*Journal of Intelligent Decision Support System (IDSS) 6 (4) (2023), pp. 209-215*

Published by: Institute of Computer Science (IOCS)

Journal of Intelligent Decision Support System (IDSS)

Journal homepage: www.idss.iocspublisher.org

# Exploratory Data Analysis (EDA) methods for healthcare classification

Hanna Willa Dhany [1], Sutarman [2], Fahmi Izhari [3]

[1,3] *University of Pembangunan Panca Budi, Medan, Indonesia*
[2] *University of Sumatera Utara, Medan, Indonesia*

**Article Info**

***Article history:***

Received Nov 09, 2023
Revised Nov 29, 2023
Accepted Nov 30, 2023

***Keywords:***

Classification;
EDA;
Healthcare;
Lifestyle;
Patients.

**ABSTRACT**

The recovery and rehabilitation of individuals, helping them regain their physical and mental well-being. Healthcare offers comfort and relief for patients with serious or terminal illnesses, focusing on improving their quality of life and managing symptoms. It plays a role in educating individuals about health risks, disease prevention, and healthy lifestyles. Healthcare contributes to medical research and innovation, leading to advancements in treatments, medications, and medical technologies. Here are some common results and findings that can be obtained through EDA in healthcare data about EDA can reveal the age, gender, and other demographic information of patients. This information is essential for understanding the population served by a healthcare facility. EDA can help identify the prevalence of different diseases or conditions within a patient population. This can assist in resource allocation and healthcare planning. EDA can show how disease rates or healthcare utilization patterns change over time. For example, it can highlight seasonal variations in the incidence of certain diseases. EDA can be used to analyze healthcare data geospatially to identify regions with higher disease prevalence, helping in targeted interventions.



***Corresponding Author:***

Sutarman,
Faculty of Mathematics and Natural Sciences,
University of Sumatera Utara,
Jalan Abdul Hakim, Padang Bulan, Medan Baru, Sumatera Utara 20222, Medan, Indonesia.
Email: sutarman@usu.ac.id

## Introduction

Healthcare is a broad and essential field that encompasses the maintenance or improvement of an individual's physical, mental, and social well-being. It includes various components such as medical services, preventive care, diagnostics, treatment, and rehabilitation. Healthcare is delivered by a diverse range of professionals and organizations, aiming to promote health, prevent illness, and provide medical interventions when necessary. It plays a critical role in society, addressing the well-being of individuals and communities, and is subject to constant advancements and challenges in an ever-evolving landscape.

Disparities in access to healthcare services exist, with many individuals lacking affordable or adequate health insurance coverage. Healthcare costs can be prohibitively high, leading to financial burdens on patients and limiting access to necessary care. Variability in the quality of healthcare services can lead to suboptimal outcomes, medical errors, and patient dissatisfaction. Health disparities based on factors like race, income, and geography persist, resulting in unequal health outcomes among different populations. As populations age, there is increased demand for healthcare services, placing pressure on

healthcare systems to meet the needs of elderly individuals. The rise in chronic conditions, such as diabetes and heart disease, requires long-term management and places strain on healthcare resources. Many regions experience shortages of healthcare professionals, including doctors, nurses, and specialists. (Rahmawati, 2023).

The integration of advanced technology can be challenging, leading to issues like data privacy concerns and unequal access to digital healthcare solutions. Complex healthcare regulations and policies can hinder the efficient delivery of care and contribute to administrative burdens. Events like pandemics and natural disasters can overwhelm healthcare systems and highlight the need for emergency preparedness. Addressing these problems in healthcare is an ongoing and complex task that involves collaboration among policymakers, healthcare providers, and the community to ensure better access, affordability, and quality of care. (Dhany, 2020).

