

Clustering of Myopia Patients Based on Severity Level, Geographic Area, Age, and Gender Using the K-Means Algorithm: A Case Study of Indra Optik Blitar

Ayuna Rahim Imannur^{1*}, Abdi Pandu Kusuma¹, Wahyu Dwi Puspitasari¹

¹Informatics Engineering, Faculty of Engineering and Information Technology, Universitas Islam Balitar, Blitar, Indonesia

Article History

Submitted: 08-07-2026

Revised: 04-08-2026

Published: 18-09-2026

DOI: 10.32664/j-intech.v14i03.2455

Keywords

Feature Weighting; Knowledge Discovery In Database; K-Means Clustering; Myopia; Optuna; Patient Segmentation

***Corresponding Author:**

ayuna.rachim@gmail.com

Abstract

Indra Optik Blitar manages a large volume of patient medical record data that has not yet been utilized to support strategic decision-making. This study aims to cluster myopia refractive error patients based on their severity level, region, age, and gender characteristics using the K-Means algorithm, and to evaluate the quality of the resulting clusters using the Silhouette Coefficient. The research process follows the Knowledge Discovery in Database (KDD) framework, encompassing data selection, preprocessing, feature engineering, Min-Max Scaling normalization, and clustering execution. From 14,863 raw data entries, 6,821 valid records were obtained after three filtering layers: demographic, geographic, and clinical. Feature engineering was performed by converting four raw refraction variables into a single Worst_SE (Worst Spherical Equivalent) variable and transforming district address data into latitude and longitude coordinates. Feature weighting was automatically optimized using the Optuna framework with the Tree-structured Parzen Estimator (TPE) algorithm through 100 optimization trials. The optimal number of clusters was determined using a combination of the Elbow Method and Silhouette Coefficient, resulting in $K = 5$. The K-Means algorithm with K-Means++ initialization achieved convergence at iteration 20, producing five clusters with distinctive characteristics. Evaluation using the Silhouette Coefficient yielded an average score of 0.4144, which, given the high dimensionality and heterogeneity of the combined clinical, demographic, and geographic features, represents a moderately well-separated cluster structure suitable for practical application. The five clusters formed represent patient segments that differ clinically, demographically, and geographically, providing a data-driven foundation for Indra Optik Blitar to design more targeted marketing strategies, eye health socialization programs, and social assistance initiatives.

1. Introduction

The Industry 4.0 and big-data era has fundamentally changed how organizations compete: data is no longer merely a transactional archive but a strategic asset, and the ability to convert raw records into actionable insight has become a key source of competitive advantage [1]. This shift is not limited to large technology firms; small and medium enterprises in retail and healthcare, including optical clinics, also accumulate substantial volumes of patient data every day. In many such organizations, however, this data is only used for administrative purposes such as billing and medical archiving, a condition commonly referred to as “data rich, information poor” [1].

Myopia (short-sightedness) is one of the most common refractive disorders encountered in optical services and carries a well-documented global burden. An estimated 1.4 billion people (22.9% of the world population) were myopic in 2000, and this figure is projected to reach 4.75 billion (49.8%) by 2050, with 938 million people (9.8%) expected to develop high myopia, a condition that substantially raises the risk of cataract, glaucoma, and retinal detachment [2]. This global trend translates directly into rising demand for local vision-correction services, placing clinics such as Indra Optik Blitar, an officially affiliated BPJS Kesehatan referral facility located on Jl. Merdeka No. 127, Blitar City, in a strategically important position.

Indra Optik Blitar records more than 11,000 patient visits per year, spanning insured, student, and general-public segments across Blitar City (population 150,371) and Blitar Regency (population 1,263,650) [3], [4], two areas with markedly different demographic and accessibility profiles. Despite this volume, interviews with the clinic's management confirmed that the archived medical records have never been processed beyond administrative use, and that promotional and health-outreach strategies are still designed generically, without any customer segmentation. This gap creates a pattern of “blind marketing” in which resources are spread evenly across patient groups instead of being prioritized toward the segments that need them most.

Clustering, an unsupervised learning technique that groups data by similarity without predefined labels, offers a direct route to closing this gap. Among clustering algorithms, K-Means is widely favored for its simplicity, computational efficiency, and ease of implementation on large datasets [5], and it has previously been applied to hospital medical records [6] and optical retail sales [7], though typically using a single demographic or transactional dimension and manually assigned feature weights, a practice that risks sub-optimal cluster separation [8]. This study extends that line of work by jointly clustering patients on three dimensions, clinical severity, demography (age, gender), and geography (sub-district location), and by replacing manual weighting with automatic feature-weight optimization. This combination is intended to let the severity of myopia remain the dominant driver of cluster formation, as is clinically appropriate [1], while still allowing spatial and demographic structure to surface, an approach still uncommon in local optical-industry clustering studies [6], [7].

