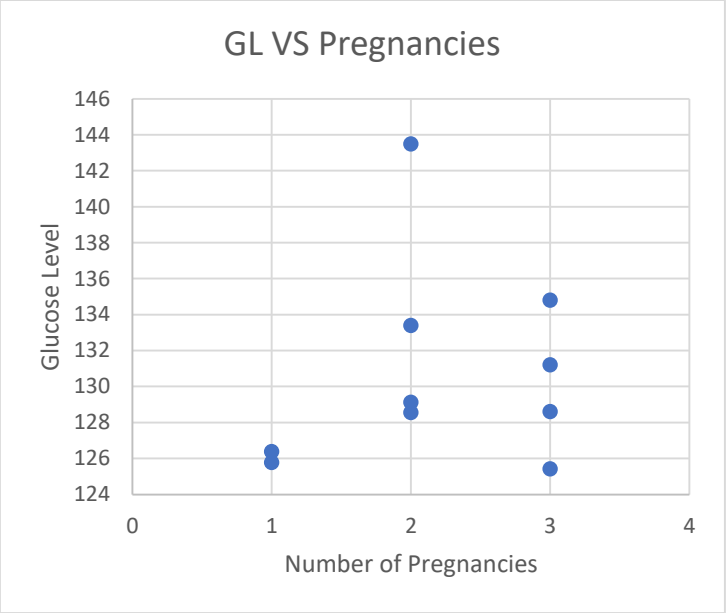
```
# A tibble: 20 × 7
   Cluster model  term                     estimate std.error statistic   p.value
   <fct>   <list> <chr>                       <dbl>     <dbl>     <dbl>     <dbl>
 1 1       <lm>   (Intercept)               78.7       8.42       9.35   1.02e-17
 2 1       <lm>   Pregnancies                0.320     0.593      0.540  5.90e- 1
 3 1       <lm>   BloodPressure              0.0152    0.0705     0.215  8.30e- 1
 4 1       <lm>   SkinThickness             NA        NA         NA     NA
 5 1       <lm>   Insulin                   NA        NA         NA     NA
 6 1       <lm>   BMI                        0.746     0.214      3.49   5.83e- 4
 7 1       <lm>   DiabetesPedigreeFunction   4.97      6.86       0.724  4.70e- 1
 8 1       <lm>   Age                        0.479     0.156      3.08   2.34e- 3
 9 2       <lm>   (Intercept)               81.3      11.1        7.33   3.94e-12
10 2       <lm>   Pregnancies               -0.355     0.822     -0.433  6.66e- 1
11 2       <lm>   BloodPressure              0.121     0.0957     1.26   2.09e- 1
12 2       <lm>   SkinThickness              0.178     0.149      1.20   2.32e- 1
13 2       <lm>   Insulin                    0.127     0.0189     6.69   1.73e-10
14 2       <lm>   BMI                       -0.269     0.234     -1.15   2.51e- 1
15 2       <lm>   DiabetesPedigreeFunction  -5.21      5.58      -0.933  3.52e- 1
16 2       <lm>   Age                        0.488     0.402      1.21   2.26e- 1
17 3       <lm>   (Intercept)               67.6      15.7        4.31   2.23e- 5
18 3       <lm>   Pregnancies               -0.520     0.639     -0.813  4.17e- 1
19 3       <lm>   BloodPressure              0.369     0.191      1.93   5.50e- 2
20 3       <lm>   SkinThickness             -0.0502    0.257     -0.196  8.45e- 1
```

Glucose = 78.72 + 0.32 * Pregnancies + 0.015 * Blood Pressure + NA * SkinThickness + NA * Insulin + 0.746 * BMI + 4.967 * DiabetesPedigreeFunction + 0.479 * Age.