Healthcare aims to prevent illness and promote healthy behaviors through measures such as vaccinations, screenings, and health education. It involves the identification and diagnosis of medical conditions and diseases through various tests, examinations, and medical evaluations. Healthcare provides medical interventions, therapies, and treatments to manage and cure diseases and health issues. After illness or injury, healthcare supports the recovery and rehabilitation of individuals, helping them regain their physical and mental well-being. Healthcare offers comfort and relief for patients with serious or terminal illnesses, focusing on improving their quality of life and managing symptoms. It plays a role in educating individuals about health risks, disease prevention, and healthy lifestyles. Healthcare contributes to medical research and innovation, leading to advancements in treatments, medications, and medical technologies. Healthcare systems work to safeguard public health by monitoring and responding to outbreaks, ensuring clean water and food, and controlling the spread of diseases. Healthcare strives to provide equal access to medical services, addressing health disparities and promoting health equity among diverse populations. In essence, the purpose of healthcare is to ensure that individuals have the opportunity to live healthy lives, receive necessary medical care, and recover from illnesses or injuries, with the ultimate goal of enhancing the overall quality of life for both individuals and society as a whole.

## Method

Exploratory Data Analysis (EDA) is a crucial step in understanding and gaining insights from healthcare data. Here's a general methodology for EDA in healthcare:

- Data Collection: Gather relevant healthcare data from various sources, which may include electronic health records, medical claims, patient surveys, and other health-related datasets.
- Data Cleaning: Address missing values, outliers, and inconsistencies in the data. Impute missing data or decide on appropriate strategies for handling them.
- Data Visualization: Create visualizations such as histograms, box plots, scatter plots, and heatmaps to explore data distributions, relationships, and patterns. Visualization helps in identifying trends and potential anomalies.
- Descriptive Statistics: Calculate summary statistics, including mean, median, standard deviation, and quartiles, to understand the central tendencies and variability of the data.
- Domain Knowledge: Leverage domain expertise to interpret the data and identify key variables that may impact healthcare outcomes. This step can guide the EDA process effectively.
- Feature Engineering: Create new features or transform existing ones to extract more meaningful information. For healthcare, this might involve age groups, risk scores, or other relevant features.
- Correlation Analysis: Examine correlations between variables to identify potential relationships, dependencies, or multicollinearity.
- Time Series Analysis: If dealing with time-series healthcare data, perform time-based analyses to uncover temporal trends and patterns.
- Outlier Detection: Identify and investigate outliers, which may indicate data entry errors, unusual patient cases, or significant health events.
- Hypothesis Testing: Formulate and test hypotheses related to healthcare outcomes, such as the

impact of specific treatments or interventions on patient health.
- Geospatial Analysis: If relevant, analyze data in a geographic context to understand regional healthcare disparities or trends.
- Machine Learning: Apply machine learning algorithms for predictive modeling or clustering to uncover patterns and make predictions related to healthcare outcomes.
- Ethical Considerations: In healthcare, it's critical to consider data privacy and ethics, ensuring that data is handled in compliance with regulations such as EDA.
- Reporting and Communication: Summarize findings and insights in clear, understandable terms, and communicate results to healthcare professionals, stakeholders, or decision-makers.
- Iteration: EDA is often an iterative process. As you uncover insights, you may need to revisit and refine your analysis.