Following the applied-research paradigm, which is oriented toward testing and evaluating a theory's capacity to solve a real practical problem [9], and taking an unsupervised, exploratory rather than hypothesis-confirming stance [10], this study aims to (1) apply the K-Means algorithm to cluster myopia patients at Indra Optik Blitar based on severity level, region, age, and gender, and (2) evaluate the resulting cluster quality using the Silhouette Coefficient. The resulting patient profiles are intended to give Indra Optik Blitar an empirical, data-driven foundation for designing targeted marketing, eye-health outreach, and social-assistance programs, replacing the intuition-based approach used until now.

The main contributions of this study are threefold: (1) it jointly clusters myopia patients across three previously unconnected dimensions, clinical severity, demography, and geography, within a single K-Means model, rather than the single-dimension approaches used in prior local studies; (2) it replaces a single, manually fixed weight per feature with automated, Optuna-optimized weighting within clinically-informed search bounds, validated jointly against the Elbow and Silhouette criteria, rather than relying on one arbitrary weight assignment; and (3) it translates the resulting cluster structure into concrete, data-driven marketing, outreach, and social-assistance recommendations for an operating optical clinic, demonstrating a practical

pathway from academic clustering analysis to operational decision support.

2. Research Method

2.1 Research Design and Location

This study is an applied, quantitative research effort conducted at Indra Optik Blitar (Jl. Merdeka No. 127, Kepanjenkidul, Blitar City), chosen for its large and representative patient base spanning both Blitar City and Blitar Regency. The analysis uses one year of retrospective electronic medical-record data (January–December 2025) to capture the most recent prevalence pattern of myopia while minimizing seasonal bias.

2.2 Data Collection

Data collection combined three techniques. In-depth interviews with the clinic's management and branch operations staff validated the business problem, namely the absence of structured patient segmentation, and confirmed the availability and structure of the medical-record database. Direct observation of the operational system verified the completeness of the demographic (date of birth, gender, address) and clinical (spherical, cylinder, and axis values for both eyes) attributes required for clustering. Documentation study, the main data-collection technique, extracted the historical medical records that form the secondary dataset used in the analysis.

2.3 Research Framework

Data processing follows the Knowledge Discovery in Database (KDD) framework, comprising Selection, Preprocessing, Transformation, Data Mining, and Evaluation, as illustrated in Fig. 1.

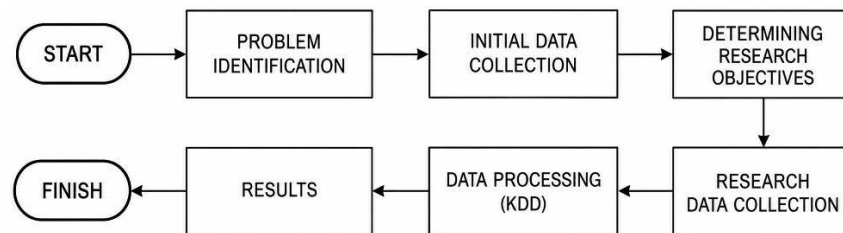


Figure 1. Research workflow following the KDD stages.

2.4 Data Selection and Preprocessing

The raw dataset comprised 14,863 medical records. Three sequential filtering layers were applied: demographic filtering removed records with incomplete identity attributes (name, age, gender); geographic filtering retained only patients residing in Blitar City or Blitar Regency to keep the case study spatially consistent and to avoid distance outliers; and clinical filtering retained only records with complete refraction values (right/left sphere and cylinder). The dataset was further restricted to patients aged 10–80 years and to myopia cases only (hyperopia and emmetropia were excluded), with a minimum correction threshold of -0.50 D. The minimum correction threshold of -0.50 D follows the International Myopia Institute's consensus definition for clinical myopia [11]. The 10–80 year age range reflects a study-specific methodological criterion rather than a universal clinical guideline: patients younger than 10 years were excluded because ocular refraction during childhood is still undergoing growth-related change and tends to fluctuate, making longitudinal comparison across patients less stable, and pediatric myopia guidelines in the literature vary considerably in their lower age bound, with some using age 6 instead. Patients older than 80 years were excluded because vision impairment in this age group is frequently caused by cataract and other age-related conditions rather than primary refractive error [12], consistent with reports that myopic shifts in very elderly populations are predominantly cataract-driven. Myopia severity was classified into three levels: low (mild) myopia, -0.25 D to -3.00 D; moderate myopia, -3.25 D to -6.00 D; and high (severe) myopia, worse than -6.00 D [13]. Patient names were anonymized to initials throughout the analysis.