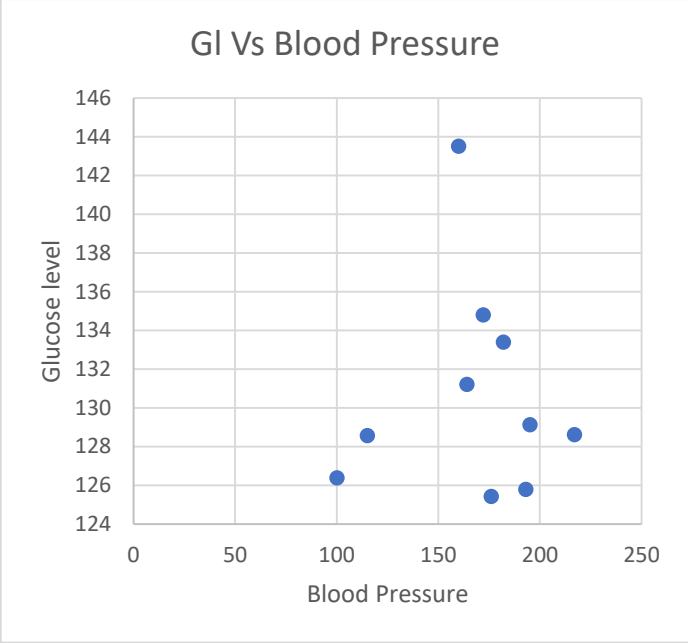
In this equation , the glucose levels of the patients are given interms of pregnancies , blood pressure, body mass index and Diabetes pedigree function with Age. Insulin and skin thickness are less significant and hence not applicable (NA) coefficient is presented. This indicates that the patients were on oral medication and not on insulin. The fat content is in limits hence, skin thickness is not a measurable parameter in the above equation.

Variable		Value
Pregnancies	A1	0.32
Blood Pressure	A2	0.015
Skin Thickness	A3	0
Insulin	A4	0
BMI	A5	0.746
Diabetes Pedegree Function	A6	4.967
Age	A7	0.479

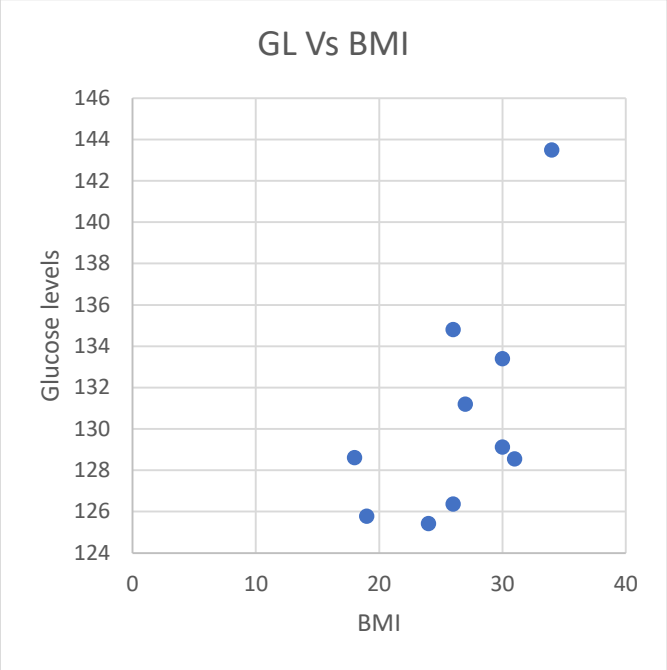
Glucose = 78.72 + 0.32*A1 + 0.015*A2 + 0.746*A5 + 4.967*A6+0.479*A7.



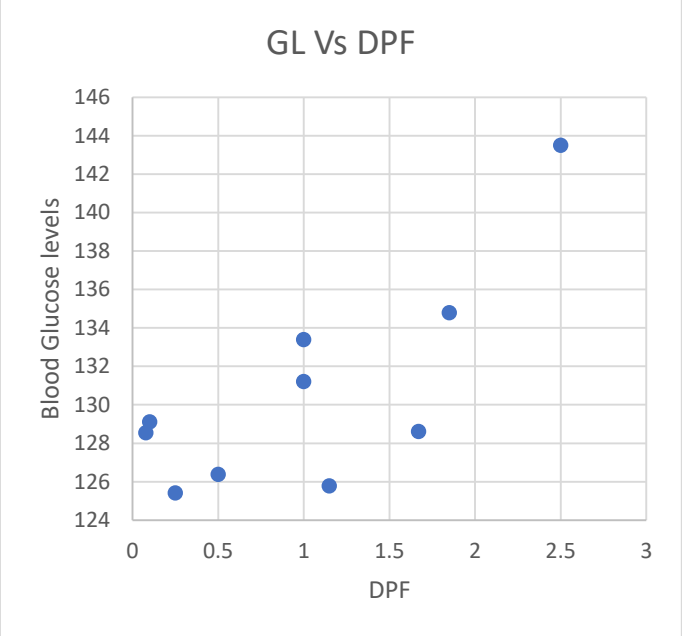
It is observed in women having a single pregnancy has high risk of pre-diabetic than with more than one pregnancy in cluster -01.



It is observed that there is some latent relationship between the Blood Pressure and Glucose levels within. Corresponding to Higher levels of BP, the risk of attaining Diabetes in come days is more likely than with those with levels of BP. Though data nalaysis with higher values of BP is indicates less levels of blood glucose levels , still there is a risk of attaining diabetes in near future as indicated in the above figure titled Glucose levels Vs Blood Pressure.



