Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Clustering of Blood Test-Based Liver Disorder Data Using K-Means and Principal Component Analysis (PCA)

Adrian Rangga Mafatihallah[1]*, Rivandi Faruqi[2], Ferdinandus Talu Tukan[3], Sumanto[4], Ade Surya Budiman[5]

[1,2,3,4,5] *Informatics Study Program, Faculty of Engineering and Informatics,Universitas Bina Sarana Informatika*
15220743@bsi.ac.id[1*] , 15220692@bsi.ac.id[2] , 15220655@bsi.ac.id[3] , sumanto@bsi.ac.id[4] , ade.aum@bsi.ac.id[5]

Abstract

Liver disease is one of the serious health disorders that requires early detection to ensure effective treatment. Laboratory examinations through blood tests are the primary method for identifying liver function abnormalities. However, the large number of variables in blood test data often complicates the analysis process. This study aims to cluster liver disorder patient data based on blood test results using the Principal Component Analysis (PCA) and K-Means Clustering methods. PCA is applied to reduce the dimensionality of the data to facilitate the clustering process, while K-Means is used to group data based on similarities in their characteristics. The dataset consists of 345 patient records with seven numerical attributes representing liver function indicators. The PCA results show that the first two principal components (PC1 and PC2) explain 59.6% of the total data variance. Clustering was performed using various numbers of clusters, with the best result obtained at K = 3, and a Silhouette Score of 0.202. Although this score is considered relatively low, the approach is capable of uncovering natural clustering patterns in the data. The results of the study indicate that the combination of PCA and K-Means can be used to assist in early medical screening for liver function disorders, although additional methods are required to improve the accuracy and validity of the clustering results.

Keywords: *Liver Disorders, PCA, K-Means, Clustering, Blood Test, Data Mining*

1. Introduction

Liver disease is a type of inflammation that attacks the liver, and is generally caused by unhealthy lifestyle habits [1]. The increasing data on liver disease requires the application of certain methods to process and obtain accurate information. This aims to improve the quality of information as well as efficiency and effectiveness in data management, so that it can facilitate the decision-making process, especially in efforts to overcome liver disease [2].The liver is the largest organ in the human body and has various vital roles. Its functions include processing nutrients from food, helping the process of breaking down fat, cleaning the blood from toxins, and maintaining stable nutrient levels in the body [3]. The liver plays a central role in the detoxification of chemical compounds through two main stages, namely the first stage involving the cytochrome P450 enzyme and the second stage in the form of a conjugation reaction. The level of enzyme activity in this process can vary up to 40-fold between individuals, influenced by genetic differences [4]. If there is a disruption in the function of liver cell synthesis, serum albumin levels will decrease (hypoalbumin), especially if there are extensive and chronic liver cell lesions [5]. Liver function tests are a collection of blood tests that aim to measure the levels of certain enzymes or proteins in the blood. This examination is usually used to detect, assess, and monitor liver disorders or damage [6]. Therefore, laboratory examination through blood tests is one of the main methods in detecting liver dysfunction early. Given the many variables in blood test data, a data analysis method is needed that can simplify the information without reducing the content of important information. One technique that can be used is Principal Component Analysis (PCA). This method can be used to reduce the dimensions of data before clustering, so that it can simplify datasets that have a large number of records [7]. After dimension reduction using PCA, the data grouping or clustering process can be carried out using the K-Means algorithm. This is a method used in data analysis and machine learning to group data into several clusters based on similar characteristics or attributes. This technique aims to reduce the distance between data in one cluster and increase the distance between

different clusters [8]. The combination of PCA and K-Means is expected to help identify hidden patterns in blood test data for liver disorder patients, so that it can support the early detection process and classification of disease risks more efficiently and accurately.

2. Page layout

Research Stages is a method used to describe the steps taken from identifying problems to getting a solution to existing problems. The stages of business understanding and deployment are not discussed because the scope of this research is on technical aspects. The stages of research carried out can be seen in Figure 1.