2.5 Feature Engineering

Because K-Means requires a single numeric value per attribute, the right- and left-eye sphere (SPH) and cylinder (CYL) values were combined using the Spherical Equivalent (SE) formula [14], [15]:

$$SE = SPH + \frac{CYL}{2} \quad (1)$$

The more negative (worse) of the two eyes was then taken as the patient's final clinical attribute, consistent with clinical practice of basing risk classification on the more severely affected eye:

$$SE_{\text{worst}} = \min(SE_{\text{od}}, SE_{\text{os}}) \quad (2)$$

where SE_{od} and SE_{os} denote the spherical equivalent of the right (Oculus Dextra) and left (Oculus Sinistra) eye. Patient sub-district addresses were geocoded into latitude/longitude coordinates using each sub-district's centroid, and gender was binary-encoded (male = 0, female = 1).

2.6 Normalization and Feature Weighting

The resulting attributes, age, gender, latitude, longitude, and Worst_SE, differ by several orders of magnitude, which would let large-magnitude attributes such as age dominate the Euclidean distance computation. Min-Max normalization was therefore applied to rescale every attribute to the [0, 1] range:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (3)$$

Equal scaling, however, does not guarantee that each attribute contributes appropriately to cluster formation. An initial manual weighting scheme produced a low Silhouette Score, so feature weights were instead optimized automatically using the Optuna framework with the Tree-structured Parzen Estimator (TPE) algorithm [16]. Optuna searched jointly over the weight of each feature and the number of clusters K, running each candidate configuration through K-Means and using the resulting Silhouette Score as the objective to be maximized, so that the search was guided adaptively toward higher-quality cluster structures rather than proceeding by random trial. Specifically, the search space for K spanned the integer range [4, 7]. Feature weight bounds were constrained asymmetrically based on domain knowledge rather than searched over a uniform range: Worst_SE was allotted the widest and highest bound, [1.60, 2.20], to ensure clinical severity could dominate cluster formation; age and the spatial coordinates (latitude, longitude) were bounded at [0.80, 1.10] and [0.85, 1.10] respectively; and gender was deliberately constrained to a low range, [0.05, 0.15], to limit its influence relative to clinical and spatial attributes. Optuna's TPE sampler therefore optimized the exact weight value within each pre-specified, clinically motivated bound rather than discovering the relative feature importance from an unconstrained search. The optimization ran for 100 trials in total, as reported in the abstract.

2.7 K-Means Clustering

Cluster assignment is based on Euclidean distance between each patient's weighted, normalized feature vector and each cluster centroid:

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (4)$$

Centroids are initialized using K-Means++ to avoid poor local optima from arbitrary initialization, and are iteratively updated as the mean of their assigned members,

$$c_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_i \quad (5)$$

until convergence, i.e., until cluster membership no longer changes between iterations.

2.8 Cluster Validation

Cluster quality was validated internally using the Silhouette Coefficient [17], which jointly considers intra-cluster cohesion, $a(i)$, the mean distance from point i to the other members of its own cluster, and inter-cluster separation, $b(i)$, the mean distance from point i to the members of the nearest neighboring cluster:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (6)$$

The score ranges from -1 to $+1$; values approaching $+1$ indicate a strong, well-separated cluster structure, while values near 0 indicate overlapping clusters. The mean of $s(i)$ across all patients is reported as the model's global Silhouette Score. For visualization purposes, PCA was applied to the raw (unweighted, unnormalized) clinical, demographic, and geographic variables, Worst_SE, age, latitude, longitude, and encoded gender, rather than to the weighted, normalized feature matrix used for clustering. This visualization therefore illustrates the natural variance structure of the underlying patient attributes rather than the geometry K-Means directly operated on.

3. Results and Discussions

3.1 Processing Outcome

Applying the three preprocessing filters described in Section 2.4 reduced the dataset from 14,863 raw records to 6,821 valid records, retaining about 45.9% of the original data (Table 1). Geographic filtering removed the largest share (4,225 records), reflecting the proportion of visiting patients residing outside the Blitar City/Regency case-study area; clinical filtering removed a further 3,203 records with incomplete refraction values, a necessary step since these four fields underlie the Worst_SE feature used in clustering.

Table 1. Data reduction through the preprocessing stages.

Preprocessing Stage	Records Remaining	Records Removed
Raw data	14,863	—
Demographic filtering	14,249	614
Geographic filtering	10,024	4,225
Clinical filtering	6,821	3,203

3.2 Feature Weighting and Optimal Cluster Number

Optuna's TPE search over the feature-weight space returned an optimal configuration with an estimated Silhouette Score of 0.4141 and a recommended $K = 5$. The optimized weights placed the highest priority on Worst_SE (1.642), consistent with the search space's clinically-motivated design, which allotted it the widest and highest bound to keep severity dominant, followed by longitude (1.093) and age (1.055), while latitude (0.984) and gender (0.051) settled near the lower end of their comparatively narrow, deliberately constrained bounds.