The variation of Glucose levels with BMI is plotted and observed that higher BMI is providing a risk of acquiring Diabetes.



Genetically acquired highly chronic diabetes can be identified by diabetes pedigree function. The variation with DPF for Blood glucose levels are estimated and identified that for cluster-01, the higher value of DPF than 2 indicates high risk of Diabetes.

Similarly, the empirical relations were developed for cluster 2 and cluster which are significant and tabulated presented.

Glucose = 81.329 -0.355 * Pregnancies + 0.121 * Blood Pressure + 0.178 * Skin Thickness + 0.127 * Insulin + -0.269 * BMI -5.206 * Diabetes PedigreeFunction + 0.488 * Age.

Variable		Value
Pregnancies (P)	A1	-0.355
Blood Pressure(BP)	A2	0.121
Skin Thickness(ST)	A3	0.178
Insulin(IN)	A4	0.127
BMI	A5	-0.269
Diabetes Pedegree Function(DPF)	A6	-5.206
Age(A)	A7	0.488

The Glucose level is given by :

GL= A1*P+A2*BP+A3*ST+A4*IN+ A5*BMI+A6*DPF +A7*A

For Cluster-03, the empirical relation is given by

Glucose = 67.619 + -0.52 * Pregnancies + 0.369 * Blood Pressure + -0.05 * Skin Thickness + 0.108 * Insulin + 0.382 * BMI + 0.644 * DiabetesPedigreeFunction + 0.297 * Age

Variable		Value
Pregnancies (P)	A1	-0.52
Blood Pressure(BP)	A2	0.369
Skin Thickness(ST)	A3	-0.05
Insulin(IN)	A4	0.108
BMI	A5	0.382
Diabetes Pedegree Function(DPF)	A6	0.644
Age(A)	A7	0.297

GL= A1*P+A2*BP+A3*ST+ A4*IN+ A5 * BMI +A6*DPF +A7*A.

5. Conclusions

It is observed that a large data set containing 768 rows with 8 columns data containing information about the parameters such as Pregnancies, Age, BMI, DPF, blood pressure, insulin dependency, with skin thickness is segregated into 12 dynamic clusters and most significant cluster is cluster-02 which contained about 235 patients have given a correlation :

Glucose = 81.329 -0.355 * Pregnancies + 0.121 * Blood Pressure + 0.178 * Skin Thickness + 0.127 * Insulin + -0.269 * BMI -5.206 * DiabetesPedigreeFunction + 0.488 * Age.

This empirical relation is related with all the variables of relevance and is used for forecasting / predicting possible diabetes for patients in coming time.

The author would like to acknowledge the Kaggle for using the PIMA Indian Diabetes dataset for analysis and would acknowledge the works of the authors who has contributed for the database repository, which is publicly available for academic research.

References

1. B. Zhou et al., "Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants," The Lancet, 2024
2. K. L. Ong et al., "Global, regional, and national burden of diabetes from 1990 to 2021," The Lancet, 2023.
3. P. Cardoso et al., "Dirichlet process mixture models to impute missing predictor data in counterfactual prediction models: an application to predict optimal type 2 diabetes therapy," BMC Medical Informatics and Decision Making, vol. 24, no. 1, p. 12, Jan. 2024.
4. P. Cardoso et al., "Dirichlet process mixture models to estimate outcomes for individuals with incomplete data," medRxiv, Jul. 2022 (preprint).
5. Y. Liang et al., "Estimating the prevalence of diabetic retinopathy in the presence of massive unlabeled EHR data," eClinicalMedicine, 2024. (Open-access version on PMC).
6. E. A. Lundeen et al., "Prevalence of diabetic retinopathy in the US in 2021," JAMA Ophthalmology, 2023. (Uses Bayesian meta-regression for subgrouped prevalence).
7. M. A. Escobar and M. West, "Bayesian density estimation and inference using mixtures," technical report/paper (classic reference; widely cited), accessed 2025.
8. S. Ghalebikesabi, C. Holmes, E. Fong, and B. Lehmann, "Quasi-Bayesian nonparametric density estimation via autoregressive predictive updates," Proceedings of Machine Learning Research, 2023.
9. N. A. Burhanuddin et al., "Clustering mixed-type data via Dirichlet process mixture models with flexible covariance," Symmetry, 2024. (Method supports density modeling for mixed clinical variables).
10. V. Inácio and M. de Carvalho, "Bayesian nonparametric inference for the overlap coefficient," Statistics in Medicine, 2022. (DPMM-based distributional estimation).