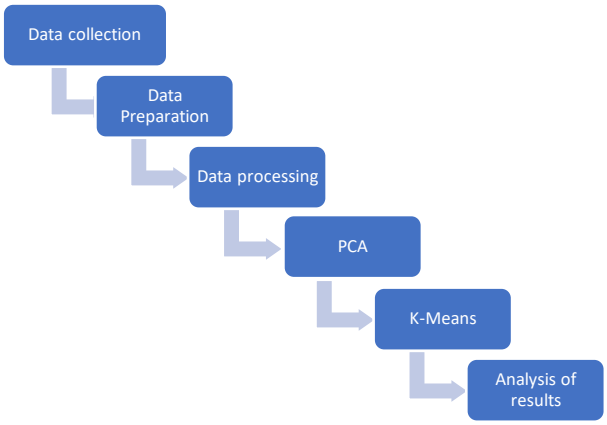

Fig. 1: Research Stages

2.1. Data collection

Data collection is an extensive procedure to gather details about a particular research topic that is carried out in a systematic manner. One thing that researchers must keep in mind is the accuracy and integrity of research data, regardless of the methods and approaches used, researchers must ensure that the data is collected correctly and honestly as it is [9]. The selection of data collection techniques is very important because it can affect the validity and reliability of research results that must be adjusted to the research objectives, the type of data needed, the resources available, and ethical considerations. A combination of several data collection techniques is also often used to gain a more complete understanding of the research problem [10]. The dataset used in this study comes from the UCI Machine Learning Repository, with the name Liver Disorders.

2.2 Preparation Data

This stage aims to ensure that the data used is of good quality, consistent, and ready for further analysis, so that it can produce accurate and reliable results in processing processes such as PCA and clustering algorithms such as K-Means.

2.3 Data Processing

Processing refers to the process of changing or manipulating a material from its original form into a different form. While data processing is the process of manipulating data to produce more useful and meaningful information[11]. Data quality improvement can be done by implementing a comprehensive data processing process[12]

2.4 PCA

Principal Component Analysis (PCA) is one of the methods in multivariate analysis that uses linear transformation to simplify the number of variables in the data. This technique is often applied to obtain primary information from large data and to understand the structure of relationships between variables.[13]

2.5 K-Means

K-Means is one of the popular algorithms in clustering techniques, because it is known to have an easy-to-understand mechanism and fairly efficient performance.[14] K-means clustering is one of the methods in data analysis or data mining that is unsupervised, namely without using data labels, and is used to group data by dividing data into several groups (partitions).[15]

2.6 Results Analysis

This study successfully clustered patient data with potential liver dysfunction using a combination of Principal Component Analysis (PCA) and K-Means Clustering methods. PCA was used to reduce the dimensionality of the data, where two principal components (PC1 and PC2) were able to explain 59.6% of the total variation, which was good enough for the visualization and clustering process.

Clustering using K-Means showed the best results when the number of clusters K = 3, with a Silhouette Score value = 0.202. This value indicates that the separation between clusters is still relatively weak, but the natural clustering pattern can still be recognized. Scatter plot visualization shows that cluster C2 is quite clearly separated, while C1 and C3 tend to overlap.

3. Results and Discussion

This study uses a dataset consisting of 345 rows of data (instances) without blank values, which reflect the results of blood laboratory tests from patients with potential liver dysfunction. There are 7 numeric attributes analyzed, namely: mcv (Mean Corpuscular Volume), alkphos (Alkaline Phosphatase), sgpt, sgot, gammagt, drinks (daily alcohol consumption), and selector. All of these attributes are relevant biochemical indicators in assessing liver function conditions. An example of the data used can be seen in Figure 2.