This $K = 5$ recommendation was cross-checked against the classical Elbow and Silhouette curves computed for $K = 2$ to $K = 10$ (Fig. 2). The Elbow curve (SSE) decreases sharply up to $K = 5$ and flattens noticeably afterward, while the Silhouette curve, although highest at the practically uninformative $K = 2$, peaks again at $K = 5$ (≈ 0.414) among the more meaningful range of $K \geq 3$. The convergence of the Optuna, Elbow, and Silhouette evidence supports $K = 5$ as the number of clusters used in the final model. A K-Means++ seed search over 200 candidate initializations selected seed 2 (Silhouette 0.4144, inertia 373.10), and the final weighted K-Means run converged after 20 iterations. The minimal 0.0003 gap between the Optuna-estimated score (0.4141) and this final evaluated score (0.4144) is not attributable to rounding or to a different data subsample; both stages operate on the identical, full 6,821-patient weighted feature matrix. The difference instead reflects a difference in seed-search rigor between the two stages: Optuna evaluated each candidate weight configuration using a single fixed random seed (random_state = 42) with ten internal K-Means++ initializations selected by lowest inertia, the scikit-learn default, whereas the dedicated post-hoc search explicitly evaluated 200 distinct seeds

by Silhouette Score directly, rather than by inertia, and selected the seed yielding the highest score. This more exhaustive, silhouette-driven seed search accounts for the small residual improvement from 0.4141 to 0.4144.

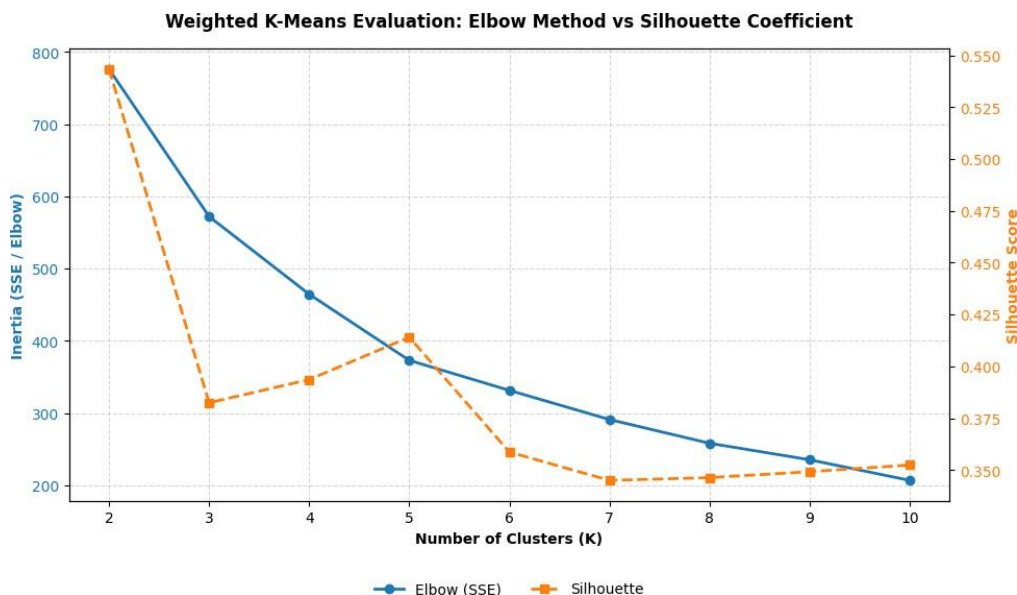


Figure 2. Elbow (SSE) and Silhouette Score across $K = 2$ to $K = 10$.

3.3 Cluster Composition and Profiles

The final weighted K-Means model produced five clusters of markedly uneven size (Table 2), a distribution that reflects the real concentration of patient characteristics in the field rather than an algorithmic artifact. Individual-level Silhouette evaluation, computed directly from $a(i)$, $b(i)$, and $s(i)$ for every patient, yielded a global average Silhouette Score of 0.4144, consistent with the Optuna estimate. Under the standard Silhouette interpretation scale, this value indicates a moderate (“fair”) but interpretable cluster structure [17].

Table 2. Patient population per cluster.

Cluster	Patients	Proportion
1	3,022	44.3%
2	110	1.6%
3	1,157	17.0%
4	891	13.1%
5	1,641	24.1%
Total	6,821	100%

Table 3 summarizes the demographic, clinical, and geographic profile of each cluster. Cluster 1 is the largest group (44.3%), dominated by young, urban, female patients with mild myopia. Cluster 3 is concentrated in the western sub-districts of Pongkok and Srengat and shows the broadest severity spread of the mild-dominant clusters, including 9.9% severe cases. Cluster 4 is the most clinically distinctive group: all of its 891 members fall in the moderate-to-severe myopia category (61.3% severe), concentrated in the 26–45-year productive-age range. Cluster 5 is the oldest group, spanning ages 36 to over 60, and is dominated by mild myopia, consistent with a presbyopia-relevant segment.