Table 1. Dataset Healthcare

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Name | Tiffany Ramirez | Ruben Burns | Chad Byrd | Antonio Frederick | Mrs. Brandy Flowers |
| Age | 81 | 35 | 61 | 49 | 51 |
| Gender | Female | Male | Male | Male | Male |
| Blood Type | O- | O+ | B- | B- | O- |
| Medical Condition | Diabetes | Asthma | Obesity | Asthma | Arthritis |
| Date of Admission | 17-11-22 | 01-06-23 | 09-01-19 | 02-05-20 | 09-07-21 |
| Doctor | Patrick Parker | Diane Jackson | Paul Baker | Brian Chandler | Dustin Griffin |
| Hospital | Wallace-Hamilton | Burke, Griffin and Cooper | Walton LLC | Garcia Ltd | Jones, Brown and Murray |
| Insurance Provider | Medicare | UnitedHealthcare | Medicare | Medicare | UnitedHealthcare |
| Billing Amount | 37490.98336 | 47304.06485 | 36874.897 | 23303.32209 | 18086.34418 |
| Room Number | 146 | 404 | 292 | 480 | 477 |
| Admission Type | Elective | Emergency | Emergency | Urgent | Urgent |
| Discharge Date | 01-12-22 | 15-06-23 | 08-02-19 | 03-05-20 | 02-08-21 |
| Medication | Aspirin | Lipitor | Lipitor | Penicillin | Paracetamol |
| Test Results | Inconclusive | Normal | Normal | Abnormal | Normal |

Dataset Information:
Each column provides specific information about the patient, their admission, and the healthcare services provided, making this dataset suitable for various data analysis and modeling tasks in the healthcare domain. Here's a brief explanation of each column in the dataset -

Name : This column represents the name of the patient associated with the healthcare record.
Age : The age of the patient at the time of admission, expressed in years.
Gender : Indicates the gender of the patient, either "Male" or "Female."
Blood Type : The patient's blood type, which can be one of the common blood types (e.g., "A+", "O-", etc.).
Medical Condition : This column specifies the primary medical condition or diagnosis associated with the patient, such as "Diabetes," "Hypertension," "Asthma," and more.
Date of Admission : The date on which the patient was admitted to the healthcare facility.
Doctor : The name of the doctor responsible for the patient's care during their admission.
Hospital : Identifies the healthcare facility or hospital where the patient was admitted.

| | | |
|---|---|---|
| Insurance Provider | : | This column indicates the patient's insurance provider, which can be one of several options, including "Aetna," "Blue Cross," "Cigna," "UnitedHealthcare," and "Medicare." |
| Billing Amount | : | The amount of money billed for the patient's healthcare services during their admission. This is expressed as a floating-point number. |
| Room Number | : | The room number where the patient was accommodated during their admission. |
| Admission Type | : | Specifies the type of admission, which can be "Emergency," "Elective," or "Urgent," reflecting the circumstances of the admission. |
| Discharge Date | : | The date on which the patient was discharged from the healthcare facility, based on the admission date and a random number of days within a realistic range. |
| Medication | : | Identifies a medication prescribed or administered to the patient during their admission. Examples include "Aspirin," "Ibuprofen," "Penicillin," "Paracetamol," and "Lipitor." |
| Test Results | : | Describes the results of a medical test conducted during the patient's admission. Possible values include "Normal," "Abnormal," or "Inconclusive," indicating the outcome of the test. |