	mcv	alkphos	sgpt	sgot	gammagt	drinks	selector
1	85	92	45	27	31	0.0	1
2	85	64	59	32	23	0.0	2
3	86	54	33	16	54	0.0	2
4	91	78	34	24	36	0.0	2
5	87	70	12	28	10	0.0	2
6	98	55	13	17	17	0.0	2
7	88	62	20	17	9	0.5	1
8	88	67	21	11	11	0.5	1
9	92	54	22	20	7	0.5	1
10	90	60	25	19	5	0.5	1

Fig. 2: Blood laboratory test data

To reduce the dimensionality and simplify the visualization and processing of the data, the Principal Component Analysis (PCA) method is used. This technique transforms the original features into linear combinations called principal components, with the aim of preserving as much variance in the data as possible.

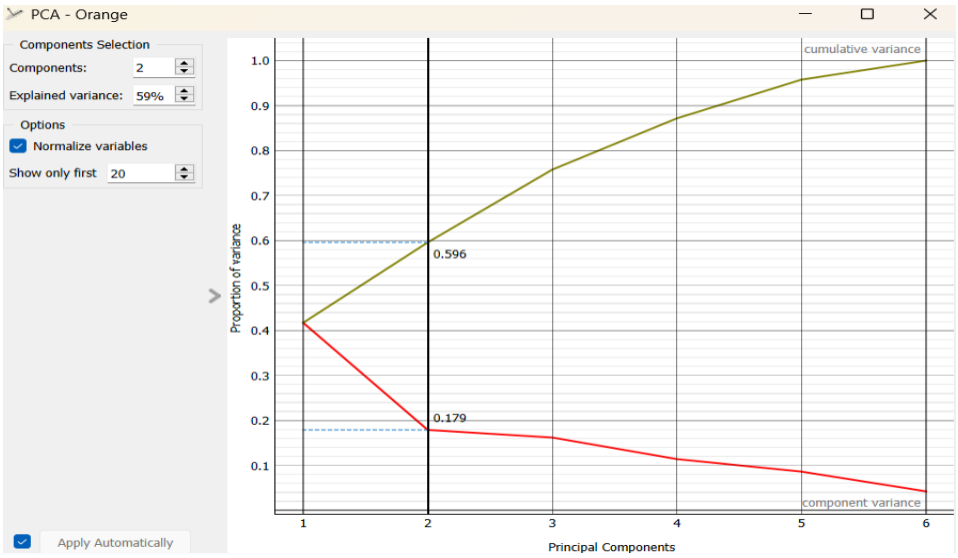

Fig. 3: PCA Results

The PCA results show that the first two principal components (PC1 and PC2) successfully explain 59.6% of the total variation in the data (PC1 = 41.7%, PC2 = 17.9%). This indicates that the important information from the six numeric attributes is successfully condensed into two dimensions, making it very useful for the subsequent visual analysis and clustering stages. These components are then used as input for the K-Means algorithm.