Table 3. Summary profile of the five K-Means clusters.

Cluster	Size (%)	Dominant Age	Dominant Severity	Dominant Region	Gender
1	3,022 (44.3%)	18–35 yr	Mild (74.6%)	Kanigoro, Sananwetan	Female 71.3%
2	110 (1.6%)	Mixed	Mild (74.5%)	Bakung (100%)	Female 72.7%
3	1,157 (17.0%)	18–35 yr	Mild (62.3%)	Ponggok, Srengat	Female 72.8%
4	891 (13.1%)	26–45 yr	Severe (61.3%)	Kanigoro, Sananwetan	Female 68.8%
5	1,641 (24.1%)	36–60+ yr	Mild (77.0%)	Sananwetan, Kanigoro	Female 70.3%

Note: Severity thresholds are mild, -0.25 D to -3.00 D; moderate, -3.25 D to -6.00 D; and severe, worse than -6.00 D [13].

Cluster 2's complete confinement to Bakung sub-district is confirmed, not merely inferred, by direct inspection of the weighted feature space: the cluster's centroid sits at $W_Lat = 0.000$, the absolute minimum of the normalized latitude range across all 6,821 patients, with near-zero within-cluster variance on both W_Lat and W_Lon , since sub-district centroid geocoding assigns every Bakung patient an identical coordinate. Members average a distance of 0.29 to their own centroid versus 1.00–1.22 to the other four centroids in the full weighted space, and a distance of essentially 0.0 versus 0.95–1.11 when latitude and longitude alone are considered, meaning geographic coordinates alone almost fully account for the separation. This separation holds even though the longitude weight (1.093) is only modest, close to the unweighted baseline of 1.0, unlike Worst_SE's substantially inflated weight (1.642): Bakung's raw coordinate distance from the other sub-district centroids is large enough on its own, since it lies roughly 50 km from Indra Optik Blitar's location at the southernmost, least populous edge of the regency, that even a near-neutral weight preserves a clearly separable gap. By contrast, the cluster's within-group spread in Worst_SE and age (SD 0.167 and 0.270) is comparable to that of the full dataset (SD 0.195 and 0.230), ruling out clinical or demographic homogeneity as the unifying factor. This indicates a distance artifact of centroid-based geocoding rather than a genuine clinical-geographic community: patients are grouped by shared sub-district location, not by shared health characteristics. An institutional referral pathway, such as a BPJS program or school-based screening specific to Bakung, cannot be excluded but is not testable with the variables recorded here.

Principal Component Analysis (PC1 = 97.2% variance, PC2 = 2.7%) visualizes this structure in two dimensions (Fig. 3). Cluster 4 separates clearly along PC2 from the remaining four clusters, confirming that Worst_SE, which loads heavily on PC1, is the dominant axis distinguishing patients, while Cluster 5 spreads further along PC1 due to its higher age values. PC1's dominant 97.2% variance share reflects a scale effect rather than the Optuna weighting scheme: because PCA was computed on raw, unnormalized values, Worst_SE, whose diopter range extends well beyond the -6.00 D severe-myopia threshold for the most affected patients, has substantially larger raw numeric variance than age (bounded 10–80 years) or the narrow-range latitude and longitude coordinates within a single regency, causing it to dominate the first principal component.

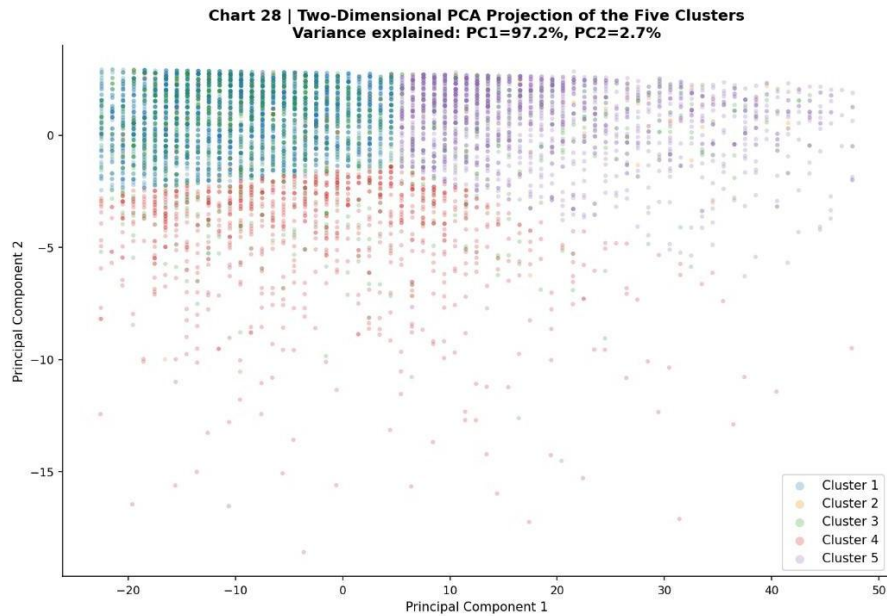


Figure 3. Two-dimensional PCA projection of the five clusters.