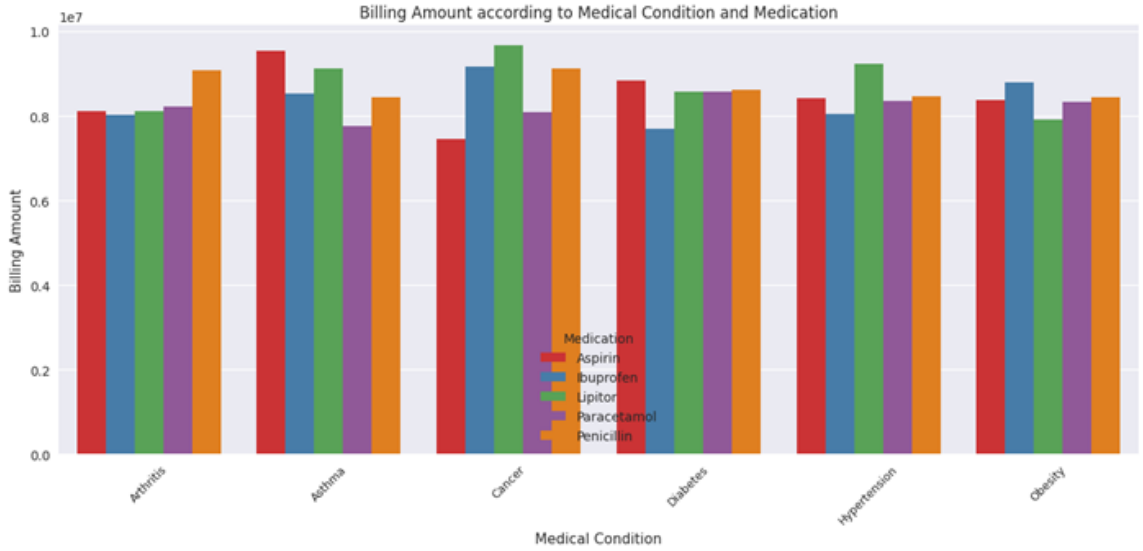
## Results and Discussions

Numerical features refer to data attributes that are represented by numbers. These features contain quantitative information and are typically used for measurements, calculations, or statistical analysis. Numerical features can include variables like age, income, temperature, and any other data that can be expressed as numbers. They are an important component in data analysis, machine learning, and statistical modeling.

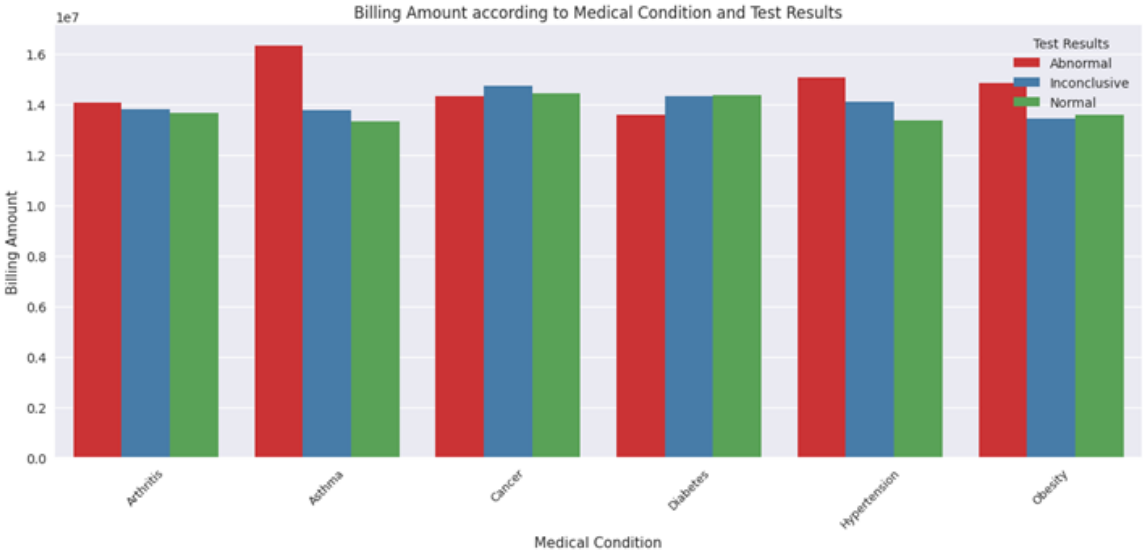
Tables 2. Numerical Features

| | Age | Billing Amount | Room Number |
|---|---|---|---|
| count | 10000 | 10000 | 10000 |
| mean | 51.4522 | 25516.8068 | 300.082 |
| std | 19.588974 | 14067.2927 | 115.806027 |
| min | 18 | 1000.18084 | 101 |
| 25% | 35 | 13506.524 | 199 |
| 50% | 52 | 25258.1126 | 299 |
| 75% | 68 | 37733.9137 | 400 |
| max | 85 | 49995.9023 | 500 |

The following are the results:




Figures 1. Billing Amount According to Medical Condition and Medication




Figures 2. Billing Amount According to Medical Condition and Test Result

The data provides an overview of patient characteristics and healthcare costs from a sample of 10,000 observations. The average age of patients in the dataset is about 51 years old, with a wide age distribution between 18 and 85 years old. This shows the age diversity within the observed population. With regards to the cost of care, the average bill amounted to 25,516.80, with significant variation as reflected by the high standard deviation value (14,067.29). The range of bills also varied significantly, with the minimum bill being around 1,000.18 and the maximum bill reaching 49,995.90. Quartile data shows that 25% of patients had bills below 13,506.52, while 75% had bills below 37,733.91. Regarding inpatient rooms, room numbers varied between 101 to 500, with an average of around 300. This analysis provides initial insights into the distribution of age, cost of care, and room numbers in the dataset, which can form the basis for further analysis and decision-making in a healthcare context.

## Conclusions

Exploratory Data Analysis (EDA) in healthcare can provide valuable insights and help inform decision-making. Here are some common results and findings that can be obtained through EDA in healthcare data about EDA can reveal the age, gender, and other demographic information of patients. This information is essential for understanding the population served by a healthcare facility. EDA can help identify the prevalence of different diseases or conditions within a patient population. This can assist in resource allocation and healthcare planning. EDA can show how disease rates or healthcare utilization patterns change over time. For example, it can highlight seasonal variations in the incidence of certain diseases. EDA can be used to analyze healthcare data geospatially to identify regions with higher disease prevalence, helping in targeted interventions. EDA can uncover patterns in healthcare costs, helping to optimize spending and resource allocation. EDA can identify outliers in healthcare data, which might represent unusual or erroneous data points that need further investigation. EDA can show how patients move through the healthcare system, from diagnosis to treatment and follow-up, which can help in improving patient care and reducing wait times. EDA can be a precursor to predictive modeling in healthcare, where patterns discovered can be used to build models for risk prediction, readmission prediction, or other clinical outcomes. EDA can help segment the patient population into different risk groups, allowing for more personalized healthcare interventions. EDA can reveal patterns of adverse events related to treatments or medications, leading to better patient safety. Future research development suggestions to focus on the integration of EDA with advanced technologies such as artificial intelligence and machine learning in the context of healthcare. Combining exploratory data analysis with more complex predictive models can enable the identification of more subtle patterns and more accurate predictions regarding health risks, clinical outcomes and individual care needs. Further explore geospatial aspects for the development of more spatially targeted intervention strategies. Further understanding of how demographic and environmental variables may interact in the context of public health could help detail more effective and locally tailored health strategies. The integration of new technologies and a focus on the development of more advanced predictive models can increase the capacity of healthcare systems to provide more responsive and personalized care to the population served.

## References

Ahmadi,M., Mohd Osman, M.H. and Aghdam, M.M. (2020). Integrated exploratory factor analysis and Data Envelopment Analysis to evaluate balanced ambidexterity fostering innovation in manufacturing SMEs. Asia Paciųic Management Review25(3),142–155.

Arsyad, Kiagus Muhammad., Yunita, Ariana., Krismartopo, Haniifah Mas'uudah., Dimar, Aghnia Syahputri., Dewi, Kartika., Madrinovella, Iktri. (2023). Revealing Insights Through Exploratory Data Analysis on Earthquake Dataset. Journal of Science and Informatics for Society Volume 1. No 1.

Chong Ho Yu. (2020). Exploratory data analysis in the context of data mining and resampling. International Journal of Psychological Research.

Dhany, Hanna Willa. (2020). Performance Analysis Similarity Matrix, Responsibility Matrix, Availability Matrix, Criterion Matrix of Affinity Propagation. Journal of Physics: Conference Series.

Dhany, Hanna Willa. (2021) Performa Algoritma K-Nearest Neighbour dalam Memprediksi Penyakit Jantung. SENATIKA.