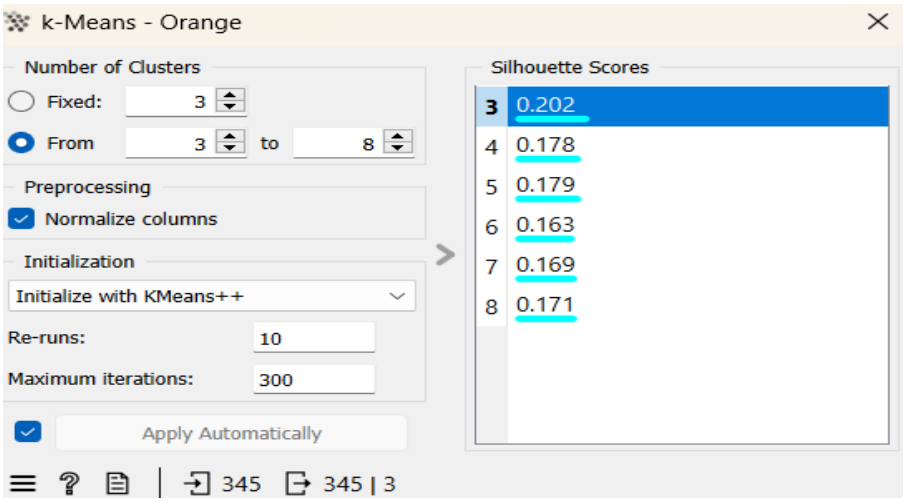

Fig. 4: K-Means

The next stage is data clustering using the K-Means algorithm, which is a partition-based unsupervised learning method. Researchers tried various numbers of clusters, from 3 to 8, and evaluated the quality of the cluster results using the Silhouette Score, a metric that measures the suitability of each data to the cluster it is in compared to other clusters. The evaluation results show that the best selection of the number of clusters is 3 clusters (K = 3), with a Silhouette Score value = 0.202. Although this value does not yet show a very strong separation between clusters (the ideal value is above 0.5), it is enough to show that the data structure contains a natural clustering pattern that can be further utilized.

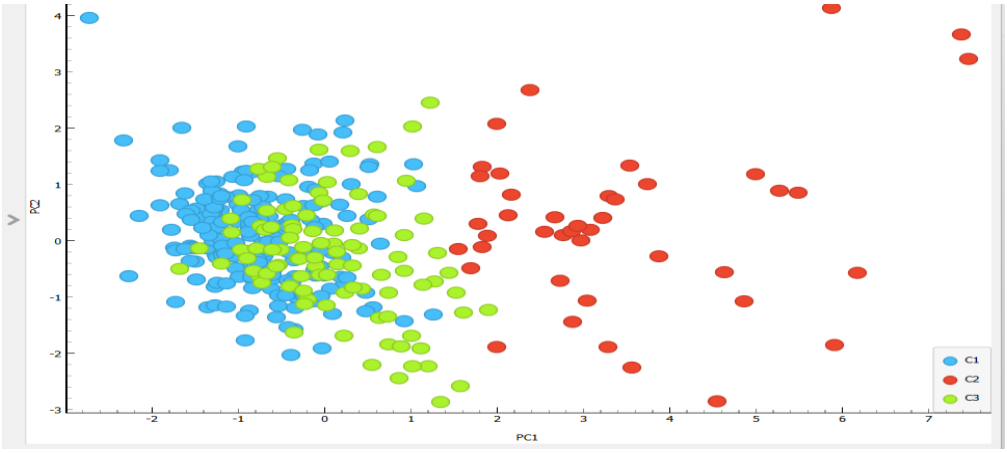

Fig. 5: Scatter Plot

To visually assess the clustering results, a scatter plot was performed based on two PCA components (PC1 and PC2). The visualization shows that the data is successfully separated into three main clusters marked with different colors:
C1 (blue)
C2 (red)
C3 (green)
Visually, cluster C2 appears clearly separated from the other two clusters, which may indicate a group with more extreme or striking biological conditions. In contrast, clusters C1 and C3 appear close to each other and even slightly overlap, indicating possible similarities in characteristics or less striking differences between the two.

4. Conclusion

Based on the results of the study, it can be concluded that the combination of the Principal Component Analysis (PCA) method and the K-Means algorithm is effective for grouping patients based on similarities in blood biochemical patterns. This approach has considerable potential in the application of medical screening, because it can accelerate the process of identifying high-risk groups for liver dysfunction without requiring complicated and time-consuming manual diagnostic procedures. However, there are several limitations that need to be considered. The relatively low silhouette score value indicates that the quality of separation between clusters is not optimal, so other approaches such as DBSCAN or Agglomerative Clustering can be considered to obtain more accurate results. In addition, because the data used does not yet have a clear medical diagnosis label, the evaluation of the clustering results can only be done indirectly. The success of the PCA application itself is greatly influenced by the distribution of variance in the data, in this study only about 59.6% of information can be maintained in the two main components, so there is still important information hidden in the unvisualized dimensions.