3.4 Comparison with Related Studies

The five-cluster outcome obtained here parallels a previous customer-segmentation study that likewise identified five clusters using a combined Elbow-Dendrogram approach [18]. This correspondence suggests that human populations, whether retail customers or medical patients, tend to separate into a similar number of distinct segments under multidimensional K-Means. The present study extends that finding by integrating clinical severity and geographic coordinates alongside demographic attributes within a single model, a combination absent from both that customer-segmentation data and earlier single-dimension clustering of hospital medical records [6]. Female patients dominate all five clusters (68.8–72.8%), consistent with the gender-skewed segments reported in hospital medical-record clustering and with K-Means customer segments led by female shoppers aged 20–40 [6], [18]. Clusters 1 and 3 here, dominated by ages 18–35 with the highest female share, closely mirror that profile, indicating that gender itself is not the main differentiator once severity and geography are accounted for.

Using sub-district latitude/longitude coordinates successfully isolated Cluster 2 (exclusively Bakung) and Cluster 3 (concentrated in Ponggok and Srengat). This result corroborates earlier evidence, derived from restaurant distribution, that coordinate-based K-Means can reveal spatial hotspots [19]. Extending that approach to patient distribution, the present study also bridges two previously separate applications of K-Means in eye care, namely the clustering of optical sales data alone and the clustering of ocular-disease image features alone [7], [20]. Combining clinical severity with a commercial patient base in one model, as done here, has not previously been reported in this domain.

The choice of K-Means over density-based alternatives is supported by evidence that K-Means produces a more cohesive structure than DBSCAN (Davies-Bouldin Index 0.39 vs. 1.39) on similarly continuous numeric data [21]. The resulting Silhouette Score of 0.4144 is lower than the score of 0.51 previously reported for three clusters formed from homogeneous, low-dimensional provincial data [22]. This difference is attributable to the higher dimensionality and heterogeneity of the present patient-level dataset, which is clinical, demographic, and spatial at once; a score in this range is therefore considered representative given the added complexity. It is further corroborated by a comparable Silhouette Score of 0.3718 reported for K-Means clustering of HIV screening data across health service units, a similarly complex, multi-domain patient-service clustering task that applied the same KDD framework used here, combined testing volume,

positivity rate, and gender-based indicators, and was published in J-INTECH [23]. The roughly two-fold improvement in Silhouette Score achieved through Optuna weighting mirrors the substantial gains reported when Optuna was applied to heart disease classification (accuracy from 77.04% to 90.16%), reinforcing that automated weight optimization is preferable to manually assigned weights [24]. Finally, relying on the Elbow curve alone would carry the subjectivity inherent in visually identifying an “elbow” point [25]. Pairing the Elbow curve with the Silhouette Coefficient therefore allowed $K = 5$ to be justified by two independent, mutually reinforcing criteria rather than visual judgment alone, in line with prior recommendations to combine both methods [26].

These comparisons translate into concrete actions for Indra Optik Blitar. Cluster 1's large young urban segment is best reached through digital and social-media marketing of standard, aesthetically oriented lenses, while Cluster 5's older patients represent a natural market for progressive or bifocal lenses supported by consultation-based service. Cluster 3's concentration in Ponggok and Srengat points to an under-served western catchment area that could justify a satellite service point or periodic outreach visits. Cluster 4, entirely composed of moderate-to-severe patients requiring costlier high-index lenses, and Cluster 2, facing limited geographic access from Bakung, are the clearest priorities for subsidized-lens or social-assistance programs, while Clusters 1 and 3, both dominated by younger patients, are the most relevant targets for school- and campus-based eye-health education.

4. Conclusions and Future Works

This study applied the K-Means algorithm, following the Knowledge Discovery in Database framework, to cluster myopia patients at Indra Optik Blitar by severity level, region, age, and gender. Sequential demographic, geographic, and clinical filtering reduced 14,863 raw medical records to 6,821 valid records, from which Optuna-optimized feature weighting, cross-validated against the Elbow and Silhouette curves, identified $K = 5$ as the optimal number of clusters. The resulting weighted K-Means model, initialized with K-Means++, converged into five clusters with clearly distinguishable identities: a large young urban segment with mild myopia (Cluster 1), a geographically singular Bakung segment (Cluster 2), a western-region segment concentrated in Ponggok and Srengat (Cluster 3), a clinically distinct moderate-to-severe segment concentrated in productive age (Cluster 4), and an older segment dominated by mild myopia (Cluster 5). Individual-level Silhouette evaluation across all 6,821 patients produced a global average score of 0.4144, indicating a moderately well-separated, interpretable cluster structure suitable for practical use.