Dhany, Hanna Willa. (2022). Classification of Fungus Types Using The K-Nearest Neighbour Algorithm. Infokum Journal.

Eka Dyar Wahyuni,Amalia Anjani Arifiyanti,Mashita Kustyani. (2019). Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining. Prosiding Nasional Rekayasa Teknologi Industri dan Informasi.

Hidiyanto, Fitra., Leksono, Shabrina., Fajar, Rizqon., Atmaja, Sigit Tri. 2022. Data Exploratory Analysis andFeature Selection of Low-Speed Wind Tunnel Data for Predicting Force andMoment of Aircraft. Journal of Industrial Technology Assessment.

Isa, Indra Griha Tofik., Zulkarnaini., Novianti, Leni., Elfaladonna, Febie., Agustri, Suzan. 2022. Exploratory Data Analysis (EDA) dalam Dataset Penerimaan Mahasiswa Baru Universitas XYZ Palembang. Smart Comp: Jurnal Orang Pintar Komputer.

Jebb AT, Parrigon S, Woo SE. (2017). Exploratory data analysis as a foundation of inductive research. Human ResourceManagement Review.

Mambang., Rinjani, Haniffah Sri., Zulfadhilah, Muhammad., Marleny, Finki Dona., Prastya, Septyan Eka., Cipta, Subhan Panji. (2022). Exploratory Data Analysis of Exact Science and Social Science Learning Contenton Digital Platform. Walisongo Journal of Information Technology, Vol.4No.2. Indonesia.

Maringka, Raissa., Kusnawi. 2021. Exploratory Data Analysis Factors Influence Mental Health in the Workplace. Cogito Smart Journal| VOL. 7

Matthew B. Courtney. (2021). Exploratory Data Analysis in Schools: A Logic Model to Guide Implementation. Kentucky Department of Education.

Mitika Chaudhary, Vinay Prakash and Neeraj Kumari. (2018). Identification Vehicle Movement Detection in Forest Area using MFCC and KNN. Proceedings of the SMART.

Muhammad Radhi, Amalia, Daniel Ryan Hamonangan Sitompul, Stiven Hamonangan Sinurat, Evta Indra. (2021). ANALISIS BIG DATA DENGAN METODE EXPLORATORY DATA ANALYSIS (EDA) DAN METODE VISUALISASI MENGGUNAKAN JUPYTER NOTEBOOK.

Nurdialit, Dwi Gustin. (2020). Exploratory Data Analysis (EDA) and Visualization Using Python. Analytics Vidhya.

Rahmawati, Erni., Wardhani, Ratih Kusuma., Tamsuri, Anas., Wiseno, Bambang. (2023). The Effect Of Health Education On The Knowledge And AttitudesOf Adolescent Health Cadres About Table Fe ConsumptionIn Sma N 1 Kediri Regency..

Samosir, Feliks Victor Parningotan., Mustamu, Loudry Palmarums., Anggara, Erik Dwi., Wiyogo, Albertus Indarko., Widjaja, Andreas. (2021). Exploratory Data Analysisterhadap KepadatanPenumpang Kereta Rel Listrik. JurnalTeknik Informatika dan Sistem Informasi.

Setiawan, Irwan. Suprihanto. (2021). Exploratory data analysis of crime report. Jurnal Manajemen Teknologi dan Informatika.

Vegari, Abi., Setia Budi. (2020). Implementasi Exploratory Data Analysis Pada Dataset Video Trending Harian YouTube. Jurnal Strategi.

Vishwakarma, Swapnil. (2023). Netflix Case Study (EDA): Unveiling Data-Driven Strategies for Streaming. Analytics Vidhya.

Yulina, Syefrida., Elviyenti, Mona. 2022. An Exploratory Data Analysis for Synchronous Online Learning Based on AFEA Digital Images. Jurnal Nasional Teknik Elektro dan Teknologi Informasi | Vol. 11, No. 2