5. Suggestion

Further research is suggested to use other clustering methods such as DBSCAN or Hierarchical Clustering to improve the accuracy of cluster separation. The use of data with medical diagnosis labels also needs to be considered so that the evaluation of the results is more objective. In addition, the application of more than two principal components or alternative dimensionality reduction methods such as t-SNE can improve information retention. A more complete and diverse dataset is also needed to obtain more representative results. Finally, the results of this study are expected to be further developed in medical decision support systems.

References

[1] "IJCCS , Vol.x, No.x, Julyxxxx, pp. 1~5ISSN: 1978-1520," vol. 8, no. 3, 2020.
[2] D. C. Adaboost and J. Majapahit, "Implementasi Data Mining Untuk Klasifikasi Penyakit Liver".
[3] R. E. Kristanty, *Solusi Herbal untuk Masalah Liver*. 2024.
[4] A. J. A. Batubara, I. Situmorang, I. R. Nasution, and N. Dumaria, "Mekanisme Detoksifikasi : Cara Sistem Ekskresi Melindungi Tubuh dari Racun," vol. 6, no. 3, pp. 184–190, 2025.
[5] A. Rosida, "Pemeriksaan laboratorium penyakit hati," pp. 123–131.
[6] T. Journal, O. Muhammadiyah, H. Kahar, F. Kedokteran, and U. Airlangga, "PENGARUH HEMOLISIS TERHADAP KADAR SERUM GLUTAMATE PYRUVATE TRANSAMINASE (SGPT) SEBAGAI SALAH SATU PARAMETER," vol. 1, no. 1, 2018.
[7] N. A. Muhaa, L. M. Mulyono, M. R. Fadhilah, and Y. Umaidah, "Klasterisasi Tren Tuberkulosis Global dengan Principal Component Analysis (PCA) dan K-Means," vol. 5, no. 1, pp. 132–143, 2025.
[8] N. A. Maori, "METODE ELBOW DALAM OPTIMASI JUMLAH CLUSTER PADA K-MEANS CLUSTERING," vol. 14, no. 2, pp. 277–287, 2023.
[9] A. P. Nugroho, *Metode Pengumpulan Data*, no. October. 2022.
[10] A. Wardhana, *Teknik Pengumpulan Data Penelitian*, no. July. 2024.
[11] O. S. Udang, M. Tabaru, E. A. M. Sampetoding, and S. Esther, "Pengolahan Data Siswa SMA Negeri 1 Sambuara Kabupaten Kepulauan Talaud Pada Aplikasi DAPODIK," vol. 6, no. 1, pp. 7–11, 2021.
[12] S. Dwididanti, D. A. Anggoro, and M. H. Sutanto, "EMITOR: Jurnal Teknik Elektro," 2019, doi: 10.23917/emitor.v22i2.15677.
[13] M. Billah, M. A. Zartesya, D. S. Prasvita, S. Komp, and M. Kom, "Penerapan Collaborative Filtering , PCA dan K-Means dalam Pembangunan Sistem Rekomendasi Film," no. April, pp. 579–587, 2021.
[14] M. R. Nugroho, I. E. Hendrawan, and P. P. Purwantoro, "Penerapan Algoritma K-Means Untuk Klasterisasi Data Obat Pada Rumah Sakit ASRI," *Nuansa Inform.*, vol. 16, no. 1, pp. 125–133, 2022, doi: 10.25134/nuansa.v16i1.5294.
[15] N. Afiasari, N. Suarna, and N. Rahaningsih, "Implementasi Data Mining Transaksi Penjualan Menggunakan Algoritma Clustering dengan Metode K-Means E-commerce K-Means melakukan analisis penerapan Data Mining dalam mengelompokkan jumlah," vol. 9, pp. 100–110, 2023.