These five profiles give Indra Optik Blitar a concrete, data-driven basis for targeted marketing, prioritized eye-health outreach, and social-assistance allocation, as detailed in Section 3.4. Indra Optik Blitar could also extend this clustering approach beyond myopia to the other refractive disorders it treats, such as hypermetropia, astigmatism, and presbyopia, so that the resulting segmentation represents its entire patient population, and could strengthen address-recording standards in its medical-record system to support finer-grained spatial analysis in the future.

The study is nonetheless limited by its use of sub-district centroid coordinates rather than precise patient addresses, and by its purely descriptive, cross-sectional scope, which does not extend to forecasting future patient behavior. Future research could therefore refine spatial resolution using API-based geocoding, such as OpenStreetMap Nominatim or the Open Source Routing Machine, combined with address text-matching or a real-time address-validation step at the point of data entry; incorporate temporal features such as visit frequency or dioptic progression to capture longitudinal patient trajectories; benchmark K-Means against DBSCAN, HDBSCAN, and Agglomerative Hierarchical Clustering with Ward linkage, each of which accommodates non-spherical cluster shapes that may better capture geographic hotspot patterns such as Bakung's; design the Optuna objective function around a specific managerial goal, such as maximizing severity-based separation for social-assistance targeting or geography-based separation for digital-marketing purposes, in closer collaboration with clinic management; and, in later operational deployments, automate monthly data extraction and re-clustering so that patient segments stay current with the clinic's most recent data rather than reflecting a single historical snapshot.

5. Declarations

Author Contributions

Ayuna Rahim Imannur: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Writing - Original Draft, Writing - Review & Editing. Abdi Pandu Kusuma: Supervision, Validation, Critical Review, Final Approval. Wahyu Dwi Puspitasari: Supervision, Validation, Critical Review, Final Approval. All authors have read and agreed to the published version of the manuscript.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Data Availability Statement

The datasets analyzed during the current study are not publicly available due to patient privacy and clinical confidentiality policies. However, anonymized data extracts may be available from the corresponding author on reasonable request, subject to non-disclosure agreements and explicit permission from the management of Indra Optik Blitar.

Use of Artificial Intelligence (AI)

The authors declare that Artificial Intelligence (AI) tools were utilized solely for language polishing, grammar correction, and formatting assistance during the preparation of this manuscript. The AI tools were not used to generate the core intellectual content, data analysis, or scientific conclusions of this study. All content and methodologies represent the original work of the authors.

Ethical Clearance

This study was conducted in accordance with the ethical principles of medical and clinical research. Administrative and ethical permission to access and utilize retrospective medical records for research purposes was officially granted by the management of Indra Optik Blitar. Patient confidentiality was strictly maintained, and all personal identifiers were fully anonymized prior to data processing and clustering analysis.

6. References

- [1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2011.
- [2] B. A. Holden *et al.*, "Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050," *Ophthalmology*, 2016, doi: 10.1016/j.ophtha.2016.01.006.
- [3] B. P. S. K. Blitar, *Indikator Kesejahteraan Rakyat Kota Blitar 2022*. 2022. [Online]. Available: <https://blitarkota.bps.go.id/id/publication/2022/12/28/ace8c12becce9e280e0a2a82/welfare-indicators-in-blitar-municipality-2022.html>
- [4] Dinas Komunikasi Informatika Statistik dan Persandian Kabupaten Blitar, "Jumlah penduduk."
- [5] I. Kamila, U. Khairunnisa, and Mustakim, "Perbandingan algoritma K-Means dan K-Medoids untuk pengelompokan," *J. Ilmiah Rekayasa dan Manajemen Sistem Informasi*, 2019, doi: 10.24014/rmsi.v5i1.7381.

- [6] A. Ali, "Klasterisasi data rekam medis pasien menggunakan metode K-Means clustering di Rumah Sakit Anwar Medika Balong Bendo Sidoarjo," *MATRIK: J. Manajemen, Tek. Inform. dan Rekayasa Komputer*, 2019, doi: 10.30812/matrik.v19i1.529.
- [7] M. Rizki and M. Mulyawan, "Penerapan metode K-Means clustering pada data penjualan Optik Chantika," *JATI (J. Mahasiswa Tek. Inform.)*, 2023, doi: 10.36040/jati.v7i2.6562.
- [8] R. C. de Amorim, "A survey on feature weighting based K-Means algorithms," *J. Classif.*, 2016, doi: 10.1007/s00357-016-9208-4.
- [9] M. O. Armandari, N. S. S. Ambarwati, and N. Hidayah, "Analisis kebutuhan bahan ajar dalam pembelajaran metode penelitian terapan di program studi kosmetik dan perawatan kecantikan," *Academy of Education Journal*, 2024, doi: 10.47200/aoej.v15i2.2493.
- [10] S. Setyaningtyas, B. I. Nugroho, and Z. Arif, "Tinjauan pustaka sistematis pada data mining: Studi kasus algoritma K-Means clustering," *J. Teknoif Tek. Inform. Institut Teknologi Padang*, 2022, doi: 10.21063/jtif.2022.V10.2.52-61.
- [11] D. I. Flitcroft *et al.*, "IMI - Defining and Classifying Myopia: A Proposed Set of Standards for Clinical and Epidemiologic Studies," *Invest. Ophthalmol. Vis. Sci.*, 2019, doi: 10.1167/iovs.18-25957.
- [12] A. Hashemi, M. Khabazkhoob, and H. Hashemi, "High prevalence of refractive errors in an elderly population; a public health issue," *BMC Ophthalmol.*, 2023, doi: 10.1186/s12886-023-02791-x.
- [13] T. Lestari, Anggunan, T. Triwahyuni, and R. Syuhada, "Studi Faktor Risiko Kelainan Miopia Di Rumah Sakit Pertamina Bintang Amin," *Jurnal Ilmiah Kesehatan Sandi Husada*, 2020, doi: 10.35816/jiskh.v11i1.275.
- [14] P. Oktarima, S. Caesarya, I. Irfani, M. W. Kuntorini, F. K. Memed, and D. V. Ginting, "Hubungan usia dan jenis kelamin dengan jenis kelainan refraksi pada anak di Pusat Mata Nasional Rumah Sakit Mata Cicendo," *Oftalmologi: J. Kesehatan Mata Indonesia*, 2021, doi: 10.11594/ojkmi.v3i2.15.
- [15] V. Galvis, A. Tello, P. A. Camacho, L. M. Gómez, J. J. Rey, and A. A. Serrano, "Definition of refractive errors for research studies: Spherical equivalent could not be enough," *J. Optometry*, 2021, doi: 10.1016/j.optom.2020.10.003.
- [16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019. doi: 10.1145/3292500.3330701.
- [17] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, 1987, doi: 10.1016/0377-0427(87)90125-7.
- [18] D. A. Jethva, S. M. Khamkar, A. A. Pachchhapur, and S. Kulkarni, "Customer segmentation analysis using K-means algorithm with Elbow method and Dendrogram," in *Proceedings of the 2023 6th International Conference on Advances in Science and Technology (ICAST)*, 2023. doi: 10.1109/ICAST59062.2023.10454952.
- [19] M. S. Rahman, T. D. Sarkar, and M. Y. Emon, "Comprehensive geospatial and statistical analysis of restaurant distribution using K-means clustering in machine learning technology," in *Proceedings of the 2024 4th International Conference Advancement in Electronics and Communication Engineering (AECE)*, 2024. doi: 10.1109/AECE62803.2024.10911593.
- [20] M. I. Mashur and Y. Salim, "Analisis performa metode cluster K-Means pada dataset Ocular Disease

Recognition," *Indonesian J. Data and Science*, 2022, doi: 10.56705/ijodas.v3i1.47.

- [21] N. P. Sutramiani, I. M. T. Arthana, P. F. Lampung, S. Aurelia, M. Fauzi, and I. W. A. S. Darma, "The performance comparison of DBSCAN and K-Means clustering for MSMEs grouping based on asset value and turnover," *J. Information Systems Engineering and Business Intelligence*, 2024, doi: 10.20473/jisebi.10.1.13-24.
- [22] Suharmanto, W. S. Utami, N. Pratiwi, and F. Muhammad, "Penerapan data mining menggunakan algoritma K-Means untuk clustering perokok usia lebih dari 15 tahun," *Bulletin of Information Technology (BIT)*, 2023, doi: 10.47065/bit.v4i4.1067.
- [23] Y. R. N. Manobi, A. De Kweldju, and J. P. P. Naibaho, "Clustering HIV Screening Data in Teluk Bintuni Using K-Means," *J-INTECH*, 2026, doi: 10.32664/j-intech.v14i02.2368.
- [24] D. Wijayanto, R. Marco, A. Sidauruk, and M. Sulistiyono, "The effect of SMOTE and Optuna hyperparameter optimization on TabNet performance for heart disease classification," *J. Sisfokom (Sistem Informasi dan Komputer)*, 2025, doi: 10.32736/sisfokom.v14i2.2348.
- [25] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP J. Wirel. Commun. Netw.*, 2021, doi: 10.1186/s13638-021-01910-w.
- [26] P. Vania and B. N. Sari, "Perbandingan metode Elbow dan Silhouette untuk penentuan jumlah kluster yang optimal pada clustering produksi padi menggunakan algoritma K-Means," *J. Ilmiah Wahana Pendidikan*, 2023, doi: 10.5281/zenodo.10081332.